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

Ali Karimnezhad and David R. Bickel

June 9, 2016

Ottawa Institute of Systems Biology

Biochemistry, Microbiology, and Immunology Department

Mathematics and Statistics Department

University of Ottawa

451 Smyth Rd.

Ottawa, Ontario K1H 8M5

dbickel@uottawa.ca

Abstract

An objective probability that a single nucleotide polymorphism (SNP) is 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 solve this reference class problem, we introduce novel 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 avoids choosing the worst reference class by depending on the separate class when it has enough SNPs for reliable LFDR estimation and depending solely on the combined class otherwise. The new game-theoretic rule also performs well relative to previous approaches. Analyzing data from a genome-wide association study of 2000 cases and 3000 controls, we find that the maximum entropy method makes notably different identifications of disease-associated SNPs than the two reference classes considered.

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 (1995) introduces the concept of false discovery rate (FDR) and since then many developments have been conducted (Efron and Tibshirani 2002; Storey 2002, 2003; Efron 2010 among others). A quantity to discover if a specific SNP is associated with the disease is to compute the posterior probability that the null hypothesis is true. But, the hypothesis posterior probability which is sometimes referred to as the local false discovery rate (LFDR), depends on some parameters that are usually unknown (Bickel, 2013; Efron et al., 2001). Thus, 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 (Efron and Tibshirani, 2002; Efron et al., 2008). 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. (2002), Pan et al. (2003) and Efron (2004, 2007) consider the estimation of LFDR based on a discrete mixture model. Also Muralidharan (2010), Padilla and Bickel (2012) and Yang et al. (2013) consider the LFDR estimation using the maximum likelihood (ML) approach. Bickel (2013) 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 (2016) 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 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. (2016) 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 as well as results from the CAD data analysis. 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 in the hypothesis set $H_{0i} : A_i = 0$ vs $H_{1i} : A_i = 1$, $i = 1, 2, \ldots, N$, where A_i is an indicator variable indicating whether H_{1i} (the ith alternative hypothesis) is true. 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-squared distribution with one degree of freedom (Yang et al., 2013).

Consider there are 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. 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 statis-

tic 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 $H_{1i} : A_i = 1$, $i \in R = \{S, C\}$. 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

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

where t_i is a realization of the test statistic T_i having the pdf $g_0(.)$ conditional on the null hypothesis and pdf $g_{d_{\text{alt}R}}(.)$ conditional on the alternative hypothesis (Efron, 2010; Bickel, 2013). 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 $g_{d_{\text{alt}R}}$ have to be estimated (Padilla and Bickel, 2012; Yang et al., 2013).

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 estimates of π_{0R} and $d_{\text{alt}R}$ by $\widehat{\pi}_{0R}$ and $\hat{d}_{\text{alt}R}$.

Let $\delta_i = \delta(t_i)$ be a decision rule based on the test statistic value t_i. For ith SNP consider the following loss functions

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

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

and

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

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. The squared error loss (SEL) 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 SEL function L_{SE} in (2) can be derived from results of estimating θ_i under the SEL 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

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

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

$$\delta_{\text{ZO}}^{\pi_R}(t_i) = \begin{cases} 1 & \text{if } \widehat{\psi}_{i,R} \leq \frac{l_{II}}{l_I + l_{II}}, \\ 0 & \text{if } \widehat{\psi}_{i,R} > \frac{l_{II}}{l_I + l_{II}}. \end{cases} \tag{4}$$

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

$$\rho\left(\widehat{P}_R^i, \delta_i\right) = \delta_i^2 \widehat{\psi}_{i,R} + (\delta_i - \widehat{\theta}_i)^2 \left(1 - \widehat{\psi}_{i,R}\right) \tag{5}$$

which leads to the following Bayes estimator

$$\delta_{\text{OR}}^{\pi_R}(t_i) = \widehat{\theta}_i(1 - \widehat{\psi}_{i,R}). \tag{6}$$

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 (5) 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 (6) 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 in subsection 2.2.

3.1 A maximum entropy method of pooling distributions

To discover SNPs associated with a specific disease, 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.

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 (2015), define the following likelihood set

$$L_S = \left\{ \tau : \frac{L(\tau)}{L(\widehat{\tau}_S)} \geq \frac{1}{2^a},\, \tau \in [0,1] \times [d_1, d_2] \right\}, \tag{7}$$

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 $\prod_{i=1}^{N_S} \left(\pi_{0R} g_0(t_i) + (1-\pi_{0R}) g_{d_{\text{alt}R}}(t_i)\right)$ over $(\pi_{0R}, d_{\text{alt}R})$.

