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

Farnoosh Abbas-Aghababazadeh[1,3], Mayer Alvo[1], David R. Bickel[1,2]*

1 Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada, **2** Ottawa Institute of Systems Biology Biochemistry, Microbiology, and Immunology Department, University of Ottawa, Ottawa, Ontario, Canada, **3** Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL, United States of America

* dbickel@uottawa.ca

Abstract

Methods of estimating the local false discovery rate (LFDR) have been applied to different types of datasets such as high-throughput biological data, diffusion tensor imaging (DTI), and genome-wide association (GWA) studies. We present a model for LFDR estimation that incorporates a covariate into each test. Incorporating the covariates may improve the performance of testing procedures, because it contains additional information based on the biological context of the corresponding test. This method provides different estimates depending on a tuning parameter. We estimate the optimal value of that parameter by choosing the one that minimizes the estimated LFDR resulting from the bias and variance in a bootstrap approach. This estimation method is called an *adaptive reference class* (ARC) method. In this study, we consider the performance of ARC method under certain assumptions on the prior probability of each hypothesis test as a function of the covariate. We prove that, under these assumptions, the ARC method has a mean squared error asymptotically no greater than that of the other method where the entire set of hypotheses is used and assuming a large covariate effect. In addition, we conduct a simulation study to evaluate the performance of estimator associated with the ARC method for a finite number of hypotheses. Here, we apply the proposed method to coronary artery disease (CAD) data taken from a GWA study and diffusion tensor imaging (DTI) data.

1 Introduction

Methods of estimating the local false discovery rate (LFDR) [1], not suffering from the bias inherent in estimating other false discovery rates [2], have been applied to various datasets such as high-throughput biological data (e.g., gene expression, proteomics, and metabolomics), diffusion tensor imaging (DTI), and genome-wide association (GWA) study [3–5]. As an example, in a GWA study, the methods of estimating the LFDR are used in order to estimate the probability that a single nucleotide polymorphism (SNP) is associated with a disease

[6–10]. In addition, in DTI brain scans, the LFDR estimates have been used to estimate the proportion of dyslexic-non-dyslexic differences [5, 11].

In many situations, the considered hypotheses are connected by a scientific context. However, ignorance of this scientific context in a data analysis can be misleading, because it may introduce bias into the LFDR estimates [11]. For example, each test in a GWA study corresponds to a specific genetic marker for which previous biological information may be available. Moreover, in a DTI study, each test corresponds to a voxel, where the voxel location can be incorporated as a scientific context.

1.1 Motivating example

We consider coronary artery disease (CAD) data [12], where $N = 394, 839$ SNPs passed the quality control (QC) filtering methods explained in Section 3.2.1. The aim of this study is to identify whether each SNP is associated with a disease. We have two components (z_i, x_i) associated with each hypothesis for $i = 1, \ldots, N$, where z_i is an observed test statistic and x_i presents the minor allele frequency (MAF). For our data, all observed test statistics are used to identify the disease-associated SNPs. Fig 1(A) the $N = 394, 839$ SNPs used to estimate the LFDR (Section 2). A total of 44 disease-associated SNPs with LFDR estimates lower than 0.2 are identified.

A set of hypotheses or features used to determine the posterior probability of a null hypothesis is called a *reference class*, and the problem of finding such a set is an example of the reference class problem (e.g., Bickel [13]). For example, considering all SNPs when estimating the LFDR for a specific MAF, instead of considering the different subsets of SNPs, is called the *combined reference class* (CRC) method (Section 2). From Fig 1(A), consider the example of x_i = 0.0306 and $z_i = -4.3971$, with an estimated LFDR of 0.1973 that is close to the threshold of 0.2. For this MAF, we define a reference class of SNPs in such a way that the MAFs are within a symmetric window around $x_i = 0.0306$, with width 2Δ. Different window widths yield different reference classes. Again, each subset of SNPs is used to estimate the LFDR. Fig 1(B) illustrates the LFDR estimates versus the reference class width. This figure shows that changing the reference class width provides different LFDR estimates, raising the important question of how we can estimate the optimal reference class in order to estimate the LFDR. For the considered CAD data, the reference class problem consists of deciding which SNPs should be used to determine whether a SNP is associated with a disease.

The hypotheses can be divided into groups based on the characteristics of the problem. For example, the CAD data can be divided into two distinct groups according to MAFs, low-frequency SNPs ($1\% \leq x_i \leq 5\%$) and common SNPs ($x_i > 5\%$). Thus, we need to determine which reference class should be used to determine the posterior probability that the SNP is not associated with the disease occurring at the MAF $x_i = 0.0306$. Should we use the entire set of SNPs, or the low-frequency SNPs [14]? In addition, SNPs can be divided into different classes, such as non-synonymous SNPs, genic SNPs, SNPs in highly conserved regions, SNPs in linkage disequilibrium with many (or few) other SNPs, or categorized SNPs based on their MAFs [12].

1.2 Previous research on the reference class problem

Many methods have been proposed for incorporating covariates into statistical techniques for testing multiple hypotheses. Bickel [15] considered the effect of selecting 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 do. Several researchers have applied the idea of incorporating a group structure and weights to improve

A.

B.
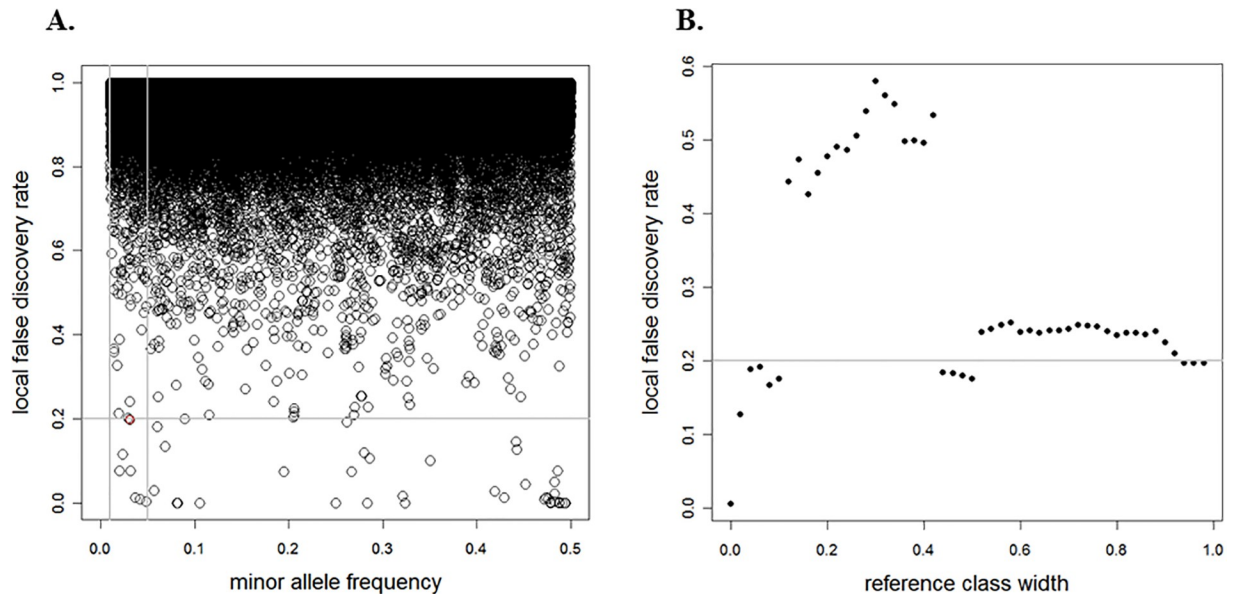
Fig 1. CAD data for LFDR estimates under the HB model versus the MAF x_i (A) and reference class width (B) for a fixed MAF x_i = 0.0306. The horizontal line represents the threshold of 0.2. The vertical lines in (A) indicate the symmetry around x_i = 0.0306 with Δ = 0.04.

