Incorporating prior knowledge about genetic variants into the analysis of genetic association data: An empirical Bayes approach

Ali Karimnezhad and David R. Bickel

Abstract—In a genome-wide association study (GWAS), the probability that a single nucleotide polymorphism (SNP) is not associated with a disease is its local false discovery rate (LFDR). The LFDR for each SNP is relative to a reference class of SNPs. For example, the LFDR of an exonic SNP can vary widely depending on whether it is considered relative to the separate reference class of other exonic SNPs or relative to the combined reference class of all SNPs in the data set. As a result, the analysis of the data based on the combined reference class might indicate that a specific exonic SNP is associated with the disease, while using the separate reference class indicates that it is not associated, or vice versa. To address that, we introduce empirical Bayes methods that simultaneously consider a combined reference class and a separate reference class. Our simulation studies indicate that the proposed methods lead to improved performance. The new maximum entropy method achieves that by depending on the separate class when it has enough SNPs for reliable LFDR estimation and depending solely on the combined class otherwise.
We used the new methods to analyze data from a GWAS of 2000 cases and 3000 controls.
R functions implementing the proposed methods are available on CRAN
<https://cran.r-project.org/web/packages/LFDREmpiricalBayes> and Shiny
<https://empiricalbayes.shinyapps.io/lfdrempiricalbayesapp>.

Index Terms—empirical Bayes estimation, local false discovery rate, maximum entropy, minimum relative entropy, reference class problem, robust Bayes action, separate analysis

1 INTRODUCTION

Discovering single nucleotide polymorphisms (SNPs) associated with a specific disease such as coronary artery disease (CAD) has been absorbing attention in recent years. In such a large-scale simultaneous hypothesis testing problem, several thousands of SNPs in a case-control study are tested together. For each SNP, the null hypothesis that the SNP is not associated with the disease is tested against the alternative hypothesis that the SNP is associated with it. Then, if the null hypothesis is rejected, the SNP is considered to be associated with the disease.

A pioneering work in the multiple hypothesis testing scheme by Benjamini and Hochberg [3] introduces the concept of false discovery rate (FDR) and since then many developments have been conducted [17, 18, 35, 36]. The posterior probability that the null hypothesis is true can be used to decide whether or not a specific SNP is associated with the disease. But the posterior probability of the null hypothesis called the local false discovery rate (LFDR) depends on some parameters that are usually unknown [9, 19]. In such cases, the LFDR as a Bayesian posterior probability needs to be estimated. A successful approach in this regard is the empirical Bayes approach of estimating LFDR which replaces estimates of the parameters on which the LFDR depends by their estimated values [16, 18]. Then, SNPs which are associated with the disease can be identified on the basis of values of estimated LFDRs.

LFDR estimation has been performed by different methods in the literature. Allison et al. [2], Pan et al. [31] and Efron [14, 15] consider the estimation of LFDR based on a discrete mixture model. Also Muralidharan [29], Padilla and Bickel [30] and

- *A. Karimnezhad is a postdoctoral fellow at the Ottawa Hospital Research Institute and the Department of Biochemistry, Microbiology, and Immunology, University of Ottawa, Ottawa, ON, Canada, K1H 8M5.*
  E-mail: a.karimnezhad@uottawa.ca
- *D. R. Bickel is with the Ottawa Institute of Systems Biology; Department of Biochemistry, Microbiology and Immunology; Department of Mathematics and Statistics; University of Ottawa, Ottawa, ON, Canada, K1H 8M5.*
  E-mail: dbickel@uottawa.ca

Yang et al. [43] consider the LFDR estimation using the maximum likelihood (ML) approach. Bickel [9] provides a summary of strengths and weakness of four major approaches to multiple hypothesis testing including two error-rate control approaches (family-wise error rate and FDR control) and two posterior probability approaches (classical and empirical Bayes). In addition, Bickel [11] points out that, compared to FDR-controlling methods, LFDR estimation leads to a lower bias.

As a motivating example, the CAD data includes information of 357,468 SNPs of which 10126 SNPs are in non-coding RNA (ncRNA). Now, there are two directions to determine whether a specific ncRNA SNP, rs7326878, is associated with the disease. One direction is to analyze only the ncRNA SNPs together, and an alternative direction is to conduct the analysis over all the available SNPs. While the analysis based on information of all the 357,468 SNPs leads to a low estimate of the corresponding LFDR (0.1563), considering only the ncRNA SNPs yields to a high estimate of the LFDR (0.8940). Analyzing all the SNPs together due to the low estimated LFDR discovers that the SNP rs7326878 is associated with the disease while limiting considering only the ncRNA SNPs together determines that this SNP is not associated with the disease. Thus, there is an uncertainty regarding whether to reject the null hypothesis that the SNP is not associated with the disease. To incorporate such prior knowledge available in the form of biological annotations, we introduce novel approaches of discovering SNPs associated with the disease based on robust Bayes and information-theoretic approaches.

For a single genetic variant such as a SNP, which other variants should be used when estimating the LFDR? They will be the variants in some reference class of which the genetic variant of interest is also a member. All reference classes include the genetic variant of interest. The problem is that the LFDR estimate strongly depends on the class. In the above example, the separate reference class is ncRNA, and the combined reference class is all the SNPs. We propose and compare several candidate solutions of this reference class problem; Aghababazadeh et al. [1] review previous solutions.

In Section 2, we introduce notation and briefly review previous empirical Bayes methods and the use of classical Bayesian decision theory with the estimated posterior distributions. We consider inference based on simultaneously considering the combined reference class and the separate reference class. Since two reference classes lead to two different posterior distributions, special methods are needed. The approach of Section 3 is to pool the two posterior distributions into a single posterior distribution for use with the classical Bayesian decision theory. The approach of Section 4 is to instead apply robust Bayes decision theory without first pooling the posterior distributions. Section 5 reports our simulation results. Results from the CAD data analysis are reported in Section 6. We end up the paper with some conclusions and discussions in Section 7.

2 Previous Empirical Bayes Methods

2.1 Notation

The procedure of discovering SNPs that are associated with the disease is as follows: For an ith SNP, $i = 1, 2, \ldots, N$, the test statistic t_i is used to either accept or reject the null hypothesis that ith SNP is associated with the disease. Let A_i be an indicator such that the null hypothesis is $H_{0i} : A_i = 0$ and the alternative hypothesis is $H_{1i} : A_i = 1$. Under the alternative hypothesis, i.e. $A_i = 1$, the ith SNP is deemed to be associated with (affected by) the underlying disease or treatment. Alternatively under the null hypothesis, i.e. $A_i = 0$, the ith SNP is supposed not to be associated with (affected by) the underlying disease or treatment. To test the hypothesis, a critical region is defined and if the test statistic t_i falls within the critical region, the corresponding null hypothesis is decided to be rejected. A common quantity to measure strength of the ith SNP association with the disease is the *odds ratio* OR_i or its $\log$ transform, i.e., $\theta_i = \log(OR_i)$, which compares the odds between individuals with different genotype or allele. In terms of θ_i, the null hypothesis stating that the ith SNP is not associated with the disease corresponds to $\theta_i = 0$, otherwise $\theta_i \neq 0$. The *OR* is usually estimated using the logistic regression and a regression coefficient β_i corresponding to the ith SNP is estimated by the maximum likelihood estimator $\widehat{\beta}_i$. Then, to test the hypothesis $H_{0i} : \beta_i = 0$ vs $H_{1i} : \beta_i \neq 0$, a Wald test statistic is defined through a function T on $\widehat{\beta}_i$, i.e., $t_i = T(\widehat{\beta}_i) = \frac{\widehat{\beta}_i^2}{\widehat{Var}(\widehat{\beta}_i)}$, where $\widehat{Var}(\widehat{\beta}_i)$ is the standard estimate of the variance of $\widehat{\beta}_i$. This test statistic under the null hypothesis has approximately a chi-square distribution with one degree of freedom [43].

Consider the case that there is some biological information leading to possibility of conducting both separate and combined analyses, as provided in the above ncRNA example. This information automatically defines separate and combined reference classes. We shall refer to the small reference class by S. We also refer to the combined reference class by C. By this definition, it is obvious that $S \subset C$. In correspondence with the separate and combined analyses, let $t_1, t_2, \ldots, t_{N_S}, t_{N_S+1}, \ldots, t_{N_C}$ denote test statistic values in which t_i is a realization of the test statistic T_i having the probability density functions (pdfs)

$g_0(.)$ and $g_{d_{\text{alt}}}(.)$, conditional on the null and non-null hypotheses, respectively. Let $S = \{1, 2, \ldots, N_S\}$, $C = \{1, 2, \ldots, N_C\}$ and $M = \{N_{S+1}, N_{S+2}, \ldots, N_C\}$ be the set of indices. Then, our goal is to test the following hypotheses $H_{0i} : A_i = 0$ vs $H_{1i} : A_i = 1$, $i \in R$, where R is the set of indices of SNPs falling in either the separate or the combined reference class. For instance, in our motivating example S stands for the ncRNA reference class and then, M will refer to all the SNPs in the combined reference class C after excluding the ncRNA SNPs.