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 (2015) 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 (7), 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

$$D\left(P^i||\widehat{P}_C^i\right) = \psi_i \log\left(\frac{\psi_i}{\widehat{\psi}_{i,C}}\right) + (1-\psi_i)\log\left(\frac{1-\psi_i}{1-\widehat{\psi}_{i,C}}\right). \tag{8}$$

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 ν_i using the equations (4) and (6), 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 (8). 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 (8), 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 $\widehat{\psi}_{i,\mathrm{ME}} = \widehat{\psi}_{i,C}$. Otherwise, if $\widehat{\psi}_{i,C} < \psi^L_{i,S}$, then $\widehat{\psi}_{i,\mathrm{ME}} = \psi^L_{i,S}$ and if $\widehat{\psi}_{i,C} > \psi^U_{i,S}$, then $\widehat{\psi}_{i,\mathrm{ME}} = \psi^U_{i,S}$.

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 (2012). 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 (2012), 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 (4) and (6), respectively. To do so, we replace $\widehat{\psi}_{i,R}$ by $\widehat{\psi}_{i,\mathrm{GT}}$.

4 INFERENCE VIA ROBUST BAYES METHODS

As observed in Sections 2 and 3, 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 (1985, 1990) and Berger et al. (1994). Also, Karimnezhad et al. (2014) and Karimnezhad and Parsian (2014) 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.*

4.1 Caution-type estimators

Decision-theoretic rules can be chosen by caution or without any caution given a viable set of prior distributions (Hurwicz, 1951, Jaffray, 1989, Bickel, 2012). In this regard, we apply an extended version of the criterion introduced by Hurwicz (1951) and followed later by Jaffray (1989) and Bickel (2015) for the hypothesis testing problem. For an ith hypothesis indicator, this criterion specifies an optimal action satisfying

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

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

$$\delta^{\text{ZO},\kappa}_{\text{opt}}(t_i) = \begin{cases} 1 & \text{if } l_{II}(1-\underline{\psi_i}-\overline{\psi_i}) \geq h_i(\kappa), \\ 0 & \text{if } l_{II}(1-\underline{\psi_i}-\overline{\psi_i}) < h_i(\kappa), \end{cases} \tag{9}$$

where $h_i(\kappa) = (l_I - l_{II})(\kappa\overline{\psi_i} + (1-\kappa)\underline{\psi_i})$.

Caution-type actions for some specific values of κ are interesting. The most cautious attitude ($\kappa = 1$), corresponds to the conditional gamma minimax strategy (Betro and Ruggeri, 1992; Watson, 1974) and the least cautious attitude ($\kappa = 0$), might be referred to as conditional gamma minimin. 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 CGM1, CGM0.5 and CGM0, respectively.

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

$$\delta_{\text{opt}}^{\text{SEL},1}(t_i) = \begin{cases} \widehat{\theta}_i(1-\overline{\psi_i}) & \text{if } \underline{\psi_i} \le \overline{\psi_i} \le \frac{1}{2}, \\ \frac{1}{2}\widehat{\theta}_i & \text{if } \underline{\psi_i} < \frac{1}{2} < \overline{\psi_i}, \\ \widehat{\theta}_i(1-\underline{\psi_i}) & \text{if } \frac{1}{2} < \underline{\psi_i} \le \overline{\psi_i}, \end{cases} \tag{10}$$

$$\delta_{\text{opt}}^{\text{SEL},0.5}(t_i) = \begin{cases} \widehat{\theta}_i(1-\underline{\psi_i}) & \text{if } \underline{\psi_i} \le \overline{\psi_i} \le \frac{1}{2}, \\ \widehat{\theta}_i\left(1-\underline{\psi_i} I_i^{[0,1]} - \overline{\psi_i} I_i^{(1,2]}\right) & \text{if } \underline{\psi_i} < \frac{1}{2} < \overline{\psi_i}, \\ \widehat{\theta}_i(1-\overline{\psi_i}) & \text{if } \frac{1}{2} < \underline{\psi_i} \le \overline{\psi_i}, \end{cases} \tag{11}$$

where for subset U of the real line,

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

and

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

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. (1990), and excellent developments can be found in Berger (1985), Insua et al. (1992) and Berger et al. (1994).