the statistical power of tests. This group structure can be used when testing multiple hypotheses by assigning weights for the hypotheses or the p-values in each group. Benjamini and Hochberg [16] used a p-value weighting method to evaluate different procedures. Genovese et al. [17] demonstrated that a p-value weighting procedure can be employed to control the FWER and FDR while increasing the statistical power of the test. Subsequently, Wasserman and Roeder [18] introduced an optimal p-value weighting procedure to control the FWER. Sun et al. [19] proposed a stratified false discovery control approach for genetic studies, in which a large number of hypotheses include inherent stratification. In addition, Efron [11] argued that analyzing separate reference classes can be legitimate from a frequentist viewpoint, and Hu et al. [20] proposed a weighting scheme based on a simple Bayesian framework that employs the proportion of null hypotheses that are true within each group. Such an approach can control the FDR for p-values with certain dependence structures. The unknown proportion of true null hypotheses is estimated within each group. Moreover, Zablocki et al. [8] 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. In contrast to Zablocki et al. [8], instead of specifying the hyperprior distributions required for a hierarchical Bayesian approach, we follow Karimnezhad and Bickel [21] and use an empirical Bayes approach to estimate an optimal reference class to improve the LFDR estimate.

In this study, we assume the prior probability to be a function of covariates (Section 2.1). Then, we propose an adaptive reference class (ARC) method for estimating the LFDR, using a bootstrap approach to estimate the optimal reference class (Section 2.2). We compare the performance of the proposed ARC method and the CRC method using the mean squared error (MSE) as the performance criterion. We prove that, under certain assumptions, the ARC method has an MSE that is asymptotically no greater than that of the CRC method (see S1 File). In addition to the asymptotic results, we conduct a simulation study to investigate the finite dataset performances of the LFDR estimators for each method in Section 3.1. Both the

asymptotic and simulation results show that, under certain assumptions, the ARC method performs well compared with the CRC method. We present an application of the ARC method on both CAD data and DTI data in Section 3.2 in order to demonstrate the practical importance of deciding between the ARC and CRC methods. Finally, we conclude the paper with a brief discussion in Section 4. The proof for the main theorem is included in S1 File.

2 Materials and methods

Suppose N null hypotheses $H_{01}, \ldots, H_{0N}$ are considered simultaneously. For example in GWA study, let H_{0i} denote the null hypothesis that the ith SNP is not associated with the disease. Under the genetic additive model [22], each SNP yields a $Wald\ \chi^2$ test statistic W_i. Under the ith null hypothesis, it holds that $W_i \sim \chi_1^2$, while under the ith alternative hypothesis, we have $W_i \sim \chi_{1,\delta}^2$, where $\delta \in (0, \infty)$ is an unknown noncentrality parameter, following the models employed in [23] and [24]. Under the ith null hypothesis, assume that $Z_i \sim N(0, 1)$, where Z_i represents the z-transform that converts the $Wald\ \chi^2$ statistic into a standard normal statistic. In addition, for DTI data, let H_{0i} denote the null hypothesis that there is no dyslexic-non-dyslexic difference for the ith voxel. Under the ith null hypothesis, assume that $Z_i \sim N(0, 1)$, where Z_i represents the z-transform which converts two-sample t-test satistic into a standard normal statistic.

The observed statistics $z = (z_1, \ldots, z_N)^T$ are considered realizations of $\mathbf{Z} = (Z_1, \ldots, Z_N)^T$. Let A_i be an indicator variable for the event that the ith alternative hypothesis H_{ai} is true. Assume that A_i's are independent and identically distributed (i.i.d.) Bernoulli$(1 - \pi_0)$ variables, where π_0 is the prior probability that the ith null hypothesis is true. Let $f_0(z_i)$ and $f_1(z_i)$ be the null and alternative density, respectively.

The posterior probability that the ith null hypothesis is true, given $Z_i = z_i$ is the LFDR [1], and is denoted as $\Psi(z_i)$, where

$$\Psi(z_i) = \mathrm{P}(A_i = 0 | Z_i = z_i) = \frac{\pi_0 f_0(z_i)}{f(z_i; \pi_0)}, \tag{1}$$

where $f(z_i; \pi_0)$ denotes the mixture density of Z_i given by

$$f(z_i; \pi_0) = \pi_0 f_0(z_i) + (1 - \pi_0) f_1(z_i). \tag{2}$$

If the null hypothesis is true, then the null density $f_0(z_i)$ of the statistic Z_i is the standard normal, and is called the *theoretical null hypothesis* [1]. The model defined in Eq (2) with the following method of estimation is called the *histogram- based* (HB) model [25]. By assuming the theoretical null hypothesis and applying the Poisson regression, the mixture density $f(z_i; \pi_0)$ is estimated by fitting a high-degree polynomial to the histogram counts, denoted by $\widehat{f}_i(\mathbf{z})$, where the estimate of the proportion of true null hypothesis is denoted by $\widehat{\pi}_0(\mathbf{z})$. The LFDR estimate $\widehat{\Psi}_i(\mathbf{z})$ is computed by substituting $\widehat{\pi}_0(\mathbf{z})$ and $\widehat{f}_i(\mathbf{z})$ into Eq (1). This model and estimation method are an example of the CRC method.

