*Collection of Biostatistics Research Archive*
COBRA Preprint Series

*Year* 2010 *Paper* 74

# Minimum Description Length and Empirical Bayes Methods of Identifying SNPs Associated with Disease

Ye Yang* David R. Bickel†

*Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology

†Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology, Department of Mathematics and Statistics, dbickel@uottawa.ca



http://biostats.bepress.com/cobra/art74

# Minimum Description Length and Empirical Bayes Methods of Identifying SNPs Associated with Disease

Ye Yang and David R. Bickel

**Abstract**

The goal of determining which of hundreds of thousands of SNPs are associated with disease poses one of the most challenging multiple testing problems. Using the empirical Bayes approach, the local false discovery rate (LFDR) estimated using popular semiparametric models has enjoyed success in simultaneous inference. However, the estimated LFDR can be biased because the semiparametric approach tends to overestimate the proportion of the non-associated single nucleotide polymorphisms (SNPs). One of the negative consequences is that, like conventional p-values, such LFDR estimates cannot quantify the amount of information in the data that favors the null hypothesis of no disease-association.

We address this problem of the semiparametric approach by proposing two simple parametric methods under the minimum description length (MDL) and empirical Bayes frameworks. The performances of the estimators corresponding to the two proposed parametric models and of the popular semiparametric model are compared by simulation to select a method for analyzing genome-wide association data.

The application of the coronary artery disease data indicates that the semiparametric method sometimes leads to overfitting due to nonparametric density estimation. Unlike semiparametric methods, the analyses based on the two parametric models can measure the amount of information in the data that favors one hypothesis over another. In multiple simulation studies, the estimators associated with the parametric mixture model consistently performs better than those of the other two models.

# Minimum description length and empirical Bayes methods of identifying SNPs associated with disease

Ye Yang [1] and David R. Bickel [1,2*]

November 29, 2010

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

**Abstract**

The goal of determining which of hundreds of thousands of SNPs are associated with disease poses one of the most challenging multiple testing problems. Using the empirical Bayes approach, the local false discovery rate (LFDR) estimated using popular semiparametric models has enjoyed success in simultaneous inference. However, the estimated LFDR can be biased because the semiparametric approach tends to overestimate the proportion of the non-associated single nucleotide polymorphisms (SNPs). One of the negative consequences is that, like conventional p-values, such LFDR estimates cannot quantify the amount of information in the data that favors the null hypothesis of no disease-association.

We address this problem of the semiparametric approach by proposing two simple parametric methods under the minimum description length (MDL) and empirical

*to whom correspondence should be addressed

Bayes frameworks. The performances of the estimators corresponding to the two proposed parametric models and of the popular semiparametric model are compared by simulation to select a method for analyzing genome-wide association data.

The application of the coronary artery disease data (Wellcome Trust Case Control Consortium, 2007) indicates that the semiparametric method sometimes leads to overfitting due to nonparametric density estimation. Unlike semiparametric methods, the analyses based on the two parametric models can measure the amount of information in the data that favors one hypothesis over another. In multiple simulation studies, the estimators associated with the parametric mixture model consistently performs better than those of the other two models.

# 1 Introduction

Genome-wide association studies employ either case-control designs or larger cohort designs. Both study designs aim to determine which SNPs are associated with the presence or absence of disease. The most common measure of association is the odds ratio. Each allele (or combination of alleles in the dominant and recessive models) of a SNP is associated with the odds of disease; the odds ratio between two alleles is the odds of disease of one allele divided by the odds of disease of the other allele. See Hirschhorn and Daly (2005) for an informative review.

Two recent developments have made genome-wide association studies feasible. First, genotyping has become more accurate and more affordable. Second, markers can now be selected on the basis of linkage disequilibrium patterns observed across the human genome (International HapMap Consortium, 2005). By identifying SNPs associated with disease, genome-wide association studies can potentially lead to novel treatments and better disease diagnosis and prevention (Wellcome Trust Case Control Consortium, 2007).