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 (9)-(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 $\sigma^2 = 0.02$. We then transform the z_i values to chi-squared values through the transformation $t_i = (\frac{z_i}{\sigma})^2$. By this transformation, under the null hypothesis t_i follows a central chi-squared distribution with one degree of freedom and under the alternative hypothesis it follows a non-central chi-squared 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 Sections 3 and 4. Finally, we define average of risks (AMSE and AR_4 in Step 9 of the Algorithm 1) to measure performance of the methods.

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 $\sigma^2 = 0.02$, $N_{0S} = 0(200)2000$ and $N_S = 2000$. By this setting, a separate reference class, say S, with $\pi_{0S} = 0(0.1)1$ 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-squared 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 $\delta_{\mathrm{ZO}}^{\pi_R}(t_i)$ with $R \in \{S, C\}$, $\delta_{\mathrm{opt}}^{\mathrm{ZO},\kappa}(t_i)$ and $\delta_{\mathrm{opt}}^{\mathrm{SEL},\kappa}(t_i)$ 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 $L_{\mathrm{ZO}}^j(A_i, \delta_i)$ and $L_{\mathrm{OR}}^j(\theta_i, \delta_i)$ introduced in (1) and (3), where δ_i is any of the applicable estimators computed in Step 6. For the L_1 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

$$\mathrm{R}_{4,i} = \frac{1}{N}\sum_{j=1}^{N} L_{\mathrm{ZO}}^j(A_i, \delta_i), \qquad \mathrm{MSE}_i = \frac{1}{N}\sum_{j=1}^{N} L_{\mathrm{OR}}^j(\nu_i, \delta_i),$$

   where the index 4 in $\mathrm{R}_{4,i}$ refers to the choice $l_1 = 4$.

9. For each of the proposed methods, compute averages of $\mathrm{R}_{4,i}$ and MSE_i over all SNPs in the separate reference class S, i.e.,

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

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 2 (or the point $\pi_{0S} = 0$ in Figures 3 and 4) for which there is a 0.50 difference between and π_{0S} and π_{0C}. In fact, from Figures 1 and 2 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 3 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 in Sections 3 and 4 for all values of π_{0S} in different settings is satisfactory. They lead to a decrease in AMSE or AR_4 values.

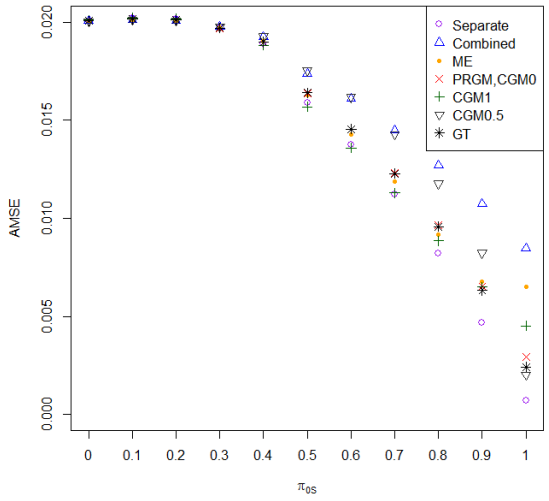

Figure 1: Plots of AMSE when $\pi_{0M} = 0$ for different values of π_{0S}.

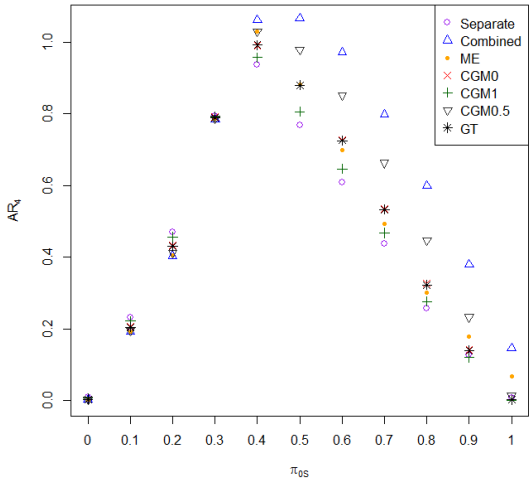

Figure 2: Plots of AR_4 when $\pi_{0M} = 0$ for different values of π_{0S}.