2.1 Proposed model

The described model in Eq (1) extends to the situation that incorporates a covariate related to the scientific context of each hypothesis test. For the CAD dataset, the covariate represents the MAF in the CAD data, while for the DTI data, the location is incorporated as a covariate. Let $\mathbf{X} = (X_1, \ldots, X_N)^T$ be i.i.d. random variables. Any test statistics are transformed to the standard normal statistic Z_i, for $i = 1, \ldots, N$. The observed statistics vector $z = (z_1, \ldots, z_N)^T$ is considered a realization of $\mathbf{Z} = (Z_1, \ldots, Z_N)^T$. Let A_i be the event that the ith alternative hypothesis H_{ai} is

true. Assume that $A_i | X_i = x_i \sim$ Bernoulli$(1 - \pi_0(x_i))$, where $\pi_0(x_i)$, the prior probability that the ith null hypothesis is true, is an unknown function of the given covariate $X_i = x_i$. We denote the posterior probability that the ith null hypothesis is true, given $Z_i = z_i$ and $X_i = x_i$ by

$$\Psi(z_i; x_i) = \text{P}(A_i = 0 | Z_i = z_i, X_i = x_i) = \frac{\pi_0(x_i) f_0(z_i)}{f(z_i; \pi_0(x_i))}, \tag{3}$$

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

$$f(z_i; x_i) = \pi_0(x_i) f_0(z_i) + (1 - \pi_0(x_i)) f_1(z_i; x_i), \tag{4}$$

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 Eq (2) is a special form of Eq (4), where the effect of the covariates is ignored. 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 Eq (3). Under the CRC method, the effects of the covariates defined in Eq (1) are ignored, while under the proposed ARC method, some assumptions are considered locally in order to estimate the LFDR $\Psi(z_i; x_i)$, defined in Eq (3).

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 given covariate $X_i = x_i$. Such a symmetric window is denoted by z_i^Δ, where

$$z_i^\Delta = \{z_j : |x_j - x_i| \leq \Delta, \ j = 1, \ldots, N\}. \tag{5}$$

Let Δ_0 denote the smallest considered value of the tuning parameter Δ. The reference class z_i^Δ contains components z_j such that their covariates are within a distance Δ of x_i. Denoting the expected dimension of the reference class z_i^Δ by d_i^Δ, we have

$$\text{d}_i^\Delta = N\text{P}(|X_j - x_i| \leq \Delta, \ j = 1, \ldots, N). \tag{6}$$

Here, d_i^Δ increases with number of null hypothesis tests N, provided that the probability is positive. For each reference class z_i^Δ, we may apply any LFDR estimation approach such as the HB model in Section 2. In contrast to the CRC method, instead of using the entire collection of observed statistics z, only the reference class z_i^Δ is used to obtain the LFDR estimate $\widehat{\Psi}_i(z_i^\Delta)$. The choice of tuning parameter Δ influences the LFDR estimate. Here, we choose the one that results in the lowest error when estimating the LFDR, which is called the *optimal* tuning parameter.

2.3 Optimal tuning parameter

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

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

$$\begin{aligned} \mu_\Delta(x_i) &= \text{E}(\widehat{\Psi}_i(z_i^\Delta) | X_i = x_i), \\ \sigma_\Delta^2(x_i) &= \text{E}[(\widehat{\Psi}_i(z_i^\Delta) - \mu_\Delta(x_i))^2 | X_i = x_i], \end{aligned} \tag{7}$$

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

with

$$\mathcal{B}_\Delta(x_i) = \mathrm{E}[(\widehat{\Psi}_i(\boldsymbol{z}_i^\Delta) - \Psi(z_i; x_i))|X_i = x_i].\tag{8}$$

Determining the optimal choice of Δ depends on the choice of the loss function used to measure the errors in the estimation of the LFDR. A good estimator is accurate in the sense that its estimates are as close to the true values as possible. Accuracy measures typically take into account the difference between the estimated value and the true value. Using the MSE as a an accuracy measure is a commonly used way to indicate how close the estimator is to the true value by incorporating both the bias and the variance [26]. Hence, using the MSE provides our criterion for defining the optimal tuning parameter Δ. The MSE for the estimator $\widehat{\Psi}_i(\boldsymbol{z}_i^\Delta)$ conditional on $X_i = x_i$ is defined as

$$\mathrm{MSE}(\widehat{\Psi}_i(\boldsymbol{z}_i^\Delta)|X_i = x_i) = \mathrm{E}[(\widehat{\Psi}_i(\boldsymbol{z}_i^\Delta) - \Psi(z_i; x_i))^2|X_i = x_i].\tag{9}$$

It can be shown that the portion of MSE that depends on Δ is given by

$$\mathrm{err}(\widehat{\Psi}_i(\boldsymbol{z}_i^\Delta)|X_i = x_i) = \sigma_\Delta^2(x_i) + \mathcal{B}_\Delta^2(x_i).\tag{10}$$

Here, we employ the errors resulting from the bias and the variance in Eq (10) to determine the optimal Δ. Denoting the optimal Δ by $\Delta_{0i}^\star$, we have

$$\Delta_{0i}^\star = \arg\inf_{\Delta \geq \Delta_0} \mathrm{err}(\widehat{\Psi}_i(\boldsymbol{z}_i^\Delta)|X_i = x_i).\tag{11}$$

To estimate $\Delta_{0i}^\star$, it is necessary to estimate the variance and the prediction bias of the LFDR estimator, which we do using the bootstrap approach.

2.4 Bootstrap estimation of the optimal tuning parameter

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

$$\boldsymbol{z}_{i,b}^\Delta = \{z_{bj}^\star : |x_{bj}^\star - x_i| \leq \Delta,\ j = 1, \ldots, N\}.\tag{12}$$

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

$$\widehat{\mu}(\Delta, B) = \frac{1}{B}\sum_{b=1}^B \widehat{\Psi}_i(\boldsymbol{z}_{i,b}^\Delta) \ \ \text{and} \ \ \widehat{\sigma}^2(\Delta, B) = \frac{1}{B-1}\sum_{b=1}^B \left(\widehat{\Psi}_i(\boldsymbol{z}_{i,b}^\Delta) - \widehat{\mu}(\Delta, B)\right)^2.\tag{13}$$

In order to estimate the prediction bias in Eq (10), we need to estimate $\pi_0(x_i)$. We propose using a reference class $\boldsymbol{z}_{i,b}^{\Delta_0}$, which contains the observed statistics z_js the covariates of which are within a distance Δ_0 of x_i. Thus, the estimator $\widehat{\mu}(\Delta_0, B)$ from Eq (13) can be used to estimate $\pi_0(x_i)$. Denoting the bootstrap estimator of the prediction bias by $\widehat{\mathcal{B}}(\Delta, \Delta_0, B)$, we have that

$$\widehat{\mathcal{B}}(\Delta, \Delta_0, B) = \widehat{\mu}(\Delta, B) - \widehat{\mu}(\Delta_0, B).\tag{14}$$