In genome-wide association studies, different samples of individuals may be genotyped for different reasons. For example, some studies are conducted in stages, beginning with one or more screening stages of a high number of markers to select loci for a validation stage of a much smaller number of markers over a possibly larger patient group (e.g. Göing *et al.* (2001); McPherson *et al.* (2007)). The validation stage requires selecting markers as having more evidence of association than was possible at the screening stage. The regions of such loci may then be sequenced to identify genes that may causally influence the onset of the disease or other trait. Alternatively, loci may be selected for potential clinical use as disease risk factors in prognosis. Regardless of whether a data set is used to scan for associations to be validated at a later stage, to validate hypothesized associations, or to perform both functions simultaneously, its analysis will involve the selection of some SNPs for their putative association with the disease.

Most commonly, p-values reported to quantify evidence for disease association are adjusted to control a family-wise error rate (e.g. Montana (2006); He *et al.* (2006)). This practice, however, tends to be too conservative to identify many SNPs associated with disease that have small odds ratios. Approaches with more statistical power include those estimating a global false discovery rate (GFDR) (e.g. Sabatti *et al.* (2003)) and Bayesian methods (e.g. Wacholder *et al.* (2004); Wellcome Trust Case Control Consortium (2007)). Empirical Bayes approaches, including *local false discovery rate* (LFDR) estimation, have the advantage of the full Bayesian approach that they can specify a posterior probability that a particular SNP is associated with disease but without depending on a prior distribution specified before observing the data.

LFDR estimation does not suffer from the main two criticisms of standard p-value approaches with regard to how the issue of multiple comparisons is handled: high false-negative rates and difficulties of interpretation. First, standard p-value adjustments for multiple comparisons incur an unacceptable loss of power due to an attempt to control the

proportion of false positives among all tests. Such traditional correction of p-values for multiple comparisons is not appropriate for the large number of tests needed for genome-wide association (GWA) data sets since they have hundreds of thousands of SNPs. As a test-wise version of the GFDR, the expected number of false positives among all statistically significant results, the estimated LFDR of each SNP's association with disease can rigorously account for multiple comparisons without that loss of power. The second criticism of p-value approaches also applies to the control of the GFDR as an alternative method for addressing the multiple-comparisons problem. That alternative requires the selection of an arbitrary threshold, not allowing assignments of different levels of confidence for significance to different tests (Ziegler *et al.*, 2008). The standard response to this criticism is to compute the *q-value*, the minimum false discovery rate threshold at which each null hypothesis would be rejected (Storey, 2002). However, the q-value can be difficult to interpret since it does not approximate the probability that the hypothesis is true (Bukszár *et al.*, 2009). Unlike the p-value and q-value, the LFDR estimate is easily interpreted as an approximate posterior probability that the null hypothesis is true (Greenwood *et al.*, 2007; Bukszár *et al.*, 2009), provided that the estimate is not essentially 0% or 100% (Bickel, 2010a). At the same time, estimation of the LFDR is built on standard principles of frequentist inference, not relying on the choice of subjective or default prior distributions, as in the purer Bayesian approaches that Wakefield (2007) and Wei *et al.* (2010) applied to GWA data.

The estimation of LFDR and GFDR is normally performed using a discrete mixture model, whether parametric (e.g., Allison *et al.* (2002); Pounds and Morris (2003); Pan *et al.* (2003); Liao *et al.* (2004); Muralidharan (2010)) or semiparametric (e.g., Efron *et al.* (2001); Efron (2004, 2007a)). The main application of semiparametric LFDR estimation has been to microarray gene expression data (e.g., Efron *et al.* (2001); Efron (2004, 2007a); Aubert *et al.* (2004)). In contrast with the large number of applications to microarray data

analyses, only a few papers have applied LFDR estimation to GWA data (e.g., Greenwood *et al.* (2007)). The estimated LFDR using the semiparametric approach is designed to be conservative, leading to the bias that Pawitan *et al.* (2005) addressed by fitting a parametric model to gene expression data. The parametric model has the additional advantage that it can quantify the strength of the statistical evidence under the likelihood paradigm, as explained below.