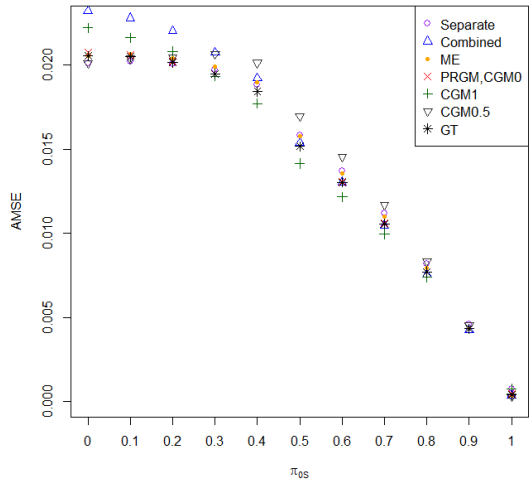

Figure 3: Plots of AMSE when $\pi_{0M} = 1$ for different values of π_{0S}.

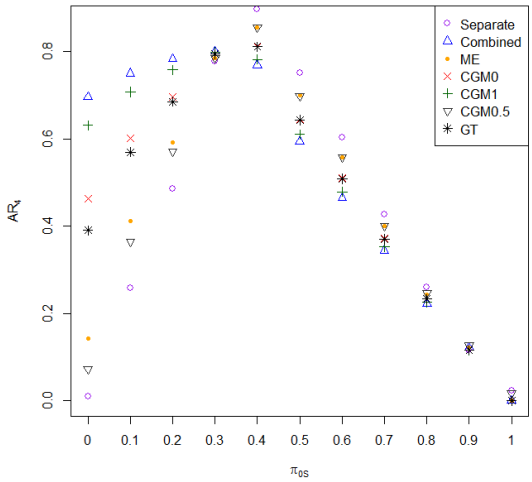

Figure 4: Plots of AR_4 when $\pi_{0M} = 1$ for different values of π_{0S}.

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}, Table 1 orders the top three estimators of $\theta_i = \log(OR_i)$ 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 Wellcome Trust Case Control Consortium (2007)[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 (Wang et al., 2010). Figure 5 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 that she/he might 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-squared distribution with one degree of freedom. The test statistics under the alternative hypothesis follow a non-central chi-squared distribution with one degree of

[1] www.wtccc.org.uk

Table 1: Comparison of the performance of the estimators for different values of π_{0S}, proportion of unassociated SNPs in the separate reference class, and π_{0M}, proportion of unassociated SNPs in the combined reference class but not inside the separate reference class. Each cell lists the best three estimators among the six estimators considered according either to averages of MSE when estimating the effect size $\theta_i = \log(\mathrm{OR}_i)$, or averages of R_4 when estimating the hypothesis indicator A_i, where $A_i = 0$, if ith SNPs is not associated with the disease and $A_i = 1$, otherwise. AMSE_d refers to averages of MSE and AR_{4d} refers to averages of R_4 of the estimator d. The averaging in AMSE_d and AR_{4d} has been performed over the 2000 SNPs in the separate reference class.