The estimator of $\mathrm{err}(\widehat{\Psi}_i(\boldsymbol{z}_i^\Delta)|X_i = x_i)$ in Eq (10) is denoted by $\widehat{\mathrm{err}}(\Delta, \Delta_0, B)$, and is computed by simply summing the bootstrap variance in Eq (13) and the squared bootstrap prediction

bias in Eq (14). Let the optimal $\Delta^\star_{0i}$ be denoted as $\widehat{\Delta}^\star_{0i}$, which is given by

$$\widehat{\Delta}^\star_{0i} = \arg \inf_{\Delta \geq \Delta_0} \widehat{\mathrm{err}}(\Delta, \Delta_0, B). \tag{15}$$

After estimating the optimal tuning parameter $\widehat{\Delta}^\star_{0i}$ (see Algorithm 1), the optimal reference class $z_i^{\widehat{\Delta}^\star_{0i}}$ is estimated from Eq (5). The class contains z_js, the covariates of which are within a distance $\widehat{\Delta}^\star_{0i}$ of x_i. Then, the optimal reference class $z_i^{\widehat{\Delta}^\star_{0i}}$ is used to estimate the LFDR in Eq (3). This LFDR estimate is denoted as $\widehat{\Psi}_i(z_i^{\widehat{\Delta}^\star_{0i}})$. The estimation methods detailed above yield two estimators. The estimator $\widehat{\Psi}_i(z)$ is related to the CRC method and $\widehat{\Psi}_i(z_i^{\widehat{\Delta}^\star_{0i}})$ is computed using the ARC method. We compare the performance of two estimators using the MSE.

Algorithm 1 Pseudo-code of the estimation of the optimal tuning parameter for one simulated dataset.

```
Input: Test statistics and covariates (z₁, x₁),..., (zₙ, xₙ); number of
bootstrap samples B; smallest value of the tuning parameterΔ₀; tuning
parameter Δ ≥ Δ₀
For i = 1,...N
```

- Build b^{th} bootstrap samples $(z^\star_b, x^\star_b) = \{(z^\star_{bj}, x^\star_{bj}), \ldots, (z^\star_{bj}, x^\star_{bj})\}$, $b = 1, \ldots, B$ and $j = 1, \ldots, N$ including pair (z_i, x_i) For $b = 1, \ldots, B$
  - Determine bootstrap reference class $z^{\Delta_0}_{i,b}$ and $z^{\Delta}_{i,b}$ in Eq (12)
  - Estimate LFDR $\widehat{\Psi}_i(z^\Delta_{i,b})$ and $\widehat{\Psi}_i(z^{\Delta_0}_{i,b})$ using HB method
- Compute $\widehat{\mu}(\Delta, B)$, $\widehat{\sigma}^2(\Delta, B)$, $\widehat{\mu}(\Delta_0, B)$, $\widehat{\mathcal{B}}(\Delta, \Delta_0, B)$ and $\widehat{\mathrm{err}}(\Delta, \Delta_0, B)$ in Eqs (10), (13) and (14)
- Minimize $\widehat{\mathrm{err}}(\Delta, \Delta_0, B)$ over $\Delta \geq \Delta_0$

```
Output: Estimate optimal tuning parameter Δ̂*₀ᵢ.
```

Let $\pi_0(x_i)$ denote the true prior probability that the ith null hypothesis is true. In GWA study, the null hypothesis means no disease association, and in the DTI study, it means no differences between dyslexic and non-dyslexic children. For a given x_0, we suppose that the unknown prior probability $\pi_0(X_i)$ is a step function of the covariate X_i, given by

$$\pi_0(X_i) = \begin{cases} \pi_{01} & \text{if } X_i \leq x_0, \\ \pi_{02} & \text{if } X_i > x_0, \end{cases} \tag{16}$$

where the prior probabilities π_{01} and π_{02} are both unknown, and $\pi_{01} \leq \pi_{02}$. This function splits the N tests into two distinct groups, such that in each group, the test statistics are i.i.d. Moreover, the simplified function in Eq (16) will have a biologically meaningful interpretation. As an example, in CAD data, the N SNPs can be divided into two distinct groups; disease-associated and not disease-associated. The two groups may have different choices of prior probabilities. In addition, under the assigned values of x_0 and Δ_0, the observed vector of covariates x may be partitioned into three regions

$$\begin{aligned} \mathcal{R}_1(x_0, \Delta_0) &= \{x_j : x_j \leq x_0 - \Delta_0\}, \\ \mathcal{R}_2(x_0, \Delta_0) &= \{x_j : x_0 - \Delta_0 < x_j < x_0 + \Delta_0\}, \\ \mathcal{R}_3(x_0, \Delta_0) &= \{x_j : x_j \geq x_0 + \Delta_0\}, \end{aligned} \tag{17}$$

for $j = 1, .., N$. Therefore, the following theorem ensures us that, under the assumptions

explained above for $\pi_0(X_i)$ Eq (16) and the region of covariates Eq (17), the proposed ARC method has an MSE that is asymptotically no greater than that of the CRC method. The proof of the theorem is given as a series of lemmas (see S1 File).

Theorem 1 *Let* $\widehat{\Psi}_i(z)$ *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*

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

where B is the number of bootstrap samples.

Remark 1 *The step function of the covariates in* Eq (16) *and the partitioning of the covariates in* Eq (17) *affect the weak consistency of* $\widehat{\mu}(\Delta_0, B)$ Eq (13) *as an estimator of* $\pi_0(X_i)$ *in* Eq (16) *(see* S1 File, *Lemma 2). Then, for* $x_i \in \mathcal{R}_2(x_0, \Delta_0)$, *the bootstrap mean* $\widehat{\mu}(\Delta_0, B)$ *is not a weakly consistent estimator of* $\pi_0(X_i)$.

Remark 2 *For* $x_i \in \mathcal{R}_3(x_0, \Delta_0)$, *similar results to Theorem 1 can be derived.*

3 Results

3.1 Simulation study

The aim of the simulation analysis presented here is to compare the finite dataset performances of the CRC and ARC methods when estimating the LFDR in Eq (3). In this section, each test statistic is assigned a prior probability that is a function of the covariate.