According to the likelihood framework, the likelihood ratio measures the degree to which the data supports one hypothesized distribution over another (Blume, 2002; Hacking, 1965; Edwards, 1969; Royall, 1997). As a ratio of probability densities, the likelihood ratio requires a single density associated with the alternative hypothesis and a single density associated with the null hypothesis. Since, however, the alternative hypothesis typically corresponds to an infinite number of probability density functions, a single density function must be selected for the numerator of the likelihood ratio. While the principle of inference to the best explanation suggests maximizing the alternative density over the parameter space (Bickel, 2010d), that approach leads to overfitting from a predictive standpoint since the density function selected depends on the same data as the likelihood. The full Bayesian solution is to assign priors in order to integrate the likelihoods associated with each hypothesis (Wei *et al.*, 2010). Without requiring any prior distribution, the minimum description length (MDL) principle also prevents overfitting by requiring that the density function not depend on the data (Grünwald, 2007).

In Section 2, we will apply two parametric mixture models that approximate MDL densities (Bickel, 2010b) to the problem of assessing the strength of evidence for association with disease. We will also apply the same models and a semiparametric model of Efron (2004) to LFDR estimation. Section 3 describes our application of those models to coronary artery disease data (Wellcome Trust Case Control Consortium, 2007). In Section 4, we report the relative performance of each of the estimators associated with the three models.

Finally, we conclude the paper in Section 5 with recommendations for the analysis of GWA data.

# 2 Methods

## 2.1 Model assumed as true

To investigate the association between the disease state and the $i$th SNP, logistic regression estimates the $i$th SNP effect $\beta_i$, the logarithm of the odds ratio for the $i$th SNP, where $i = 1, \ldots, N$, and $N$ is the total number of the measured SNPs. For the $i$th SNP, we perform the Wald test with $\beta_i = 0$ as the null hypothesis and with $\beta_i \neq 0$ as the alternative hypothesis. The Wald test statistic is $T_i = \widehat{\beta}_i^2 / \left(\widehat{\text{Var}}\left(\widehat{\beta}_i\right)\right)$, where $\widehat{\beta}_i$ is the maximum likelihood estimator of $\beta_i$ and $\widehat{\text{Var}}\left(\widehat{\beta}_i\right)$ is the standard estimate of the variance of $\widehat{\beta}_i$ (Hosmer and Lemeshow, 2000).

Let $g_\delta$ denote the probability density function admitted by the noncentral $\chi^2$ distribution of one degree of freedom and of a noncentrality parameter value $\delta$. For all $i = 1, \ldots, N$, according to Shieh (2005), the Wald test statistic $T_i$ is approximately distributed according to the probability density function $g_{\delta_i}$ with noncentrality parameter

$$\delta_i = \beta_i^2 / \text{Var}\left(\widehat{\beta}_i\right), \tag{1}$$

where $\text{Var}\left(\widehat{\beta}_i\right)$ is the variance of $\widehat{\beta}_i$. The equation indicates that the value of $\delta_i$ is determined by the parameter of interest $\beta_i$ such that $\delta_i = 0$ if $\beta_i = 0$ and $\delta_i > 0$ if $\beta_i \neq 0$. Therefore, the Wald test statistics for all measured SNPs are regarded as reduced data that are highly informative about the parameter of interest. We defined the true association

Table 1: Model types of Section 2.2 and the specific models defined in Section 2.3.

| | Parametric | Semiparametric |
|---|---|---|
| Mixture | PMM | SMM |
| Nonmixture | PNM | N/A |

indicator

$$a_i = \begin{cases} 0 & \text{if } \delta_i = 0 \\ 1 & \text{if } \delta_i > 0 \end{cases} \tag{2}$$

to indicate whether or not the $i$th SNP is associated with a certain disease. Then the proportion of the non-associated SNPs is $p_0 = \#(a_i = 0)/N$, where $\#(a_i = 0)$ is the number of truly non-associated SNPs.

## 2.2 Types of models for estimation

In order to evaluate the strength of evidence and to estimate $a_i$ and $p_0$, the assumed true model of Section 2.1 must be simplified. The main simplifying assumption is that any observed Wald test statistic $t$ for the disease-associated SNPs was sampled from the same distribution. These simplified models, to be specified in Section 2.3, differ in whether they are parametric models and in whether they are mixture models. Table 1 shows the categorization of each model.