2.2 Inference from single posterior distribution

In this subsection, first we consider inference based on a single reference class, either the combined reference class or the separate reference class. Since each reference class leads to a single posterior distribution, classical Bayesian decision theory applies. To provide the prerequisite material, for a given reference class R, suppose $P(A_i = 0) = \pi_{0R}$ and $P(A_i = 1) = 1 - \pi_{0R}$. Thus, the LFDR w.r.t. the reference class R is

$$\psi_{i,R} = \frac{\pi_{0R} g_0(t_i)}{\pi_{0R} g_0(t_i) + (1 - \pi_{0R}) g_{\text{alt}}(t_i)},$$

where t_i is a realization of the test statistic T_i having the pdf $g_0(.)$ conditional on the null hypothesis and pdf $g_{\text{alt}}(.)$ conditional on the alternative hypothesis (Bickel [9], Efron [17]). In practice, g_0 is usually known (it can be pdf of standard normal, student or a chi-square distribution with some degrees of freedom) but π_0 and g_{alt} have to be estimated [30, 43].

The hypothesis indicator A_i conditional on the test statistic t_i follows a Bernoulli distribution with probability of success $1 - \psi_{i,R}$, i.e., $P(A_i = 0|t_i) = \psi_{i,R}$ and $P(A_i = 1|t_i) = 1 - \psi_{i,R}$. We shall refer to this posterior distribution by P_R^i. We refer to an estimated LFDR for ith SNP by $\widehat{\psi}_{i,R}$ and in this case, we replace P_R^i by $\widehat{P}_R^i$. We shall denote estimate of π_{0R} by $\widehat{\pi}_{0R}$.

Let $\delta_i = \delta(t_i)$ be a decision rule based on the test statistic value t_i. This decision rule leads to an estimate of either the hypothesis indicator A_i or log of the odds ratio, i.e., $\theta_i = \log(OR_i)$. For ith SNP consider the following loss functions

$$L_{\text{ZO}}(A_i, \delta_i) = \begin{cases} 0 & \text{if } \delta_i = A_i \in \{0, 1\}, \\ l_I & \text{if } \delta_i = 1, A_i = 0, \\ l_{II} & \text{if } \delta_i = 0, A_i = 1, \end{cases} \tag{1}$$

$$L_{\text{SE}}(A_i, \delta_i) = (\delta_i - A_i)^2, \quad -\infty < \delta_i < +\infty, \; A_i \in \{0, 1\}, \tag{2}$$

and

$$L_{\text{OR}}(\theta_i, \delta_i) = (\delta_i - \theta_i)^2, \quad -\infty < \theta_i, \delta_i < +\infty. \tag{3}$$

The above loss functions measure inaccuracy of the estimation tasks of either A_i or θ_i. The term ZO in L_{ZO} is abbreviation of "zero- one", and represents that both the estimated parameter A_i and an estimator δ_i take 0 or 1 values. The L_{ZO} loss is useful in hypothesis testing terminology, in which $l_I, l_{II} (> 0)$ are the loss due to making type I and type II errors, respectively. Both L_{SE} and L_{OR} are squared error loss (SEL) functions in the sense that both measure squared distance between the desired parameter (either θ_i or A_i) and an estimator δ_i. The term OR in L_{OR} is abbreviation of "odds ratio" and the loss function L_{OR} in (3) is measuring penalties in estimating the ith log *OR*, i.e., θ_i by the estimator δ_i. It will be apparent that results of estimating A_i under the loss function L_{SE} in (2) can be derived from results of estimating θ_i under the function L_{OR} in (3). To take readers in track, we will concentrate on estimating θ_i.

Let $\rho\left(\widehat{P}_R^i, \delta_i\right) = E\left[L(\eta_i, \delta_i)|T_i = t_i\right]$ denote the posterior risk of $\delta_i \in D$ associated with the prior π_R where D is a set of possible actions and $E[.]$ stands for expectation w.r.t. the conditional density of $\eta_i|T_i = t_i$, $\eta_i \in \{A_i, \theta_i\}$. It is well-known that a Bayes estimate w.r.t. a given prior under a specific loss function would be obtained by minimizing the posterior loss

$$\rho\left(\widehat{P}_R^i, \delta_i\right) = L_{\text{ZO}}(0, \delta_i)\widehat{\psi}_{i,R} + L_{\text{ZO}}(1, \delta_i)\left(1 - \widehat{\psi}_{i,R}\right) \tag{4}$$

w.r.t. δ_i. Taking this fact in mind, it can be verified that the Bayes estimate of each hypothesis indicator A_i, $i \in I_R$, under the L_{ZO} loss (1) is given by

$$\delta_{\text{ZO}}^{\pi_R}(t_i) = \begin{cases} 1 & \text{if } \widehat{\psi}_{i,R} \leq \frac{l_{II}}{l_I + l_{II}}, \\ 0 & \text{if } \widehat{\psi}_{i,R} > \frac{l_{II}}{l_I + l_{II}}. \end{cases} \tag{5}$$

In the same procedure, the posterior risk under the SEL function (3) corresponding to the ith hypothesis can be written as

$$\rho\left(\widehat{P}_R^i, \delta_i\right) = \delta_i^2 \widehat{\psi}_{i,R} + (\delta_i - \widehat{\theta}_i)^2\left(1 - \widehat{\psi}_{i,R}\right) \tag{6}$$

which leads to the following Bayes estimator

$$\delta_{\text{OR}}^{\pi_R}(t_i) = \widehat{\theta}_i(1 - \widehat{\psi}_{i,R}). \tag{7}$$

It is easy to verify that the posterior risk under the SEL function (2) is

$$\rho\left(\widehat{P}_R^i, \delta_i\right) = \delta_i^2 \widehat{\psi}_{i,R} + (\delta_i - 1)^2\left(1 - \widehat{\psi}_{i,R}\right),$$

which is in fact the same the posterior risk in (6) when considering $\widehat{\theta}_i = 1$. Thus the Bayes estimate of the hypothesis indicator A_i under the SEL function (2) is the same as the Bayes estimator in (7) except that we replace $\widehat{\theta}_i$ by 1.

3 Inference via Pooling Posterior Distributions

The following information-theoretic methods pool posterior distributions corresponding to different reference classes into a single distribution for use with the Bayes rule described earlier.

3.1 A maximum entropy method of pooling distributions

To discover associated SNPs, we propose a new maximum entropy (ME) approach, which compares two likelihood functions constructed based on two given models. In practice, there is a lack of knowledge that specifies whether the separate reference class S or the combined reference class C should be used, to get more reliable estimates of LFDRs for the SNPs. This approach provides a *selected reference class* using both separate and combined analyses in favor of a given data set. Then, giving credit to the selected reference class, an estimate of LFDR is computed for each SNP. We refer to the ME estimate of LFDR by *BME*, the Bayes estimator relevant to the selected reference class.

We consider the density under the null and alternative hypotheses are $g_0(.)$ and $g_{d_{\text{alt}R}}(.)$, respectively. Here $g_0(.)$ refers to the central chi-square density with 1 degree of freedom and $g_{d_{\text{alt}R}}(.)$ refers to a non-central chi-square density with 1 degree of freedom and non-centrality parameter $d_{\text{alt}R}$. The index R emphasizes that the non-centrality parameter can depend on the reference class R. Further, we denote an estimate of $d_{\text{alt}R}$ by $\hat{d}_{\text{alt}R}$.

The procedure is as follows: for all SNPs associated with the separate reference class S, consider the likelihood function

$$L(\tau) = \prod_{i \in S} \left(\pi_0 g_0(t_i) + (1-\pi_0) g_{d_{\text{alt}}}(t_i)\right),$$

where $\tau = (\pi_0, d_{\text{alt}})$. Based on a model checking approach and following Bickel [10], define the following likelihood set

$$L_S = \left\{ \tau : \frac{L(\tau)}{L(\widehat{\tau}_S)} \geq \frac{1}{2^a}, \tau \in [0,1] \times [d_1, d_2] \right\}, \quad (8)$$

where a is a predetermined threshold, d_1 and d_2 are prespecified limits of the non-centrality parameter d_{alt}, and $\widehat{\tau}_S = (\widehat{\pi}_{0S}, \widehat{d}_{\text{alt}S})$ is obtained through maximizing the joint mixture density $\prod_{i=1}^{N_S} (\pi_{0S} g_0(t_i) + (1-\pi_{0S}) g_{d_{\text{alt}}}(t_i))$ over (π_0, d_{alt}).

Different positive values can be chosen for a indicating grades of evidence against the separate reference class S and in favor of its alternative. We choose $a = 3$, considering strong evidence against the separate reference class and in favor of its alternative, see Bickel [10] for more details. We choose $d_1 = 0.1$ and $d_2 = 50$ to ensure that our procedure considers a rich interval for estimating the parameter $d_{\text{alt}R}$.