We assume that the proportion of disease-associated tends to be very small. Then, we present several simulation studies, each with a different value of $x_0 \in [0.05, 0.40]$. The datasets are simulated as follows. In each simulation, we randomly generate 1000 datasets, each corresponding to an artificial case-control study. For each dataset, we simultaneously generate both the auxiliary information and the observed Wald χ^2 test statistics, 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, the true prior probability $\pi_0(x_i)$ is determined according the given value of x_0 as a function of the observed covariate in Eq (16). From (16), let $\bar{\pi}_0 = \mathrm{E}(\pi_0(X_i))$ for $i = 1,\ldots, N$, where $\bar{\pi}_0 \in [0.60, 0.95]$. To generate the observed statistics, we generate each $A_i \sim \mathrm{Bernoulli}(1 - \pi_0(x_i))$ 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 $\delta \in [1.5, 7]$ is assigned. The Wald χ^2 test statistics when $A_i = 0$ are sampled from χ^2_1. The *Wald* χ^2 test statistics are then transformed into z-values.

Each dataset has N pairs (z_i, x_i). The total number of pairs N is equal to 300, 000. In each simulation, a pair (z_i, x_i) is selected randomly from each dataset to estimate $\Psi(z_i; x_i)$. 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 in order to determine $\widehat{\Delta}_{0i}^\star$. We consider the range $\Delta_0 \in (0, x_0)$, and set $B = 1000$. Thus, under the HB model described in Section 1.1, we compute the estimators $\widehat{\Psi}_i(z)$ and $\widehat{\Psi}_i(z_i^{\widehat{\Delta}_{0i}^\star})$. The conditional MSE approximations used to measure the performances of the estimators are given by

$$\widehat{\mathrm{MSE}}(\widehat{\Psi}_i|\mathcal{R}_r(x_0, \Delta_0)) = \frac{1}{\#\{x_i \in \mathcal{R}_r(x_0, \Delta_0)\}} \sum_{x_i \in \mathcal{R}_r(x_0, \Delta_0)} \left(\widehat{\Psi}_i - \Psi(z_i; x_i)\right)^2 \quad r = 1, 2, 3,$$

and the marginal MSE approximations are computed as follows:

$$\widehat{\text{MSE}}\left(\widehat{\Psi}_i\right) = \frac{1}{1000}\sum_{i=1}^{1000}\left(\widehat{\Psi}_i - \Psi(z_i; x_i)\right)^2,$$

where $\widehat{\Psi}_i = \widehat{\Psi}_i(z)$ for the CRC method and $\widehat{\Psi}_i = \widehat{\Psi}_i(z_i^{\widehat{\Delta}_{0i}^\star})$ for the ARC method. The relative MSE of the two estimators is a convenient measure for comparing the MSEs. The conditional and marginal relative MSEs are denoted by

$$\text{ReMSE}_{\text{cond}} = \frac{\widehat{\text{MSE}}(\widehat{\Psi}_i(z_i^{\widehat{\Delta}_{0i}^\star})|\mathcal{R}_r(x_0, \Delta_0))}{\widehat{\text{MSE}}(\widehat{\Psi}_i(z)|\mathcal{R}_r(x_0, \Delta_0))} \text{ and } \text{ReMSE}_{\text{marg}} = \frac{\widehat{\text{MSE}}(\widehat{\Psi}_i(z_i^{\widehat{\Delta}_{0i}^\star}))}{\widehat{\text{MSE}}(\widehat{\Psi}_i(z))}, \quad (18)$$

respectively. From Fig 2, we observe that the performance of the ARC method depends on the Δ_0 values and the region of the 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 $\mathcal{R}_3(x_0, \Delta_0)$ that follow the results in Theorem 1. Then, increasing $\bar{\pi}_0$, Fig 2(D) shows that the
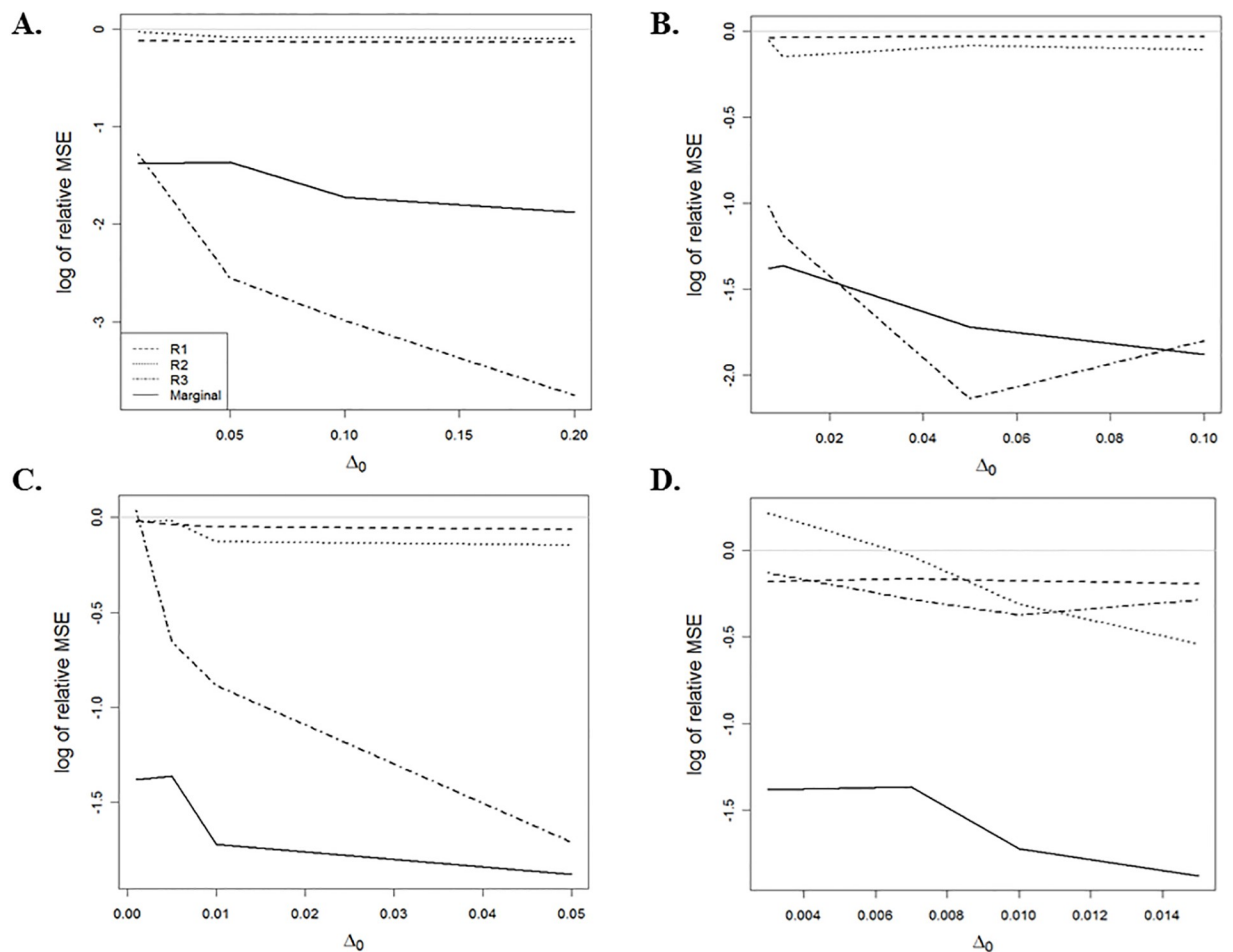
Fig 2. The $\log_2$ value of the relative MSE conditional on regions and marginal versus different values of Δ_0; for (A) $\bar{\pi}_0 = 0.60$, (B) $\bar{\pi}_0 = 0.80$, (C) $\bar{\pi}_0 = 0.90$ and (D) $\bar{\pi}_0 = 0.95$, where $\bar{\pi}_0 = \text{E}(\pi_0(X_i))$ for $i = 1, \ldots, N$.

https://doi.org/10.1371/journal.pone.0206902.g002

Table 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.