### 2.2.1 Parmetric versus semiparametric

For the two parametric models, we assume that the Wald test statistics for the disease-associated SNPs are generated from a noncentral $\chi^2$ distribution of 1 degree of freedom and of a single noncentrality parameter $\delta_i = \delta_{\text{alt}}$ for all $i$ corresponding to disease-associated SNPs. Bukszár *et al.* (2009) also assigned a single noncentrality parameter value to all SNPs not associated with disease. Similarly, for microarray data, a three-component mixture

model performs as well as more complex models according to Muralidharan (2010) with the three components corresponding to null, negative and positive groups. Since the Wald test statistics of Section 2.1 can never be negative, the three components in that microarray model are analogous to the two components of this paper. The two parametric models can also work well with test statistics from other distributions, e.g., the Student $t$ family of distributions used in microarray studies.

For the semiparametric model, the density function of the transformed Wald test statistics for the disease-associated SNPs is estimated nonparametrically. Without modifying the method of estimation, we will use a $\chi^2$ distribution because our test statistics are $\chi^2$-distributed under the null hypothesis.

### 2.2.2 Mixture versus nonmixture

In the two mixture models, for the $i$th SNP, we use the symbol $A_i$ to represent a random association indicator, and we denote the prior probability that the $i$th SNP is not associated with disease by $\pi_0 = P(A_i = 0)$. In the nonmixture model, as in the model assumed to be true (Section 2.1), we use $a_i$ to represent the fixed association indicator, with $a_i = 1$ if the $i$th SNP is associated with a certain disease and with $a_i = 0$ if it is not.

## 2.3 Models for estimation

The following three models will be used to estimate each $a_i$ of the true model specified in Section 2.1. For economy of notation, the same symbol $\widehat{a}_i$ will denote the estimator of $a_i$ for each of the estimating models.

### 2.3.1 Semiparametric mixture model (SMM)

For any observed Wald test statistic $t_i$, $z_i = \Phi^{-1}(G_0(t_i))$ is called the $z$-transformed statistic, where $\Phi$ is the standard normal cumulative distribution function (cdf) and $G_0$

is the cdf of the central $\chi^2$ distribution of 1 degree of freedom. The observed Wald test statistics $t_1, \ldots, t_N$ are transformed into $z_1, \ldots, z_N$, and for the $i$th SNP, the density is a mixture of the form

$$f(z_i) = \pi_0 f_0(z_i) + (1 - \pi_0) f_1(z_i), \tag{3}$$

where $f_0$ is the density function of $z$ for the non-associated SNPs, and $f_1$ is that for the disease-associated SNPs.

For the $i$th SNP, LFDR is the posterior probability

$$\mathrm{LFDR}_{\mathrm{SMM}}(z_i) = \mathrm{P}(A_i = 0 | z_i) = \frac{\pi_0 f_0(z_i)}{f(z_i)}, \tag{4}$$

which may be interpreted as a level of certainty in the hypothesis that the $i$th SNP is not associated with disease ($A_i = 0$) and is associated with disease ($A_i = 1$). The density function $f_0$ is assumed to be standard normal, $N(0,1)$, called the *theoretical null*. Another approach is to estimate the null density (Efron, 2004, 2007a), but such estimation unnecessarily increases the variance of the estimates of interest for the application of this paper, as explained in Section 3. Following Efron (2004), we use a nonparametric method to estimate $f$ by $\widehat{f}$ and an empirical Bayes method to estimate $\pi_0$ by $\widehat{\pi}_0$, and then compute $\widehat{\mathrm{LFDR}}_{\mathrm{SMM}}(z_i)$ by substituting $\widehat{\pi}_0$ and $\widehat{f}$ into $\pi_0$ and $f$ in equation (4). The natural estimate of $f_1$ is then $\widehat{f}_1 = \left(\widehat{f} - \widehat{\pi}_0 f_0\right) / (1 - \widehat{\pi}_0)$. We choose the estimator $\widehat{a}_i = 1 - \widehat{\mathrm{LFDR}}_{\mathrm{SMM}}(z_i)$ to estimate $a_i$ of equation (2).