Let ψ_i be the LFDR for ith SNP computed based on values of π_0 and d_{alt} belonging to the likelihood set (8), and suppose P^i is the corresponding conditional distribution for the indicator variable A_i. Since each pair (π_0, d_{alt}) in the likelihood set leads to an estimate of LFDR ψ_i, changing values of (π_0, d_{alt}) leads to an interval of LFDR, say $[\psi_{i,S}^L, \psi_{i,S,}^U]$. Now, for each ψ_i, consider the following relative entropy function

$$D\left(P^i || \widehat{P}_C^i\right) = \psi_i \log\left(\frac{\psi_i}{\widehat{\psi}_{i,C}}\right) + (1-\psi_i)\log\left(\frac{1-\psi_i}{1-\widehat{\psi}_{i,C}}\right). \quad (9)$$

Then $\widehat{\psi}_{i,\text{ME}}$, the ME estimate, is the value of ψ_i that minimizes $D\left(P^i || \widehat{P}_C^i\right)$ over the interval $[\psi_{i,S}^L, \psi_{i,S,}^U]$. Once it is computed, we calculate estimates of the parameters A_i and θ_i using the equations (5) and (7), respectively.

To clarify the above procedure, suppose $\widehat{\tau}_C = (\widehat{\pi}_{0C}, \widehat{d}_{\text{alt}C}) \in L_S$. Then, it is obvious that $\widehat{\psi}_{i,\text{ME}} = \widehat{\psi}_{i,C}$ minimizes the relative entropy $D\left(P^i || \widehat{P}_C^i\right)$ in (9). Hence this procedure selects the combined reference class as the appropriate reference class. In fact, in this case, the interval $[\psi_{i,S}^L, \psi_{i,S,}^U]$ is too wide and the separate reference class does not have enough SNPs for reliable estimates of LFDRs. But if $\widehat{\tau}_C \notin L_S$, then in correspondence with any P^i that minimizes the relative entropy in (9), a reference class will be determined by the mentioned procedure. In this case the interval $[\psi_{i,S}^L, \psi_{i,S,}^U]$ is sufficiently narrow and the separate reference class has enough SNPs for reliable estimates of LFDRs. If $\widehat{\psi}_{i,C} \in [\psi_{i,S}^L, \psi_{i,S}^U]$, then $\widehat{\psi}_{i,\text{ME}} = \widehat{\psi}_{i,C}$. Otherwise, if $\widehat{\psi}_{i,C} < \psi_{i,S}^L$, then $\widehat{\psi}_{i,\text{ME}} = \psi_{i,S}^L$ and if $\widehat{\psi}_{i,C} > \psi_{i,S}^U$, then $\widehat{\psi}_{i,\text{ME}} = \psi_{i,S}^U$.

The ME estimate of LFDR has the following interesting properties:

- if a tends to 0, the likelihood set contains only one point which is $\widehat{\tau_S}$ and thus, $\widehat{\psi}_{i,\text{ME}} = \widehat{\psi}_{i,S}$;
- if a tends to ∞, the likelihood set is equal to the whole area $[0,1] \times [0, +\infty)$ and thus, $\widehat{\psi}_{i,\text{ME}} = \widehat{\psi}_{i,C}$;
- for any other value of a, the interval $[\psi_{i,S}^L, \psi_{i,S}^U]$ will be constructed. If $\widehat{\psi}_{i,C} \in [\psi_{i,S}^L, \psi_{i,S}^U]$, then $\widehat{\psi}_{i,\text{ME}} = \widehat{\psi}_{i,C}$. Otherwise, if $\widehat{\psi}_{i,C} < \psi_{i,S}^L$, then $\widehat{\psi}_{i,\text{ME}} = \psi_{i,S}^L$ and if $\widehat{\psi}_{i,C} > \psi_{i,S}^U$, then $\widehat{\psi}_{i,\text{ME}} = \psi_{i,S}^U$.

3.2 A game-theoretic method of pooling distributions

The problem of deriving an optimal rule can be considered as a game-theoretic approach introduced by Bickel [8]. The theory is based on three players that establish a set of density functions to be combined and

the result is a linear combination of distributions with optimized weights, the values of which are based on information theory.

Following Corollary 2 of Bickel [8], we compute the wights $w_{i,S}$ and $w_{i,C}$ associated with the separate and combined reference classes, respectively, and derive the combined game-theoretic (GT) estimate of LFDR as $\widehat{\psi}_{i,\mathrm{GT}} = w_{i,S}\widehat{\psi}_{i,S} + w_{i,C}\widehat{\psi}_{i,C}$. Then, for an ith SNP, we compute estimates of the parameters A_i and θ_i using the equations (5) and (7), respectively. To do so, we replace $\widehat{\psi}_{i,R}$ by $\widehat{\psi}_{i,\mathrm{GT}}$.

Below we provide an idea on how to compute the above wights $w_{i,S}$ and $w_{i,C}$. Let $w_1 = w_{i,S}$ and $w_2 = w_{i,C}$. Followoing Corollary 2 of Bickel [8], the pairs (w_1, w_2) are computed by maximizing

$$\sum_{i=1}^{2} w_i \sum_{j=1}^{N_S}\sum_{k=0}^{1} \widehat{P}_i(A_j = k|t_j) \log \frac{\widehat{P}_i(A_j = k|t_j)}{P^{(w_1,w_2)}(A_j = k|t_j)}$$

over $\{(w_1, w_2) \in [0,1] \times [0,1] : \ w_1 + w_2 = 1\}$. In the above expression,

$$P^{(w_1,w_2)}(A_j = k|t_j) = w_1 P_1(A_j = k|t_j) + w_2 P_2(A_j = k|t_j),$$

where $P_i(A_j = k|t_j)$ represents the LFDR for SNP j computed based on either separate reference class ($i = 1$) or combined reference class ($i = 2$). $\widehat{P}_i(A_j = k|t_j)$ stands for an estimated version of $P_i(A_j = k|t_j)$.

4 Inference via robust Bayes Methods

As observed earlier, the Bayes solution depends on the choice of prior π_R which stems either from a chosen reference class R or from a distribution pooled over all reference classes. Robust Bayes analysis deals with the problem of uncertainty propagation in terms of the distribution and while giving credit to all models provided in a given distribution, is aimed at global prevention against bad choices. Excellent discussions are provided in Berger [4, 5] and [6]. Also, Karimnezhad and Parsian [25, 26] and Karimnezhad et al. [27] provide developments in different contexts.

In this section, we provide novel approaches to discover SNPs associated with a specific disease. To do so, we assume the reference class R varies over the set $\{S, C\}$. Obviously, for a chosen reference class R, the corresponding prior π_R varies over the set of priors $\Pi_R = \{\pi_S, \pi_C\}$. We shall refer to the robust Bayes analyses by *average analysis*.

Recalling our motivating example, we observed that for a particular SNP there might be more than one reference class leading to possibly different LFDR estimates. Comparing the resulting LFDR estimates with a pre-determined threshold such as 0.2 can lead us to an uncertainty regarding whether that SNP is associated with the disease or not. Since reference classes lead to different posterior distributions, we analyze behaviour of the corresponding posterior risks to be able to make a sensible decision.

In the rest of this section, we follow decision-theoretic approaches that analyze behaviour of posterior risks associated with each reference class to estimate the hypothesis indicator A_i or the parameter $\theta_i = \log(OR_i)$. This approach in subsection 4.1 involves choosing a pre-determined caution parameter κ and deriving an optimal action minimizing a combination of minimum and maximum of posterior risks when reference classes are allowed to be either separate or combined classes. The approach in subsection 4.2 is slightly different. It introduces an alternative estimator to the Bayes estimator which plays a key role in making inference from single posterior distribution discussed earlier in subsection 2.2. Instead of minimizing the posterior risk, the estimate that we introduce in subsection 4.2 uses the regret of choosing an estimate instead of the traditional Bayes estimate. It leads to an optimal estimate by minimizing that regret when reference classes are allowed to be either separate or combined classes.

4.1 Caution-type estimators

Decision-theoretic rules can be chosen by caution or without any caution given a viable set of prior distributions [8, 22, 24]. In this regard, we apply an extended version of the criterion introduced by Hurwicz [22] and followed later by Jaffray [24] and Bickel [10] for the hypothesis testing problem. For an ith hypothesis indicator, this criterion specifies an optimal action satisfying

$$\delta^{\kappa}_{\mathrm{opt}}(t_i) = \arg \inf_{\delta_i \in D} \{\kappa \max_{R\in\{S,C\}} \rho(\widehat{P}^i_R, \delta_i) + (1-\kappa) \min_{R\in\{S,C\}} \rho(\widehat{P}^i_R, \delta_i)\}, \quad (10)$$

where $\rho(\widehat{P}^i_R, \delta_i)$ is the posterior risk of a decision δ_i w.r.t. the prior π_R, and $\kappa \in [0,1]$ is the parameter encoding the caution.

Define $\underline{\psi_i} = \min_{R\in\{S,C\}} \widehat{\psi}_{i,R}$ and $\overline{\psi_i} = \max_{R\in\{S,C\}} \widehat{\psi}_{i,R}$. The caution-type estimate of each hypothesis indicator A_i, $i \in R$, under the L_{ZO} loss function (1) is derived by

$$\delta^{\mathrm{ZO},\kappa}_{\mathrm{opt}}(t_i) = \begin{cases} 1 & \text{if } l_{II}(1 - \underline{\psi_i} - \overline{\psi_i}) \geq h_i(\kappa), \\ 0 & \text{if } l_{II}(1 - \underline{\psi_i} - \overline{\psi_i}) < h_i(\kappa), \end{cases} \quad (11)$$

where $h_i(\kappa) = (l_I - l_{II})(\kappa\overline{\psi_i} + (1-\kappa)\underline{\psi_i})$. For a proof, see Appendix A.