	In estimating θ_i		In estimating A_i	
π_{0S}	$\pi_{0M} = 0$	$\pi_{0M} = 1$	$\pi_{0M} = 0$	$\pi_{0M} = 1$
$[0, 0.1[$	$\mathrm{AMSE}_{\mathrm{CGM0.5}} < \mathrm{AMSE}_{\mathrm{ME}} < \mathrm{AMSE}_{\mathrm{GT}}$	$\mathrm{AMSE}_{\mathrm{CGM0.5}} < \mathrm{AMSE}_{\mathrm{ME}} < \mathrm{AMSE}_{\mathrm{GT}}$	$\mathrm{AR}_{4\mathrm{CGM0.5}} < \mathrm{AR}_{4\mathrm{ME}} < \mathrm{AR}_{4\mathrm{GT}}$	$\mathrm{AR}_{4\mathrm{CGM0.5}} < \mathrm{AR}_{4\mathrm{ME}} < \mathrm{AR}_{4\mathrm{GT}}$
$[0.1, 0.2[$	$\mathrm{AMSE}_{\mathrm{CGM0.5}} < \mathrm{AMSE}_{\mathrm{ME}} < \mathrm{AMSE}_{\mathrm{GT}}$	$\mathrm{AMSE}_{\mathrm{CGM0.5}} < \mathrm{AMSE}_{\mathrm{GT}} < \mathrm{AMSE}_{\mathrm{PRGM}}$	$\mathrm{AR}_{4\mathrm{ME}} < \mathrm{AR}_{4\mathrm{CGM0.5}} < \mathrm{AR}_{4\mathrm{GT}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{PRGM}} < \mathrm{AR}_{4\mathrm{GT}}$
$[0.2, 0.3[$	$\mathrm{AMSE}_{\mathrm{CGM0.5}} < \mathrm{AMSE}_{\mathrm{ME}} < \mathrm{AMSE}_{\mathrm{GT}}$	$\mathrm{AMSE}_{\mathrm{GT}} < \mathrm{AMSE}_{\mathrm{PRGM}} < \mathrm{AMSE}_{\mathrm{ME}}$	$\mathrm{AR}_{4\mathrm{ME}} < \mathrm{AR}_{4\mathrm{CGM0.5}} < \mathrm{AR}_{4\mathrm{GT}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{GT}} < \mathrm{AR}_{4\mathrm{PRGM}}$
$[0.3, 0.4[$	$\mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{PRGM}} < \mathrm{AMSE}_{\mathrm{GT}}$	$\mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{GT}} < \mathrm{AMSE}_{\mathrm{PRGM}}$	$\mathrm{AR}_{4\mathrm{ME}} < \mathrm{AR}_{4\mathrm{CGM0.5}} < \mathrm{AR}_{4\mathrm{GT}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{GT}} < \mathrm{AR}_{4\mathrm{PRGM}}$
$[0.4, 0.5[$	$\mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{PRGM}} < \mathrm{AMSE}_{\mathrm{GT}}$	$\mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{PRGM}} < \mathrm{AMSE}_{\mathrm{GT}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{PRGM}} < \mathrm{AR}_{4\mathrm{GT}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{GT}} < \mathrm{AR}_{4\mathrm{PRGM}}$
$[0.5, 0.6[$	$\mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{ME}} < \mathrm{AMSE}_{\mathrm{GT}}$	$\mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{GT}} < \mathrm{AMSE}_{\mathrm{PRGM}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{PRGM}} < \mathrm{AR}_{4\mathrm{GT}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{GT}} < \mathrm{AR}_{4\mathrm{PRGM}}$
$[0.6, 0.7[$	$\mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{ME}} < \mathrm{AMSE}_{\mathrm{GT}}$	$\mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{GT}} < \mathrm{AMSE}_{\mathrm{PRGM}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{ME}} < \mathrm{AR}_{4\mathrm{GT}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{GT}} < \mathrm{AR}_{4\mathrm{PRGM}}$
$[0.7, 0.8[$	$\mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{ME}} < \mathrm{AMSE}_{\mathrm{GT}}$	$\mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{GT}} < \mathrm{AMSE}_{\mathrm{PRGM}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{ME}} < \mathrm{AR}_{4\mathrm{GT}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{GT}} < \mathrm{AR}_{4\mathrm{PRGM}}$
$[0.8, 0.9[$	$\mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{ME}} < \mathrm{AMSE}_{\mathrm{GT}}$	$\mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{GT}} < \mathrm{AMSE}_{\mathrm{PRGM}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{ME}} < \mathrm{AR}_{4\mathrm{GT}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{GT}} < \mathrm{AR}_{4\mathrm{PRGM}}$
$[0.9, 1[$	$\mathrm{AMSE}_{\mathrm{GT}} < \mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{PRGM}}$	$\mathrm{AMSE}_{\mathrm{GT}} < \mathrm{AMSE}_{\mathrm{CGM1}} < \mathrm{AMSE}_{\mathrm{PRGM}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{GT}} < \mathrm{AR}_{4\mathrm{PRGM}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{GT}} < \mathrm{AR}_{4\mathrm{PRGM}}$
1	$\mathrm{AMSE}_{\mathrm{CGM0.5}} < \mathrm{AMSE}_{\mathrm{GT}} < \mathrm{AMSE}_{\mathrm{PRGM}}$	$\mathrm{AMSE}_{\mathrm{CGM0.5}} < \mathrm{AMSE}_{\mathrm{ME}} < \mathrm{AMSE}_{\mathrm{GT}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{GT}} < \mathrm{AR}_{4\mathrm{PRGM}}$	$\mathrm{AR}_{4\mathrm{CGM1}} < \mathrm{AR}_{4\mathrm{ME}} < \mathrm{AR}_{4\mathrm{PRGM}}$