### 2.3.2 Parametric mixture model (PMM)

For the $i$th SNP, the density $g(t_i)$ is defined by replacing $f(z_i)$, $f_0(z_i)$ and $f_1(z_i)$ in equation (3) with $g(t_i)$, $g_0(t_i)$ and $g_{\delta_{\mathrm{alt}}}(t_i)$. The log-likelihood of PMM with unknown parameters $\pi_0$ and $\delta_{\mathrm{alt}}$ is thus

$$\log L\left(\pi_0, \delta_{\text{alt}}\right) = \sum_{i=1}^{N} \log\left[\pi_0 g_0\left(t_i\right) + \left(1-\pi_0\right) g_{\delta_{\text{alt}}}\left(t_i\right)\right],$$

which is maximized numerically by the *maximum likelihood estimates* $\widehat{\pi}_0$ and $\widehat{\delta}_{\text{alt}}$. The $g_{\widehat{\delta}_{\text{alt}}}$ is an estimate of $g_{\delta_{\text{alt}}}$, and $\widehat{g}$ is the corresponding estimate of $g$. For the $i$th SNP, $\widehat{a}_i = 1 - \widehat{\text{LFDR}}_{\text{PMM}}\left(t_i\right)$ estimates $a_i$, where $\widehat{\text{LFDR}}_{\text{PMM}}\left(t_i\right) = \widehat{\pi}_0 g_0\left(t_i\right) / \widehat{g}\left(t_i\right)$ is the estimator of the LFDR.

Under the two parametric models, the log-likelihood ratio for the $i$th SNP,

$$\Delta_i\left(t_i\right) = \log_2\left(g_{\delta_{\text{alt}}}\left(t_i\right) / g_0\left(t_i\right)\right) \tag{5}$$

is regarded as the ideal *information for discrimination* favoring the hypothesis of association ($\delta_i = \delta_{\text{alt}}$) over that of non-association ($\delta_i = 0$) (Kullback, 1968). The information is called *ideal* because the $\delta_{\text{alt}}$ is unknown. Here we chose the binary logarithm ($\log_2$) to facilitate interpretation. Following the evidence levels of that Bickel (2010c) considered, the discrimination information indicates strong evidence ($\Delta_i\left(t_i\right) > 3$), very strong evidence ($\Delta_i\left(t_i\right) > 5$) and overwhelming evidence ($\Delta_i\left(t_i\right) > 7$); see Royall (1997) for slightly different names of the first two grades. Negative discrimination information, $\Delta_i\left(t_i\right) < 0$, indicates evidence in favor of non-association, which cannot be integrated by p-values since they can only quantify the evidence against non-association (Wei *et al.*, 2010).

For the $i$th SNP, the *approximate information for discrimination* is obtained by substituting $g_{\widehat{\delta}_{\text{alt}}}\left(t_i\right)$ into $g_{\delta_{\text{alt}}}\left(t_i\right)$ in equation (5). That approximation also closely approximates an MDL-based information for discrimination (Bickel, 2010b).

### 2.3.3 Parametric nonmixture model (PNM)

Under the assumption that for the $i$th SNP, the observed Wald test statistic $t_i$ is sampled from $g_0(t_i)$ if $a_i = 0$ or from $g_{\delta_{\text{alt}}}(t_i)$ if $a_i = 1$, the density of $t_i$ is

$$g(t_i, \delta_{\text{alt}}, a_i) = a_i g_0(t_i) + (1 - a_i) g_{\delta_{\text{alt}}}(t_i), \tag{6}$$

and the log likelihood with the unknown parameters $\delta_{\text{alt}}$ and $a_1, \ldots, a_N$ is

$$\log L(\delta_{\text{alt}}, a_{1,\ldots}, a_N) = \sum_{i=1}^{N} \log\left(g(t_i, \delta_{\text{alt}}, a_i)\right). \tag{7}$$

The maximum likelihood estimates of the parameters $a_1, \ldots, a_N$ are computed by substituting

$$\widehat{a}_i^{\text{naive}} = \begin{cases} 0 & g_{\widetilde{\delta}_{\text{alt}}}(t_i) \leqslant g_0(t_i) \\ 1 & g_{\widetilde{\delta}_{\text{alt}}}(t_i) > g_0(t_i) \end{cases}, \tag{8}$$

into $a_i$ in (7), where $\widetilde{\delta}_{\text{alt}}$ is the value of $\delta_{\text{alt}}$ that maximizes the likelihood defined by equation (7). The estimated density functions $\widetilde{g}$ and $g_{\widetilde{\delta}_{\text{alt}}}$ are obtained by substituting $\widehat{a}_i^{\text{naive}}$ and $\widetilde{\delta}_{\text{alt}}$ for $a_i$ and $\delta_{\text{alt}}$ in equation (6).