Caution-type actions for some specific values of κ are interesting. The least cautious attitude ($\kappa = 0$), might be referred to as conditional gamma minimin, and the most cautious attitude ($\kappa = 1$), corresponds to the conditional gamma minimax strategy [7, 39].

Another caution-type action with caution parameter $\kappa = 0.5$, which in fact provides a balance between the conditional gamma minimax and conditional gamma minimin can be considered. We refer to these three interesting cases by CGM0, CGM1 and CGM0.5, respectively.

Now, considering the SEL loss function (3), after some algebraic manipulations it can be proved that the CGM0, CGM1 and CGM0.5 estimates of θ_i, $i \in R$, are respectively given by

$$\delta_{\text{opt}}^{\text{SEL},0}(t_i) = \widehat{\theta}_i \left(1 - \frac{1}{2}\left(\underline{\psi_i} + \overline{\psi_i}\right)\right), \tag{12}$$

$$\delta_{\text{opt}}^{\text{SEL},1}(t_i) = \begin{cases} \widehat{\theta}_i(1-\overline{\psi_i}) & \text{if } \underline{\psi_i} \le \overline{\psi_i} \le \frac{1}{2}, \\ \frac{1}{2}\widehat{\theta}_i & \text{if } \underline{\psi_i} < \frac{1}{2} < \overline{\psi_i}, \\ \widehat{\theta}_i(1-\underline{\psi_i}) & \text{if } \frac{1}{2} < \underline{\psi_i} \le \overline{\psi_i}, \end{cases} \tag{13}$$

and

$$\delta_{\text{opt}}^{\text{SEL},0.5}(t_i) = \begin{cases} \widehat{\theta}_i(1-\underline{\psi_i}) & \text{if } \underline{\psi_i} \le \overline{\psi_i} \le \frac{1}{2}, \\ \widehat{\theta}_i\left(1 - \underline{\psi_i} I_i^{[0,1]} - \overline{\psi_i} I_i^{(1,2]}\right) & \text{if } \underline{\psi_i} < \frac{1}{2} < \overline{\psi_i}, \\ \widehat{\theta}_i(1-\overline{\psi_i}) & \text{if } \frac{1}{2} < \underline{\psi_i} \le \overline{\psi_i}, \end{cases} \tag{14}$$

where for subset U of the real line,

$$I_i^U = \begin{cases} 0 & \text{if } \underline{\psi_i} + \overline{\psi_i} \in U, \\ 1 & \text{if } \underline{\psi_i} + \overline{\psi_i} \notin U. \end{cases}$$

For a proof behind the equations (12)-(14) see Appendix B.

4.2 Posterior regret gamma minimax estimator

Another common approach to overcome with the prior uncertainty in the Bayesian framework is called posterior regret gamma minimax (PRGM) approach which has been used and appreciated for a very long time. The context of conditional Gamma minimax regret rules are developed by Zen et al. [44], and excellent developments can be found in Berger [5], Insua et al. [23] and Berger et al. [6].

Suppose the realization t_i is an observation of a random variable T_i and $\delta_i^{\pi_R} = \delta^{\pi_R}(t_i)$ is the Bayes rule w.r.t. the prior π_R, $R \in \{S, C\}$. Each reference class corresponds to a different prior distribution and thus to a different posterior distribution. The posterior regret of a rule δ_i is defined by $r(\delta_i, \delta_i^{\pi_R}) = \rho(\widehat{P}_R^i, \delta_i) - \rho(\widehat{P}_R^i, \delta_i^{\pi_R})$. Informally, for each fixed i, $r(\delta_i, \delta_i^{\pi_R})$ measures the loss of optimality due to choosing the decision δ_i instead of the Bayes rule $\delta_i^{\pi_R}$. We say $\delta_{\text{PRGM}}(t_i)$ is a PRGM rule if it minimizes $\max_{R\in\{S,C\}} r(\delta_i, \delta_i^{\pi_R})$, i.e.,

$$\delta_{\text{PRGM}}(t_i) = \arg \inf_{\delta_i \in D} \max_{R\in\{S,C\}} r(\delta_i, \delta_i^{\pi_R}).$$

It can be verified that the PRGM estimate of each θ_i, $i \in R$, under the SEL loss function (3) is equal to the caution-type action $\delta_{\text{Opt}}^{\text{SEL},0}(t_i)$ in (12). Once again, it would be easy to verify that caution-type estimates of the hypothesis indicator A_i under the SEL function (2) will be obtained by replacing $\widehat{\theta}_i$ in (11)-(12) by 1.

5 Simulation Studies

5.1 Simulation settings

To illustrate behavior of the proposed estimators of LFDR, we conduct a simulation study as summarized in Algorithm 1. In our simulation study we consider one separate reference class (S) and one combined reference class (C) in which S consists of 2000 SNPs with some proportion of disease affection $\pi_{0S} \in \{0, 0.1, \ldots, 1\}$ and C consists of 4000 SNPs ($S \subset C$). For the 2000 SNPs in the complement of separate reference class, denoted by M, we suppose proportion of disease affection is $\pi_{0M} \in \{0, 1\}$. Using the fact the log of *OR* follows a normal distribution, we generate a sequence of z_i values which under null hypothesis that there is no association between SNPs and a specific disease, follow a normal distribution with mean 0 and variance $\sigma^2 = 0.02$, and under the alternative hypothesis follow normal distribution with mean $\log(1.25)$ and variance $\sigma^2 = 0.02$. These means and variances are unknown for the purpose of estimation; they are only used to simulate the data. We then transform the z_i values to chi-square values through the transformation $t_i = (\frac{z_i}{\sigma})^2$. By this transformation, under the null hypothesis t_i follows a central chi-square distribution with one degree of freedom and under the alternative hypothesis it follows a non-central chi-square distribution with one degree of freedom and non-centrality parameter $d_{\text{alt}S} = d_{\text{alt}M}\left(\frac{\log(1.25)}{\sigma}\right)^2$. Once the test statistics are generated, we apply the ML approach to estimate the corresponding LFDRs based on the methods developed in this paper. Finally, we define average of risks (AMSE and AR_4 in Step 9 of the Algorithm 1) to measure performance of the methods.

5.2 Simulation results

We carried out different simulations with different parameters as shown by Figures 1-4. From the results we observed that if there is a significant difference between π_{0S} and π_{0C}, there is a difference between performance of the resulting separate and combined analyses. For example look at the results associated with the point $\pi_{0S} = 1$ in Figures 1 and 3 (or the point $\pi_{0S} = 0$ in Figures 2 and 4) for which there is a 0.50 difference between and π_{0S} and π_{0C}.

Algorithm 1 Summary of simulation methodology.

1) Take $j = 1$.
2) Generate $z_1, z_2, \ldots, z_{N_{0S}}, z_{N_{0S}+1}, \ldots, z_{N_S}$ such that for $i = 1, 2, \ldots, N_{0S}$, $z_i \sim N(\log(1.25), \sigma^2)$ and for $i = N_{0S+1}, N_{0S+2}, \ldots, N_S$, $z_i \sim N(0, \sigma^2)$, where $\sigma^2 = 0.02$, $N_{0S} = 0, 200, \ldots, 2000$ and $N_S = 2000$. By this setting, a separate reference class, say S, with $\pi_{0S} = 0, 0.1, \ldots, 1$ is constructed.
3) Generate $z_{N_S+1}, z_{N_S+2}, \ldots, z_{N_{0M}}, z_{N_{0M}+1}, \ldots, z_{N_C}$ such that for $i = N_S+1, N_S+2, \ldots, N_{0M}$, $z_i \sim N(\log(1.25), \sigma^2)$ and for $i = N_{0M+1}, N_{0M+2}, \ldots, N_C$, $z_i \sim N(0, \sigma^2)$, where $\sigma^2 = 0.02$, $N_{0M} = 2000, 4000$ and $N_C = 4000$. This way, the reference class M with $\pi_{0M} = 0, 1$ is constructed. Obviously, the combined reference class C is the union of S and M.
4) Compute $t_i = (\frac{z_i}{\sigma})^2$ to construct the chi-square test statistics.
5) Estimate the corresponding LFDRs by $\widehat{\psi}_{i,R}$, $R \in \{S, C\}$.
6) Using the estimated LFDR computed in Step 5, compute $\delta_{\text{ZO}}^{\pi_R}(t_i)$ with $R \in \{S, C\}$, $\delta_{\text{opt}}^{\text{ZO},\kappa}(t_i)$ and $\delta_{\text{opt}}^{\text{SEL},\kappa}(t_i)$ with $\kappa = 0, 0.5, 1$, the game-theoretic and the ME LFDR estimators. Replace $\widehat{\theta}_i$ by corresponding z_i generated in Steps 2 and 3, wherever needed.
7) Compute the losses $L_{\text{ZO}}^j(A_i, \delta_i)$ and $L_{\text{OR}}^j(\theta_i, \delta_i)$ introduced in (1) and (3), where δ_i is any of the applicable estimators computed in Step 6. For the L_{ZO} loss consider $l_I = 4$ and $l_{II} = 1$.
8) Increase j by 1 and repeat Steps 2 to 7 for $N = 1000$ times. Compute

$$\text{R}_{4,i} = \frac{1}{N}\sum_{j=1}^{N} L_{\text{ZO}}^j(A_i, \delta_i),$$

$$\text{MSE}_i = \frac{1}{N}\sum_{j=1}^{N} L_{\text{OR}}^j(\nu_i, \delta_i),$$

where the index 4 in $\text{R}_{4,i}$ refers to the choice $l_I = 4$.
9) For each of the proposed methods, compute averages of $\text{R}_{4,i}$ and MSE_i over all SNPs in the separate reference class S, i.e.,

$$\text{AR}_4 = \frac{1}{N_S}\sum_{i=1}^{N_S} \text{R}_{4,i}, \qquad \text{AMSE} = \frac{1}{N_S}\sum_{i=1}^{N_S} \text{MSE}_i$$