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

https://doi.org/10.1371/journal.pone.0206902.t001

MSE approximation for the proposed ARC method is greater than that of the CRC method in the region $\mathcal{R}_2(x_0, \Delta_0)$, for some Δ_0.

Fig 2 shows that the ARC method has a smaller marginal MSE approximation than that of the CRC method. From Table 1, if we consider the true prior probabilities for all SNPs to be independent of the covariate, that is, $\pi_0(x_i) = \pi_0$ for $i = 1, \ldots, N$, then the CRC method has a smaller MSE than that of the ARC method. In such cases, the CRC method should be used instead of the ARC method to analyze the data.

3.2 Real data analysis

We apply both the ARC and CRC methods to the CAD and DTI datasets, and compare the disease-associated SNPs and dyslexic-non-dyslexic difference voxels under each method, respectively. The purpose of this comparison is to demonstrate the practical difference between the methods rather than to determine which method performs better.

3.2.1 CAD data analysis. The CAD dataset originating from the United Kingdom includes 500, 568 SNPs genotyped for 2, 000 cases, and 3, 000 combined on 22 autosomal chromosomes. The control individuals come from two groups: 1500 individuals from the 1958 British Birth Cohort (58C), and 1500 individuals selected from UK blood services (UKBS) controls. Following [12], we use quality control filtering methods to exclude SNPs based on the exact Hardy-Weinberg equilibrium (HWE) test as well as individuals or SNPs with many missing genotypes. The following three filters are applied sequentially. For the first filter, an SNP fails when the proportion of missing data proportion is greater than 0.05 or when the minor allele frequency (MAF) is smaller than 0.05 and the missing data proportion is greater than 0.01. For the second filter, SNPs with p-values smaller than 0.05 using the exact HWE test in the combined controls are rejected. Finally, for the third filter, we reject SNPs with p-values smaller than 5×10^{-7} using trend tests and general genotype tests between each case and the combined controls. We also excluded SNPs with MAFs smaller than 0.01. A total of $N = 394839$ SNPs passed these quality control filtering methods and are used to identify disease-associated SNPs. We apply both the ARC and the CRC methods to the CAD data and compare the disease-associated SNPs under each method.

The CAD related data introduced in Section 1.1, with $N = 394, 839$ SNPs, is employed in the following statistical analysis. Under the CRC method, all observed statistics z are considered to estimate the LFDR, where 44 disease-associated SNPs are identified with LFDR lower than 0.2. The MAF is incorporated as a covariate. Under the ARC method, the optimal reference class is estimated for each MAF, which depends on the choice of Δ_0. Fig 3(A) presents the LFDR estimates under the CRC method versus the ARC method when $\Delta_0 = 0.001$. The results show that, 160 SNPs are disease-associated based on the ARC method, while the CRC method detects 44 disease-associated SNPs. From Fig 3(B), we find that changing Δ_0 has a direct effect on the number of disease-associated SNPs. Under the ARC method, increasing the value of Δ_0 brings the proportion of disease-associated SNPs closer to the corresponding proportion under the CRC method.
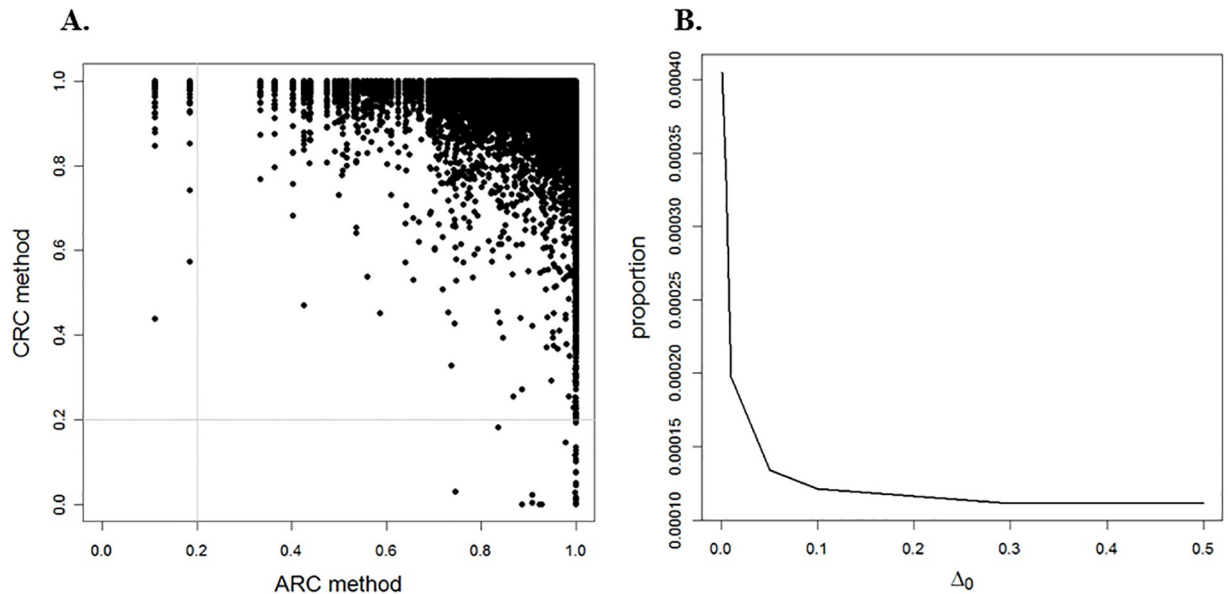
Fig 3. CAD data analysis. (A) presents the LFDR estimate under the ARC method for $\Delta_0 = 0.001$ versus that for the CRC method and (B) illustrates the proportion of disease-associated SNPs under the ARC method when the LFDR estimate is less than 0.2 versus $\Delta_0 \in (0, 0.50)$.