For the $i$th SNP, the maximum likelihood estimator $\widehat{a}_i^{\text{naive}}$ in equation (8) is a naive estimator for $a_i$ in that the decisions based on $\widehat{a}_i^{\text{naive}}$ do not perform well due to its binary nature. We use another estimator of $a_i$, $\widehat{a}_i = 1 - \widehat{p}_0 g_0(t_i) / \widetilde{g}_i(t_i)$, where $\widehat{p}_0 = \#\left(\widehat{a}_i^{\text{naive}} = 0\right)/N$ is an estimator of $p_0$ (Bickel, 2010b).

For the $i$th SNP, the approximate information for discrimination is computed by substituting $g_{\widetilde{\delta}_{\text{alt}}}(t_i)$ into $g_{\delta_{\text{alt}}}(t_i)$ in equation (5) (Bickel, 2010b).

# 3 Application

The Coronary artery disease (CAD) data from Wellcome Trust Case Control Consortium (2007) included 490,032 SNPs genotyped for 1988 cases and 3004 combined controls on 22 autochromosomes. We excluded SNPs with 10% missing genotypes, with minor allele frequencies <5% and with p-values smaller than 0.05 from a Hardy-Weinberg equilibrium exact test (Purcell *et al.*, 2007). We also excluded the samples genotyped for less than 95% of the SNPs. A total of $N = 335{,}169$ SNPs and 4934 samples (1947 cases, 2987 controls) passed those quality control filters and were used in the following statistical analysis.

We computed the Wald test statistics for all $N$ SNPs. In order to implement SMM, we transformed the observed Wald test statistics into $z$-values. Since the central region of the histogram of the $z$-values matches $N(0,1)$ very well, $f_0$ is considered to be the theoretical null. For PMM and PNM, the observed statistics were used to estimate the model parameters. These estimates are shown in Table 2. The estimates of $p_0$ under SMM and PMM were denoted by $\widehat{\pi}_0$ in Section 2.3, and SMM does not estimate $\delta_{\text{alt}}$ since instead estimates $f_1$ using the nonparametric method mentioned in Section 2.

We now introduce some additional notation to compare the three models with respect to the CAD data under the assumed true model of Section 2.1, in which the Wald test statistic for each disease-associated SNP is sampled from a noncentral $\chi^2$ distribution of 1 degree of freedom and of a potentially different noncentrality parameter value. The *average probability density function* of the Wald test statistics for all disease-associated SNPs, represented by $\overline{f}_1$, is defined as the mean of the probability density function of each of its statistics. The estimated $\overline{f}_1$ under SMM is equal to $\widehat{f}_1$. For the two parametric models, $\overline{f}_1$ was estimated by $g_{\widehat{\delta}_{\text{alt}}}$ under PMM and by $g_{\widetilde{\delta}_{\text{alt}}}$ under PNM. In Figure 1, the solid line is the estimated $f_1$ under SMM, where the $z$-values are the transformed Wald test statistics from the CAD data. For PMM and PNM, we simulated 500,000 Wald test

Table 2: The estimation of parameters $p_0$ and $\delta_{\text{alt}}$ in SMM, PMM and PNM in the CAD data.

| Models | Estimated $p_0$ | Estimated $\delta_{\text{alt}}$ |
|---|---|---|
| SMM | 0.98 | N/A |
| PMM | 0.90 | 1.01 |
| PNM | 0.77 | 3.11 |

Table 3: The number of SNPs with evidence in favor of non-association (the negative approximate information for discrimination) and with very strong evidence (approximate information for discrimination larger than 5) and overwhelming evidence (approximate information for discrimination larger than 7) in favor of association in the CAD data.

| | Negative | Very strong | Overwhelming |
|---|---|---|---|
| PMM | 234,411 | 35 | 19 |
| PNM | 259,343 | 779 | 82 |

statistics $t$ from $g_{\widehat{\delta}_{\text{alt}}}$ and $g_{\widetilde{\delta}_{\text{alt}}}$ respectively, where the values of $\widehat{\delta}_{\text{alt}}$ and $\widetilde{\delta}_{\text{alt}}$ are from Table 2. Then we transformed the $t$ values into $z$ values and to numerically approximate each density function of $z$ values. All three estimated density functions in Figure 1 were normalized into the interval between 0 and 1.