In fact, from Figures 1 and 3 we observe that when $\pi_{0S} \geq 0.4$, the separate analysis outperforms the combined analysis and when $\pi_{0S} < 0.4$, the combined analysis outperforms the separate analysis. The converse behavior observed in Figures 2 and 4. But, if there is no such significant difference, making a decision based on only the separate analysis or the combined analysis could be a challenge. From the Figures 1-4 we observe that performance of the proposed estimators for all values of π_{0S} in different settings is satisfactory. They lead to a decrease in AMSE or AR_4 values. We observe that it is not possible to claim one of the methods always performs better the other methods. However, based on different values of π_{0S} and π_{0M}, Figures 5-8 order the top three estimators of the effect size $\theta_i = \log(OR_i)$ and the hypothesis indicator A_i in terms of their AMSE and AR_4 defined in Step 9 of Algorithm 1.

6 Coronary Artery Disease Data Analysis

To illustrate behavior of the LFDR estimates with incorporated information, we analyze the CAD data from [40][1]. The data include 500,568 SNPs genotyped for 2000 cases and 3,000 combined controls. After the quality control filters performed, 357,468 SNPs with minor allele frequencies greater then 0.05 are retained on 22 autosomal chromosomes and 1926 cases and 2938 controls individuals. Our interest concentrates on incorporating biological information in identifying the SNPs that are associated with the disease.

Functional annotation was performed and different categories were assigned to the SNPs as reference classes using the ANNOVAR software [38]. Figure 9 shows SNPs distribution regarding different reference classes. It is observed that some SNPs are assigned to more than one reference class. For example, of 10126 ncRNA SNPs, 692 found to be exonic. Now, if one is interested in analyzing exonic SNPs, there are different reference classes to consider. To estimate the corresponding LFDRs, a separate analysis would suggest using information of the exonic SNPs while a combined analysis would suggest using information of either ncRNA or all of the SNPs. For these 692 SNPs, we treat the exonic and the ncRNA reference classes as separate and combined reference classes.

Following the ML approach in LFDR estimation, we computed values of test statistic t_i which under the null hypothesis that there is no association between SNPs and CAD disease follow a central chi-square distribution with one degree of freedom. The test statistics under the alternative hypothesis follow a non-central chi-square distribution with one degree of freedom and non-centrality parameter $d_{\text{alt}R}$. Considering S and C for the reference classes including

1. www.wtccc.org.uk

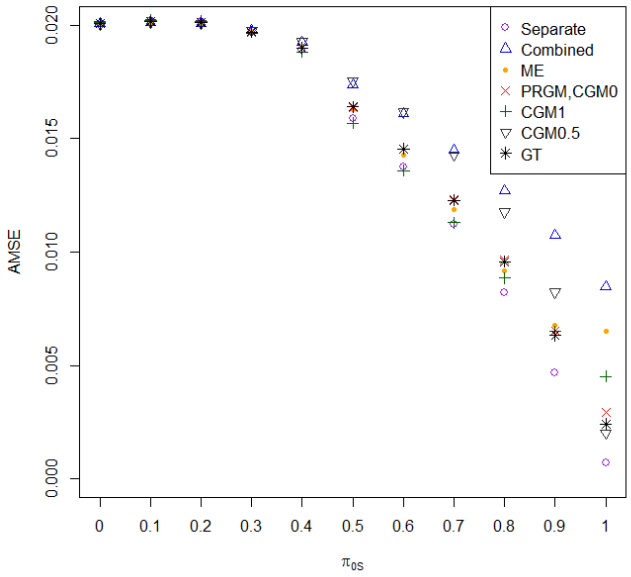

Fig. 1. Plots of AMSE when $\pi_{0M} = 0$ for different values of π_{0S}.

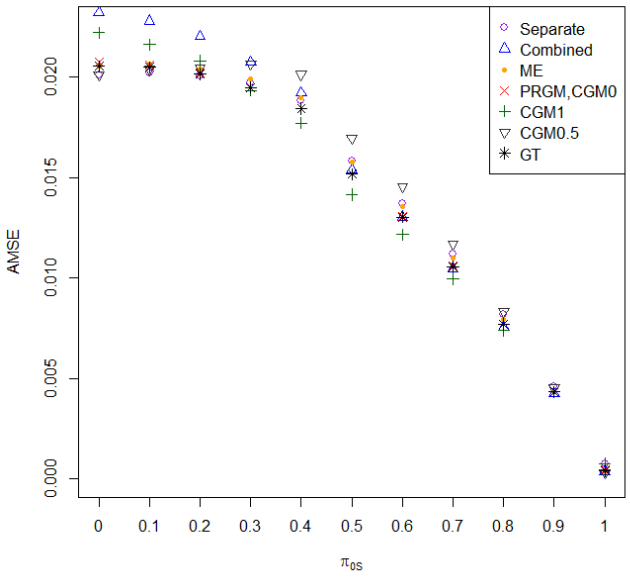

Fig. 2. Plots of AMSE when $\pi_{0M} = 1$ for different values of π_{0S}.

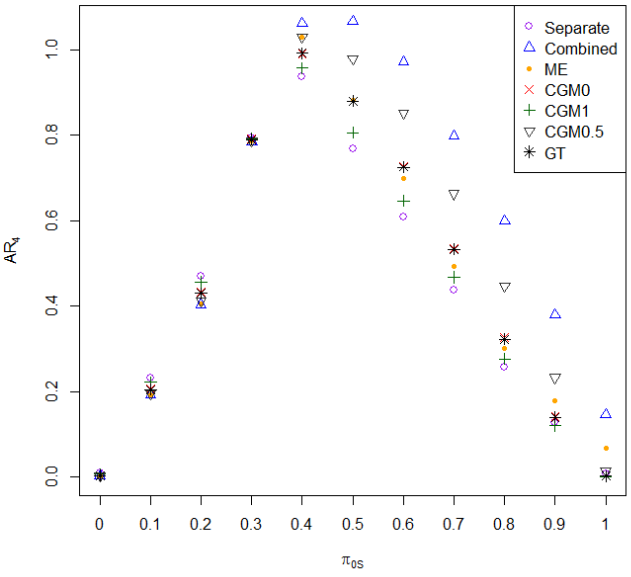

Fig. 3. Plots of AR_4 when $\pi_{0M} = 0$ for different values of π_{0S}.

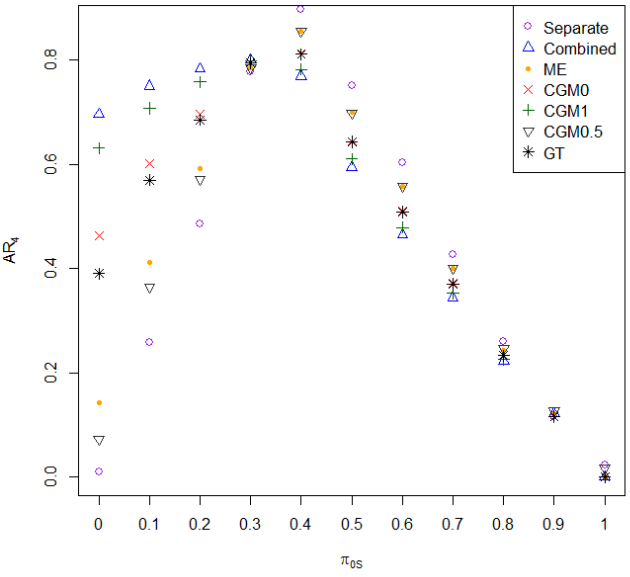

Fig. 4. Plots of AR_4 when $\pi_{0M} = 1$ for different values of π_{0S}.

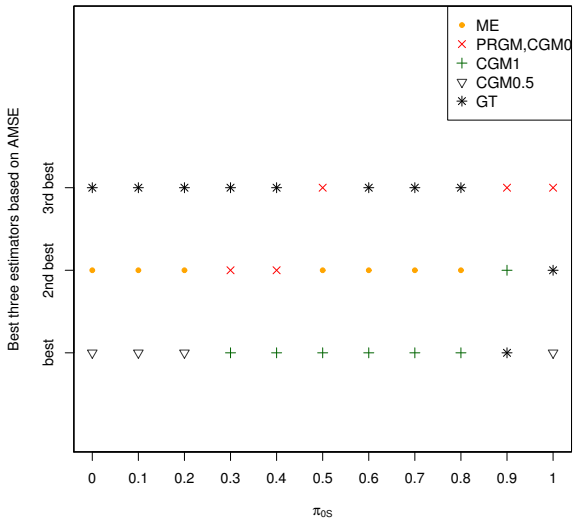

Fig. 5. Best three estimators of the effect size $\theta_i = \log(\text{OR}_i)$ when $\pi_{0M} = 0$.