3.2.2 DTI data analysis. Schwartzman et al. [5] used 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 provide information regarding the direction of diffusion. The measured diffusivity; that is, the diffusion coefficient, relates the diffusive flux to a concentration gradient [27], and has units of (mm^2/s). In this study, 12 children were tested and divided equally in each group (i.e., dyslexic or non-dyslexic group). Each child received DTI brain scans in N = 15443 locations, with each represented by its own voxel's response. The aim is to determine the dyslexic-non-dyslexic difference at the ith voxel (location), in relation to reading development in children aged 7-13 [28]. Each test corresponds to a specific voxel. We have two components (z_i, x_i) associated with each hypothesis for i = 1,...,N, where z_i is an observed test statistic that compares the dyslexic children with those who are not (see in Section 2), and x_i is the location (i.e., the distance from back of brain to the front). We apply both the ARC and CRC methods to the DTI data and compare the dyslexic-non-dyslexic difference voxels under each method. The DTI brain scans data with a total of N = 15443 locations, each represented by its own voxel's response, is employed in the following statistical analysis.

Let $W_i = Z_i^2$. The observed statistics $\boldsymbol{w} = (w_1, w_2,..., w_N)^T$ are considered as a realization of $\boldsymbol{W} = (W_1, W_2,..., W_N)^T$. Under the ith null hypothesis, it holds that $W_i \sim \chi_1$, while under the ith alternative hypothesis $W_i \sim \chi_{1,\delta}$, where $\delta \in (0, \inf)$ is an unknown noncentrality parameter, according to the models employed in [23, 24]. Therefore, the LFDR in Eq (1) is defined as follows

$$\Psi(w_i) = P(A_i = 0 | W_i = w_i) = \frac{\pi_0 g_0(w_i)}{g(w_i; \pi_0, \delta)}, \tag{19}$$

where $g_0(w_i) \sim \chi_1$ (i.e., null density), and $g(w_i; \pi_0, \delta)$ denotes the mixture density given by

$$g(w_i; \pi_0, \delta) = \pi_0 g_0(w_i) + (1 - \pi_0)g_1(w_i; \delta), \tag{20}$$
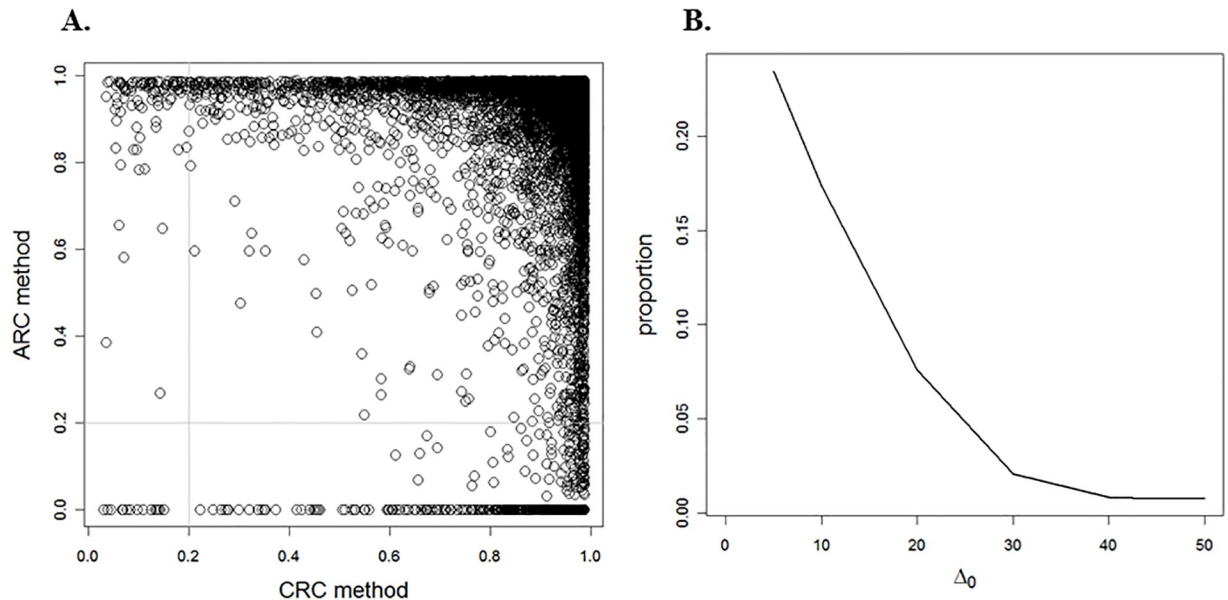
Fig 4. Brain data analysis. (A) presents the LFDR estimate under the ARC method for $\Delta_0 = 20$ versus that for the CRC method and (B) illustrates the proportion of dislexic-non-dyslexic difference voxels under the ARC method when the LFDR estimate is less than 0.2 versus $\Delta_0 \in (0, 50)$.

where $g_1(w_i; \delta)$ represents the unknown alternative density. Under the CRC method, all observed statistics $\boldsymbol{w}$ are considered to estimate the LFDR using the *type II maximum likelihood estimation* (MLE) model [24]

$$l(\pi_0, \delta) = \sum_{i=1}^{N} \log(\pi_0 g_0(w_i) + (1 - \pi_0)g_1(w_i; \delta)), \tag{21}$$

where $\widehat{\pi}_0 = 0.923$ and $\widehat{\delta} = 3.7$, and 119 dyslexic-non-dyslexic difference voxels are identified with LFDR lower than 0.2. Then, the brain location is incorporated as a covariate. Under the ARC method, the optimal reference class is estimated for each location, which depends on a choice of Δ_0. Fig 4(A) presents the LFDR estimates under the CRC method versus the ARC method when $\Delta_0 = 20$. We observe from Fig 4(B), 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.

4 Discussion and conclusion

In this study, we employ a novel approach that incorporates a covariate (i.e., a scientific context corresponding to each hypothesis test) to improve the LFDR estimate when identifying alternative hypotheses. Using 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 given in Eq (16), Theorem 1 states that the ARC method has an MSE asymptotically no greater than that of the CRC method. It would be interesting to investigate whether this result holds for a general prior probability. We leave this topic for future research. In addition, the simulation indicates that the ARC method performs in comparison to the CRC method for a finite number of hypotheses. Our simulation results confirm that for 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 approximation than that of the CRC method (see Fig 2). Moreover, we could not prove the weak consistency of $\widehat{\mu}(\Delta_0, B)$ as an estimator of $\pi_0(x_i)$ in region $\mathcal{R}_2(x_0, \Delta_0)$ (see S1 File, Lemma 2). The ARC method was applied to both CAD and DTI datasets, as illustrated in Figs 3 and 4. Regardless of LFDR estimation methods (i.e., HB and MLE), 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. This suggests that further investigation may be necessary on how the tuning parameter Δ_0 can be controlled to improve results.