The dark grey horizontal line illustrates the full width at half maximum (FWHM) for the estimated $\overline{f}_1$ under three models. Since there is more than one noncentrality parameter value under the true model, $\overline{f}_1$ is wider than the probability density function of each of its SNPs. However, as seen in Figure 1, the FWHM for the SMM-estimated density is much smaller than those of the other two estimated densities, each of which is in turn narrower than $\overline{f}_1$.

Figure 2 displays the approximate discrimination information (Section 2.3) for all measured SNPs, where the estimated odds ratio for the $i$th SNP is $e^{\widehat{\beta}_i}$. Table 3 reports the numbers of SNPs with very strong or overwhelming evidence in favor of association and of those with evidence in favor of non-association.

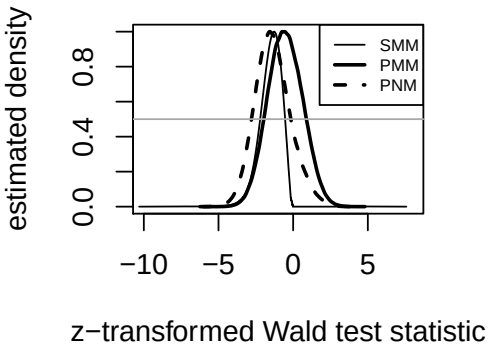


Figure 1: The estimated density functions of the z-transformed Wald test statistics for the disease-associated SNPs in the CAD data under three models.

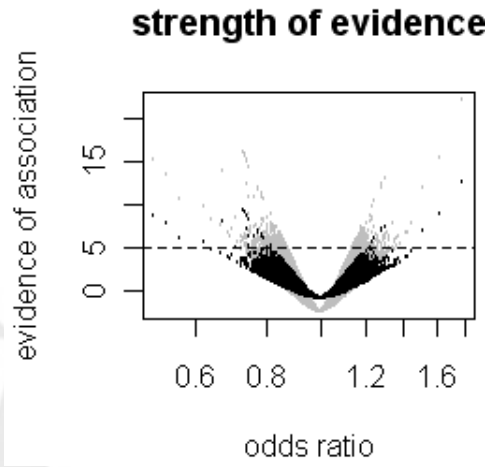


Figure 2: The binary logarithm ($\log_2$) of the likelihood ratio versus the estimated odds ratio for all measured SNPs in the CAD data under PMM (black) and PNM (grey).

# 4 Simulation studies

The aim of the following simulation studies is to compare the performances of the estimators under the three models of Section 2. First, we used the Kullback-Leibler divergence to compare the deviations between the true distribution and the estimated distribution for the disease-associated SNPs in the two parametric models (PMM and PNM). Second, we compared the abilities of all three models to estimate $p_0$ and $a_i$.

All simulations were performed under the assumed true model of Section 2.1. According to equation (1), each $\delta_i$ depends on the logarithm of the odds ratio, and SNPs with different values of the odds ratio do not necessarily have the same value of $\delta_i$. The simulation studies are divided into two sets, one with multiple values of $\delta_{\text{alt}}$ and the other with a single value of $\delta_{\text{alt}}$ for each study. In the former set, we assume that the observed Wald test statistics for the disease-associated SNPs between disease-associated genes are generated from different distributions with different values of $\delta_{\text{alt}}$ (Section 4.1). In the latter set, the observed Wald test statistics for all disease-associated SNPs are sampled from the same distribution with a single value of $\delta_{\text{alt}}$ (Section 4.2). In both sets, we assume that the observed Wald test statistics for non-associated SNPs are from the same distribution with $\delta_i = 0$. Bukszár *et al.* (2009) similarly simulated test statistics but used different estimators and different measures of performances than those employed herein.