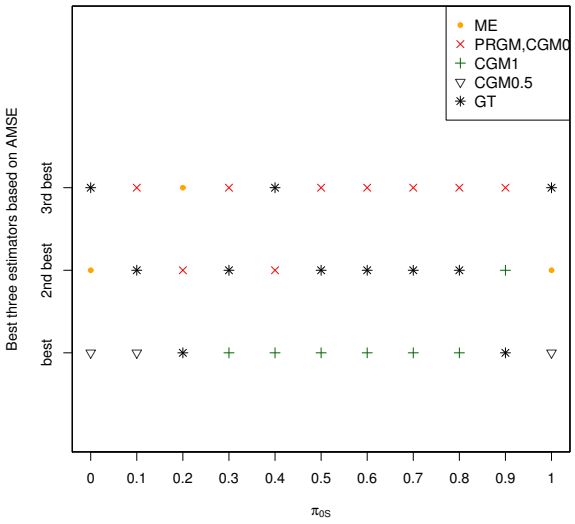

Fig. 6. Best three estimators of the effect size $\theta_i = \log(\text{OR}_i)$ when $\pi_{0M} = 1$.

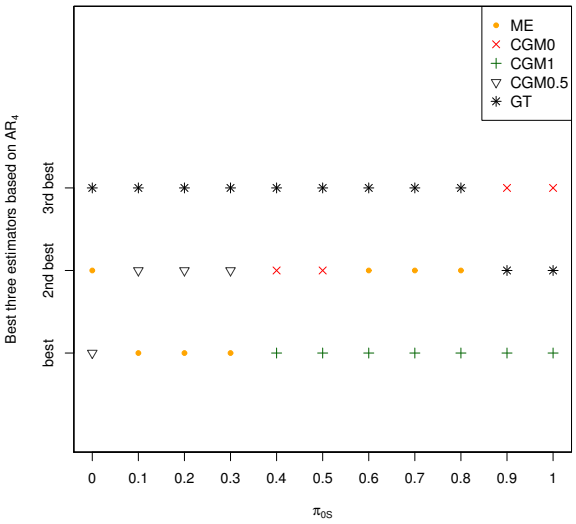

Fig. 7. Best three estimators of the hypothesis indicator A_i when $\pi_{0M} = 0$.

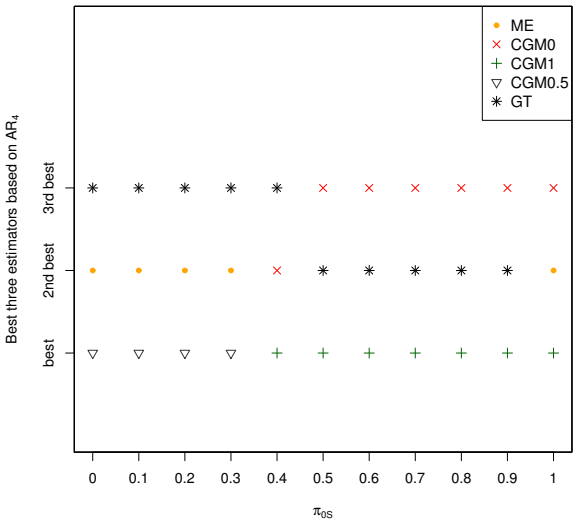

Fig. 8. Best three estimators of the hypothesis indicator A_i when $\pi_{0M} = 1$.

the exonic and ncRNA SNPs respectively, we get $\widehat{\pi}_{0S} = 0.9759$, $\widehat{d}_{\text{alt}S} = 12.2394$, $\widehat{\pi}_{0C} = 0.9978$ and $\widehat{d}_{\text{alt}C} = 33.8456$. Figure 10 provides estimated LFDR values based on the different methods. The LFDR estimates w.r.t. exonic SNPs fall on the horizontal axis and LFDR estimates of the same SNPs regarding the different approaches are shown on the vertical axis.

Significant difference and discrepancies in the LFDRs estimated values (and thus in determining associated SNPs) is realized from Figure 10. Considering the 20 percent threshold, we observe the separate analysis leads to identifying more SNPs associated with the CAD disease than the combined analysis. For example, the separate analysis leads to identifying two SNPs with estimates of LFDR close to 0.15 (rs7186668, rs9926237) while the combined and average analyses, and the BME suggest that these SNPs are not associated with the CAD disease.

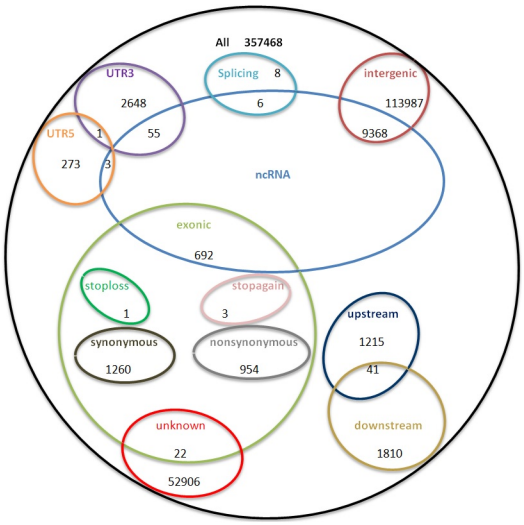

Fig. 9. SNPs distribution based on the functional annotation.

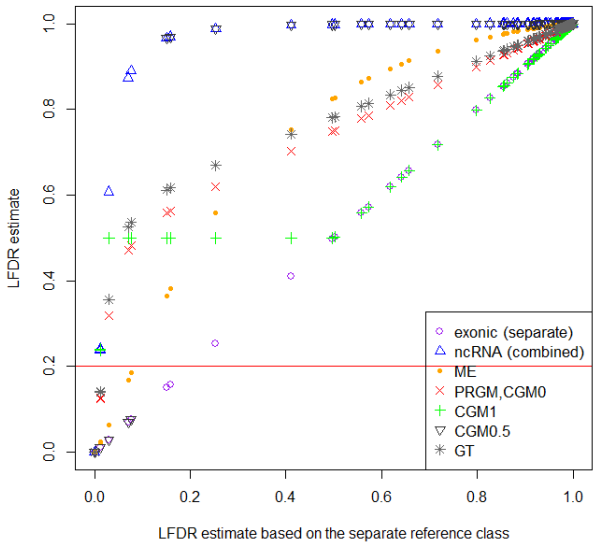

Fig. 10. Estimated LFDRs for 692 exonic SNPs. The LFDR estimates w.r.t. exonic SNPs fall on the horizontal axis and LFDR estimates of the same SNPs regarding the different approaches are shown on the vertical axis.

The proposed methods have also been applied to another GWAS data set [28].

7 Conclusion and Discussion

We conducted some simulation studies with different settings and measured performance of the estimates by the average of risks (AMSE and AR_4 in Step 9 of the Algorithm 1) and observed that the proposed methods lead to improved performance.

We observe that the new ME method is the only method considered that is based on comparing the likelihood functions and takes into account the reliability of the separate reference class. We provide examples in which the new ME method depends on the separate reference class when it has enough SNPs for reliable estimation relative to the reliability of the combined reference class and depends on the combined class otherwise.

For an example of the case when a separate reference class has enough SNPs, see behavior of the ME estimator at the point $\pi_{0S} = 0$ in Figures 2 and 4 for which $\pi_{0C} = 0.5$. The AMSE and AR_4 of the ME estimates are very close to those of the estimated values based on the separate reference class, rather than the combined reference class. This, as expected, roots from a bias of zero ($E[\widehat{\pi}_{0S} - \pi_{0S}]$) in estimating the LFDRs based on the separate reference class and a bias of 0.5 ($E[\widehat{\pi}_{0C} - \pi_{0S}]$) when using the combined reference class ($E[\widehat{\pi}_{0C}] = \pi_{0C} = \pi_{0S} + 0.5$). This leads to an increase in the MSE_i in Step 8 of Algorithm 1 which can be expressed as the sum of variance and the bias squared. Thus, at this point, 2000 SNPs in the separate reference class are enough to get reliable estimates and the ME estimate gives more weight to the separate reference class.

For an example of the case when 2000 SNPs in the separate reference are not enough for deriving reliable estimates, relative to the reliability of the combined reference class, see behavior of the ME estimator at the point $\pi_{0S} = 1$ in Figures 2 and 4 for which $\pi_{0C} = 1$ and the ME estimator chooses the combined reference class for a reliable estimation.

Among our faster and simpler estimation methods that consider separate and combined reference classes without depending on the reliability of estimation, the GT estimator performs very well in the sense that it is one of the best three estimators in Figures 5-8 for almost all values of π_{0S} and π_M.