Supporting information

S1 File. The proof of Theorem 1 proceeds by a series of lemmas.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: Mayer Alvo, David R. Bickel.

Investigation: Farnoosh Abbas-Aghababazadeh, David R. Bickel.

Methodology: Mayer Alvo, David R. Bickel.

Project administration: Mayer Alvo, David R. Bickel.

Resources: Mayer Alvo, David R. Bickel.

Supervision: Mayer Alvo, David R. Bickel.

Validation: Farnoosh Abbas-Aghababazadeh.

Visualization: Farnoosh Abbas-Aghababazadeh.

Writing – original draft: Farnoosh Abbas-Aghababazadeh.

Writing – review & editing: Farnoosh Abbas-Aghababazadeh, Mayer Alvo, David R. Bickel.

References

1. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association. 2001; 96(456):1151–1160. https://doi.org/10.1198/016214501753382129

2. Bickel DR. Correcting false discovery rates for their bias toward false positives. Working Paper, University of Ottawa, deposited in uO Research at http://hdlhandlenet/10393/34277. 2016;.

3. Efron B. Size, power and false discovery rates. The Annals of Statistics. 2007; p. 1351–1377. https://doi.org/10.1214/009053606000001460

4. Ploner A, Calza S, Gusnanto A, Pawitan Y. Multidimensional local false discovery rate for microarray studies. Bioinformatics. 2005; 22(5):556–565. https://doi.org/10.1093/bioinformatics/btk013 PMID: 16368770

5. Schwartzman A, Dougherty RF, Taylor JE. Cross-subject comparison of principal diffusion direction maps. Magnetic Resonance in Medicine. 2005; 53(6):1423–1431. https://doi.org/10.1002/mrm.20503 PMID: 15906307

6. Noble WS. How does multiple testing correction work? Nature biotechnology. 2009; 27(12):1135–1137. https://doi.org/10.1038/nbt1209-1135 PMID: 20010596

7. Adkins DE, Åberg K, McClay JL, Bukszár J, Zhao Z, Jia P, et al. Genomewide pharmacogenomic study of metabolic side effects to antipsychotic drugs. Molecular psychiatry. 2011; 16(3):321–332. https://doi.org/10.1038/mp.2010.14 PMID: 20195266

8. Zablocki RW, Schork AJ, Levine RA, Andreassen OA, Dale AM, Thompson WK. Covariate-modulated local false discovery rate for genome-wide association studies. Bioinformatics. 2014; 30(15):2098–2104. https://doi.org/10.1093/bioinformatics/btu145 PMID: 24711653

9. van den Oord EJ. Controlling false discoveries in genetic studies. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 2008; 147(5):637–644. https://doi.org/10.1002/ajmg.b.30650

10. Liao J, Lin Y, Selvanayagam ZE, Shih WJ. A mixture model for estimating the local false discovery rate in DNA microarray analysis. Bioinformatics. 2004; 20(16):2694–2701. https://doi.org/10.1093/bioinformatics/bth310 PMID: 15145810

11. Efron B. Simultaneous inference: When should hypothesis testing problems be combined? The Annals of Applied Statistics. 2008; p. 197–223. https://doi.org/10.1214/07-AOAS141

12. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. https://doi.org/10.1038/nature05911 PMID: 17554300

13. Bickel DR. Minimax-Optimal Strength of Statistical Evidence for a Composite Alternative Hypothesis. International Statistical Review. 2013; 81(2):188–206. https://doi.org/10.1111/insr.12008

14. Mei S, Karimnezhad A, Forest M, Bickel DR, Greenwood C. The performance of a new local false discovery rate method on tests of association between coronary artery disease (CAD) and genome-wide genetic variants. PLoS ONE. 2017; 12:e0185174. https://doi.org/10.1371/journal.pone.0185174 PMID: 28931044

15. Bickel DR. 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. 2004; 3(1):1–20. https://doi.org/10.2202/1544-6115.1043

16. Benjamini Y, Hochberg Y. Multiple hypotheses testing with weights. Scandinavian Journal of Statistics. 1997; 24(3):407–418. https://doi.org/10.1111/1467-9469.00072

17. Genovese CR, Roeder K, Wasserman L. False discovery control with p-value weighting. Biometrika. 2006; 93(3):509–524. https://doi.org/10.1093/biomet/93.3.509

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

19. Sun L, Craiu RV, Paterson AD, Bull SB. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. Genetic epidemiology. 2006; 30(6):519–530. https://doi.org/10.1002/gepi.20164 PMID: 16800000

20. Hu JX, Zhao H, Zhou HH. False discovery rate control with groups. Journal of the American Statistical Association. 2010; 105(491). https://doi.org/10.1198/jasa.2010.tm09329 PMID: 21931466

21. Karimnezhad A, Bickel DR. Incorporating prior knowledge about genetic variants into the analysis of genetic association data: An empirical Bayes approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2018;.

22. Lewis CM. Genetic association studies: design, analysis and interpretation. Briefings in Bioinformatics. 2002; 3(2):146–153. https://doi.org/10.1093/bib/3.2.146 PMID: 12139434

23. Bukszár J, McClay JL, van den Oord EJCG. Estimating the posterior probability that genome-wide association findings are true or false. Bioinformatics. 2009; 25:1807–1813. https://doi.org/10.1093/bioinformatics/btp305 PMID: 19420056

24. Yang Y, Aghababazadeh FA, Bickel DR. Parametric estimation of the local false discovery rate for identifying genetic associations. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2013; 10(1):98–108. https://doi.org/10.1109/TCBB.2012.140

25. Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. Journal of the American Statistical Association. 2004; 99(465):96–104. https://doi.org/10.1198/016214504000000089

26. Walther BA, Moore JL. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography. 2005; 28(6):815–829. https://doi.org/10.1111/j.2005.0906-7590.04112.x

27. Sundgren P, Dong Q, Gomez-Hassan D, Mukherji S, Maly P, Welsh R. Diffusion tensor imaging of the brain: review of clinical applications. Neuroradiology. 2004; 46(5):339–350. https://doi.org/10.1007/s00234-003-1114-x PMID: 15103435

28. Deutsch GK, Dougherty RF, Bammer R, Siok WT, Gabrieli JD, Wandell B. Children's reading performance is correlated with white matter structure measured by diffusion tensor imaging. Cortex. 2005; 41(3):354–363. https://doi.org/10.1016/S0010-9452(08)70272-7 PMID: 15871600

29. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome biology. 2004; 5(10):R80. https://doi.org/10.1186/gb-2004-5-10-r80 PMID: 15461798