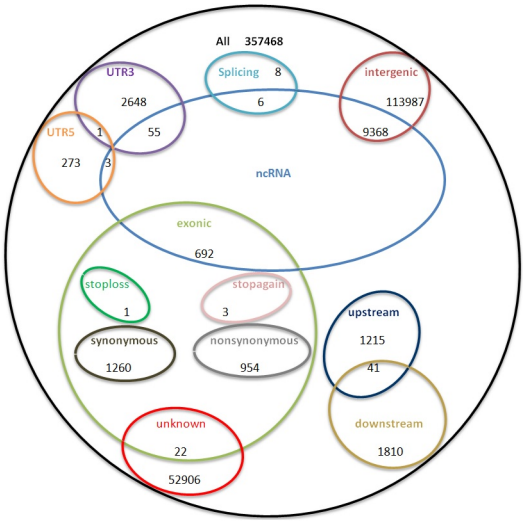

Figure 5: SNPs distribution based on the functional annotation.

freedom and non-centrality parameter $d_{\text{alt}R}$. Considering S and C for the reference classes including 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 6 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.

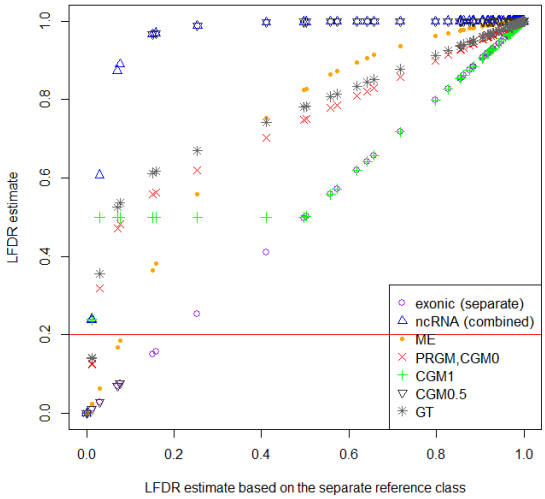

Figure 6: Estimated LFDRs for 692 exonic SNPs. The LFDR estimates w.r.t. exonic SNPs fall on the horizontal axis and LFDR estimates of the same SNPs regarding the different approaches are shown on the vertical axis.

Significant difference and discrepancies in the LFDRs estimated values (and thus in determining associated SNPs) is realized from Figure 6. 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.

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 1 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 3 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 squared of the bias. 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 3 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 appears in almost all cells of Table 1.

In estimating θ_i, the GT estimator appears in all the corresponding 22 cells in Table 1. The CGM1 and PRGM estimators perform well due to their appearance in 14 and 13 cells of the 22 cells, respectively.

In estimating A_i, we observe that the GT estimator performs very well due to its appearance in 21 cells of the corresponding 22 cells in Table 1. The CGM1 and PRGM estimators perform well due to their appearance in 17 and 14 cells of the 22 cells in Table 1, 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.

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 (4) which has been considered in Efron (2010)

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).

Acknowledgments

The authors would like to thank Dr. Marta Padilla for her assistance with `R` syntax. The authors are also grateful to Dr. Majid Nikpay for his assistance with real data preparation. We used the following packages of R (R Development Core Team, 2008): `Biobase` (Gentleman et al., 2004) and `qvalue` (Dabney et al., 2011) from Bioconductor (Gentleman et al., 2004); `locfdr` (Efron et al., 2011), `fBasics` (Wuertz, 2010), and `distr` (Ruckdeschel et al., 2006) from the CRAN repository. Functional annotation was performed by Dr. Majid Nikpay using the ANNOVAR software (Wang et al., 2010). This research was partially supported by the Canadian Institutes of Health Research 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.

References

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.

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.

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.

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

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

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.

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

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

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.

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.

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.

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.

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

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

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

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

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

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

Efron, B., et al., 2008. Microarrays, empirical bayes and the two-groups model. Statistical science 23 (1), 1–22.

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.

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

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

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

Karimnezhad, A., Niazi, S., Parsian, A., 2014. Bayes and robust Bayes prediction with an application to a rainfall prediction problem. Journal of the Korean Statistical Society 43 (2), 275–291.

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.

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

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.

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.

R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ruckdeschel, P., Kohl, M., Stabla, T., Camphausen, F., May 2006. S4 classes for distributions. R News 6 (2), 2–6.

Storey, J. D., 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society. Series B: Statistical Methodology 64, 479–498.

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.

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.

Watson, S. R., 1974. On Bayesian inference with incompletely specified prior distributions. Biometrika 61 (1), 193–196.

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.

Wuertz, D., 2010. fbasics: Rmetrics - markets and basic statistics. Reference Manual, R package version 2110.79.

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.

Zen, M.-M., DasGupta, A., et al., 1990. Estimating a binomial parameter: is robust Bayes real Bayes? Purdue University. Department of Statistics.