In estimating θ_i, the GT estimator appears in all the corresponding 22 positions in Figures 5 and 6. The CGM1 and PRGM estimators perform well due to their appearance in 14 and 13 positions of the 22 positions, respectively.

In estimating A_i, we observe that the GT estimator performs very well due to its appearance in 21 positions of the corresponding 22 positions in Figures 7 and 8. The CGM1 and CGM0 estimators perform well due to their appearance in 14 and 11 positions of the

22 positions in Figures 7 and 8, respectively.

We analyzed the CAD data set to estimate the LFDR of each of the exonic SNPs. We considered the exonic and ncRNA SNPs to define a separate and a combined reference class and observed a significance different in the results. While the data analysis using the separate reference class identified 7 SNPs associated with the disease, the combined reference class suggested that only one of these SNPs is associated with the disease. The ME method, as the only method considered that takes into account the reliability of the separate reference class, led to discovery of only two SNPs that are actually associated with the disease.

Among the proposed methods, the ME method may be considered a reasonable default since it performs well in most of our simulations and since it is the only method considered that incorporates estimates of the reliability of the candidate reference classes. However, other methods may perform better, depending on the number of features and the values of π_{0S}, π_{0C}, $d_{\text{alt}S}$ and $d_{\text{alt}C}$. Users with information about those quantities may consult Figures 5-8 to decide which method to use to analyze their data.

An anonymous referee pointed out that the reference class problem could be addressed with Bayesian model averaging instead of the maximum entropy and robust Bayes approaches considered here. Indeed, a hierarchical Bayes approach like of Veyrieras et al. [37] and Wen et al. [41] may be optimal given the joint prior distribution of all unknowns. For example, the hierarchical model of Veyrieras et al. [37] requires prior distributions of effect sizes in addition to a prior probability for each SNP. Unfortunately, the prior probabilities and prior distributions needed for fully Bayesian hierarchical models are not always known precisely, which is why Pickrell [32] resorts to empirical Bayes estimation and cross validation.

However, the priors underlying hierarchical models are known to belong to some set, other approaches, such as maximum entropy and robust Bayes methods, may obtain unique estimates. Those methods would then apply at the higher level of priors rather than at the level of the reference classes in the current paper. In practice, there is the usual tradeoff between simplicity and how much one can rely on assumptions about prior distributions. The methods proposed here and in Aghababazadeh et al. [1] are simple but are not optimal if the relevant priors are known, in which case fully Bayesian methods would be optimal; otherwise, applying the methods of this paper to hierarchical models might be worth the added complexity. A more straightforward extension of the proposed approach would generalize it to biological problems involving more than two candidate reference classes. Appendix C explains an extension in another direction.

We emphasize that our theoretical developments are general and can be applied in some problems that there are more than two reference classes. For example, given nested reference classes, the ME method may be successively applied from the largest class to the smallest. Also we emphasize that our theoretical results based on the L_{ZO} loss function are general and one might choose different values for l_I and l_{II}. Our interest was to choose $l_I = 4$ and $l_{II} = 1$ which gives a 20 percent threshold in (5) which has been considered in [17] as a conventional threshold for reporting interesting cases in different real data sets. We should add that the same results are observable under the L_{SE} loss function (2).

APPENDIX A
PROOF BEHIND EQUATION (11).

From (4) we have

$$\max_{R\in\{S,C\}} \rho(\widehat{P}_R^i, \delta_i) = \begin{cases} l_I\overline{\psi_i} & \text{if } \delta_i = 1, \\ l_{II}\left(1-\underline{\psi_i}\right) & \text{if } \delta_i = 0, \end{cases}$$

and

$$\min_{R\in\{S,C\}} \rho(\widehat{P}_R^i, \delta_i) = \begin{cases} l_I\underline{\psi_i} & \text{if } \delta_i = 1, \\ l_{II}\left(1-\overline{\psi_i}\right) & \text{if } \delta_i = 0. \end{cases}$$

Thus,

$$\begin{aligned} &\delta_{\text{opt}}^{ZO,\kappa}(t_i) \\ &= \arg\inf_{\delta_i\in\mathcal{D}} \begin{cases} l_I(\kappa\overline{\psi_i} + (1-\kappa)\underline{\psi_i}) & \text{if } \delta_i = 1, \\ l_{II}(\kappa(1-\underline{\psi_i}) + (1-\kappa)(1-\overline{\psi_i})) & \text{if } \delta_i = 0. \end{cases} \end{aligned}$$

This leads to an optimal caution-type estimate of the ith hypothesis indicator as provided in (11).

APPENDIX B
PROOF BEHIND EQUATIONS (12)-(14).

Here we suppose that $\widehat{\theta}_i$ is positive. The proof for the case when $\widehat{\theta}_i$ is negative is the same and hence omitted.

To derive the conditional gamma minimim action $\delta_{\text{opt}}^{\text{SEL},0}(t_i)$ under the SEL function (2), notice from (6) that if $\delta_i < \frac{\widehat{\theta}_i}{2}$,

$$\begin{aligned} &0.5\min_{R\in\{S,C\}} \rho(\widehat{P}_R^i, \delta_i) + 0.5\max_{R\in\{S,C\}} \rho(\widehat{P}_R^i, \delta_i) \\ &\quad = 0.5(2\delta_i - \widehat{\theta}_i)\widehat{\theta}_i\overline{\psi_i} + 0.5(2\delta_i - \widehat{\theta}_i)\widehat{\theta}_i\underline{\psi_i} + (\delta_i - \widehat{\theta}_i)^2, \end{aligned}$$

and similarly, if $\delta_i \geq \frac{\widehat{\theta}_i}{2}$,

$$\begin{aligned} &0.5\min_{R\in\{S,C\}} \rho(\widehat{P}_R^i, \delta_i) + 0.5\max_{R\in\{S,C\}} \rho(\widehat{P}_R^i, \delta_i) \\ &\quad = 0.5(2\delta_i - \widehat{\theta}_i)\widehat{\theta}_i\underline{\psi_i} + 0.5(2\delta_i - \widehat{\theta}_i)\widehat{\theta}_i\overline{\psi_i} + (\delta_i - \widehat{\theta}_i)^2. \end{aligned}$$

Summarizing the above equations, we observe that for any δ_i,

$$0.5 \min_{R\in\{S,C\}} \rho(\widehat{P}_R^i, \delta_i) + 0.5 \max_{R\in\{S,C\}} \rho(\widehat{P}_R^i, \delta_i)$$
$$= 0.5(2\delta_i - \widehat{\theta}_i)\widehat{\theta}_i\underline{\psi_i} + 0.5(2\delta_i - \widehat{\theta}_i)\widehat{\theta}_i\overline{\psi_i} + (\delta_i - \widehat{\theta}_i)^2.$$

After some algebraic manipulations, we reach the equation(12).

To derive the conditional gamma minimax action $\delta_{\text{opt}}^{\text{SEL},1}(t_i)$, a discussion on values of δ_i in the posterior risk function (6) leads to the following statement

$$\max_{R\in\{S,C\}} \rho(\widehat{P}_R^i, \delta_i)$$
$$= \begin{cases} \left(2\delta_i - \widehat{\theta}_i\right)\widehat{\theta}_i\overline{\psi_i} + \left(\delta_i - \widehat{\theta}_i\right)^2 & \text{if } \delta_i \geq \frac{\widehat{\theta}_i}{2}, \\ \left(2\delta_i - \widehat{\theta}_i\right)\widehat{\theta}_i\underline{\psi_i} + \left(\delta_i - \widehat{\theta}_i\right)^2 & \text{if } \delta_i < \frac{\widehat{\theta}_i}{2}. \end{cases}$$

Now, some algebraic manipulations lead to the equation (14).

To derive the caution-type action with $\kappa = 0.5$, i.e. $\delta_{\text{opt}}^{\text{SEL},0.5}(t_i)$, note that

$$\min_{R\in\{S,C\}} \rho(\widehat{P}_R^i, \delta_i)$$
$$= \begin{cases} \left(2\delta_i - \widehat{\theta}_i\right)\widehat{\theta}_i\underline{\psi_i} + \left(\delta_i - \widehat{\theta}_i\right)^2 & \text{if } \delta_i \geq \frac{\widehat{\theta}_i}{2}, \\ \left(2\delta_i - \widehat{\theta}_i\right)\widehat{\theta}_i\overline{\psi_i} + \left(\delta_i - \widehat{\theta}_i\right)^2 & \text{if } \delta_i < \frac{\widehat{\theta}_i}{2}. \end{cases}$$

Defining $k = \underline{\psi_i} + \overline{\psi_i} \in [0, 2]$ followed by come calculations leads us to (14).

APPENDIX C
A REFERENCE CLASS PROBLEM IN IDENTIFYING CAUSAL SNPS

This Appendix gives an example of how to extend the approach of the main text to hierarchical models.

By taking linkage disequilibrium (LD) into account, Pickrell [32] endeavors to build a model to identify the common characteristics of SNPs that causally influence a trait. Pickrell assumes that an N-SNP genome can be split into blocks of K SNPs each, and assigns the total number to N/K blocks such that each the number of SNPs in the block that causally influences the trait is either 0 or 1 (Bickel et al. [12] make a related assumption for gene network reconstruction from gene expression data). The probability of observing the data $\vec{t} = (t_1, t_2, \ldots, t_N)$ is

$$g(\vec{t}) = \prod_{k=1}^{N/K} (1 - \Pi_k)\, g_k^0(\vec{t}) + \Pi_k g_k^1(\vec{t}),$$

where Π_k is the prior probability that the kth block contains a causal SNP associated with the trait, g_k^0 is the probability function given that no causal SNPs are associated with the trait in the kth block, and g_k^1 is the probability function given that a causal SNP in the kth block is associated with the trait. The latter probability function is specified by

$$g_k^1(\vec{t}) = \sum_{i\in S_k} \pi_{ik} g_k^1(t_i),$$

where S_k is the set of SNPs in the kth block, π_{ik} is the prior probability that SNP i is the causal SNP associated with the trait in the kth block, and $g_k^1(t_i)$ is the probability of sample t_i given that the ith SNP in the kth block is causal and associated with the trait.

The following reference class problem may arise that can be addressed using the approach of our main text. The N SNPs of the genome might be split into two subclasses according to some known characteristics of the SNPs. Each subclass would be large enough to in turn be split into multiple blocks of K SNPs each. Should the combined class of N SNPs be treated as the reference class, as per Pickrell [32], or should each of the two subclasses be considered reference classes for separate analyses?

ACKNOWLEDGMENTS

The authors would like to thank Marta Padilla for her assistance with `R` syntax. The authors are also grateful to Majid Nikpay for his assistance with real data preparation. We used the following packages of R [33]: `Biobase` [21] and `qvalue` [13] from Bioconductor [21]; `locfdr` [20], `fBasics` [42], and `distr` [34] from the CRAN repository. Functional annotation was performed by Majid Nikpay using the ANNOVAR software [38]. AK made his contributions as a postdoctoral fellow at the University of Ottawa. This research was partially supported by the Canadian Institutes of Health Research (Grant No. 123508), the Natural Sciences and Engineering Research Council of Canada (RGPIN/356018-2009), and the Faculty of Medicine of the University of Ottawa. This study makes use of the data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of this data set is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. We thank the two anonymous referees for feedback that led to clarifications.

REFERENCES

[1] Aghababazadeh, F. A., Alvo, M., Bickel, D. R., 2016. Estimating the local false discovery rate via a bootstrap solution to the reference class problem. Working Paper, University of Ottawa, deposited in uO Research at http://hdl.handle.net/10393/34295.

[2] Allison, D. B., Gadbury, G. L., Heo, M., Fernandez, J. R., Lee, C. K., Prolla, T. A., Weindruch, R., 2002. A mixture model approach for the analysis of microarray gene expression data. Computational Statistics and Data Analysis 38, 1–20.

[3] Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 57, 289–300.

[4] Berger, J., 1990. Robust Bayesian analysis: sensitivity to the prior. Journal of Statistical Planning and Inference 25, 303–328.

[5] Berger, J. O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer, New York.

[6] Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., et al., 1994. An overview of robust bayesian analysis. Test 3 (1), 5–124.

[7] Betro, B., Ruggeri, F., 1992. Conditional Γ-minimax actions under convex losses. Communications in Statistics - Theory and Methods 21, 1051–1066.

[8] Bickel, D. R., 2012. Game-theoretic probability combination with applications to resolving conflicts between statistical methods. International Journal of Approximate Reasoning 53, 880–891.

[9] Bickel, D. R., 2013. Simple estimators of false discovery rates given as few as one or two p-values without strong parametric assumptions. Statistical applications in genetics and molecular biology 12 (4), 529–543.

[10] Bickel, D. R., 2015. Inference after checking multiple Bayesian models for data conflict and applications to mitigating the influence of rejected priors. International Journal of Approximate Reasoning 66, 53–72.

[11] Bickel, D. R., 2016. Correcting false discovery rates for their bias toward false positives. Working Paper, University of Ottawa, deposited in uO Research at http://hdl.handle.net/10393/34277.

[12] Bickel, D. R., Montazeri, Z., Hsieh, P. C., Beatty, M., Lawit, S. J., Bate, N. J., 2009. Gene network reconstruction from transcriptional dynamics under kinetic model uncertainty: a case for the second derivative. Bioinformatics 25 (6), 772–779.

[13] Dabney, A., Storey, J. D., with assistance from Gregory R. Warnes, 2011. qvalue: Q-value estimation for false discovery rate control. Reference Manual, R package version 1.26.0.

[14] Efron, B., 2004. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association 99, 96–104.

[15] Efron, B., 2007. Correlation and large-scale simultaneous significance testing. Journal of the American Statistical Association 102, 93–103.

[16] Efron, B., 2008. Microarrays, empirical Bayes and the two-groups model. Statistical science 23 (1), 1–22.

[17] Efron, B., 2010. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge University Press, Cambridge.

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

[19] 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.

[20] Efron, B., Turnbull, B. B., Narasimhan, B., 2011. locfdr: Computes local false discovery rates. Reference Manual, R package version 1.1-7.

[21] Gentleman, R. C., Carey, V. J., Bates, D. M., et al., 2004. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 5, R80.

[22] Hurwicz, L., 1951. Optimality criteria for decision making under ignorance. Cowles Commission Discussion Paper 370.

[23] Insua, D. R., Ruggeri, F., Vidakovic, B., 1992. Some results on posterior regret Γ-minimax estimation. Statistics and Decisions 13, 315–351.

[24] Jaffray, J.-Y., 1989. Généralisation du critère de l'utilité espérée aux choix dans l'incertain régulier. RAIRO: Recherche opérationnelle 23, 237–267.

[25] Karimnezhad, A., Parsian, A., 2014. Robust Bayesian methodology with applications in credibility premium derivation and future claim size prediction. AStA Advances in Statistical Analysis 98 (3), 287–303.

[26] Karimnezhad, A., Parsian, A., 2018. Most stable sample size determination in clinical trials. Statistical Methods & Applications. https://doi.org/10.1007/s10260-017-0419-6

[27] Karimnezhad, A., Lucas, P. J., Parsian, A., 2017. Constrained parameter estimation with uncertain priors for Bayesian networks. Electronic Journal of Statistics 11(2), 4000–4032.

[28] Mei S, Karimnezhad A, Forest M, Bickel D. R., Greenwood C. M. T. (2017) 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 12(9): e0185174. https://doi.org/10.1371/journal. pone.0185174

[29] Muralidharan, O., 2010. An empirical bayes mixture method for effect size and false discovery rate estimation. The Annals of Applied Statistics, 422–438.

[30] Padilla, M., Bickel, D. R., 2012. Estimators of the local false discovery rate designed for small numbers of tests. Statistical Applications in Genetics

and Molecular Biology 11 (5), art. 4.
[31] Pan, W., Lin, J., Le, C. T., 2003. A mixture model approach to detecting differentially expressed genes with microarray data. Functional & integrative genomics 3 (3), 117–124.
[32] Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. The American Journal of Human Genetics 94 (4), 559-573.
[33] R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
[34] Ruckdeschel, P., Kohl, M., Stabla, T., Camphausen, F., May 2006. S4 classes for distributions. R News 6 (2), 2–6.
[35] Storey, J. D., 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society. Series B: Statistical Methodology 64, 479–498.
[36] Storey, J. D., 2003. The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. The Annals of Statistics 31 (6), pp. 2013–2035.
[37] Veyrieras, J. B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., Pritchard, J. K., 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genetics 4 (10), e1000214.
[38] Wang, K., Li, M., Hakonarson, H., 2010. Annovar: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 38 (16), e164–e164.
[39] Watson, S. R., 1974. On Bayesian inference with incompletely specified prior distributions. Biometrika 61 (1), 193–196.
[40] Wellcome Trust Case Control Consortium, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678.
[41] Wen, X., Luca, F., Pique-Regi, R., 2015. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. PLoS genetics 11 (4), e1005176.
[42] Wuertz, D., 2010. fbasics: Rmetrics - markets and basic statistics. Reference Manual, R package version 2110.79.
[43] Yang, Y., Aghababazadeh, F. A., Bickel, D. R., 2013. Parametric estimation of the local false discovery rate for identifying genetic associations. IEEE/ACM Transactions on Computational Biology and Bioinformatics 10, 98–108.
[44] Zen, M.-M., DasGupta, A., et al., 1990. Estimating a binomial parameter: is robust Bayes real Bayes? Purdue University. Department of Statistics.

Ali Karimnezhad is a Postdoctoral Fellow at the Ottawa Hospital Research Institute and the University of Ottawa. He received his BSc from the University of Tehran in 2007, his MSc from the Allameh Tabataba'i University in 2010, and his PhD from the University of Tehran in 2014. He has published several papers in the area of Bayes and robust Bayes inference. His main research interests include robust Bayesian analysis, Bayesian parameter and structure learning, and bioinformatics.

David R. Bickel is an Associate Professor at the Ottawa Institute of Systems Biology. His research interests include statistical genomics and the foundations of statistics. After applying his statistical methods to the analysis of gene expression data, he has developed methods in response to problems with analyzing genome-wide association data and to measurements for smaller numbers of biological features (e.g., gene ontology terms or molecular species).