## 4.1 Multiple values of $\delta_{\text{alt}}$ per study

The data were simulated as follows. There are 2000 effect-size groups affecting a certain disease, and the number of SNPs in each effect-size group is evenly allocated. We performed 12 simulation studies, each with a different value of $p_0$ and with the number of the SNPs per effect-size group equal to one of the two integers closest to $(1 - p_0)N/2000$, where in each simulation study, $N$ is equal to 335,169, the total number of the mea-

sured SNPs in Section 3. For each of the 2000 disease-associated effect-size groups, a noncentrality parameter $\delta_j > 0$ was assigned. For the first two effect-size groups, $\delta_1 = \widehat{\delta}_{\text{alt}} = 1.01$ and $\delta_2 = \widetilde{\delta}_{\text{alt}} = 3.11$, numbers that also appear in Table 2, whereas for each $j = 3, \ldots, 2000$, $\delta_j$ was generated from the uniform distribution between 0.5 and 5.

In each simulation study, we randomly generated 50 data sets, each corresponding to an artificial case-control study. For each data set, we used the true model of Section 2 to simulate the Wald test statistics, where the Wald test statistics for the SNPs within the $j$th effect-size group associated with disease were sampled from the noncentral $\chi^2$ distribution of 1 degree of freedom and of a noncentrality parameter $\delta_j$, and the Wald test statistics for the other non-associated SNPs were sampled from the noncentral $\chi^2$ distribution of 1 degree of freedom and of the noncentrality parameter equal to 0.

The Kullback-Leibler (KL) divergence is the expected difference between the ideal information for discrimination and the approximate information for discrimination defined in Section 2.3.2, e.g., for the $i$th SNP,

$$\text{KL}_i = \text{E}\left[\log_2\left(g_{\delta_{\text{alt}}}(t_i)/g_{\widehat{\delta}_{\text{alt}}}(t_i)\right)\right] = \text{E}\left[\Delta_i(t_i) - \widehat{\Delta}_i(t_i)\right],$$

where $\widehat{\Delta}_i(t_i) = \log_2\left(g_{\widehat{\delta}_{\text{alt}}}(t_i)/g_0(t_i)\right)$ is the approximate information for discrimination under PMM. Here, $g_{\widehat{\delta}_{\text{alt}}}$ is fixed, and the expectation value is over $T_i$, which is distributed according to $g_{\delta_{\text{alt}}}$. This KL divergence measures the performance of approximating the ideal information for discrimination. The left panel of Figure 3 displays the numerically approximated KL divergence for each simulation study averaged over all 50 data sets and over all the disease-associated SNPs.

The bias for each estimator of $p_0$ is $\text{E}[\widehat{p}_0 - p_0]$. The left panel of Figure 4 shows the bias estimated by averaging $\widehat{p}_0 - p_0$ over all 50 data sets. The estimated standard deviation of

$\widehat{p}_0$ in each simulation study is <0.001% for all three models.

For the $i$th SNP, the frequentist risk for $\widehat{a}_i$ is $R\left(a_i, \widehat{a}_i\right)$= E$[l\left(a_i, \widehat{a}_i\right)]$, the expected loss, where $\widehat{a}_i$ is the estimate of $a_i$ defined for each model in Section 2.3 and $l\left(a_i, \widehat{a}_i\right)$ is the loss of $\widehat{a}_i$. In this paper, we consider the most widely used loss functions used to evaluate the performance of probability assessments (Bernardo and Smith, 1994), namely the quadratic loss $l\left(a_i, \widehat{a}_i\right) = \left(\widehat{a}_i - a_i\right)^2$, and the logarithmic loss

$$l\left(a_i, \widehat{a}_i\right) = \begin{cases} -\log_2\left(1 - \widehat{a}_i\right) & a_i = 0 \\ -\log_2 \widehat{a}_i & a_i = 1 \end{cases}.$$

The expected quadratic loss and the expected logarithmic loss are called the *mean square error* (MSE) and the *logarithmic risk*, respectively. In Figure 5, the infinite logarithmic risk under SMM is due to the infinite logarithmic risk for the disease-associated SNPs since we have infinite loss for the disease-associated SNPs whenever $\widehat{\mathrm{LFDR}}_{\mathrm{SMM}}\left(z_i\right)$ is equal to 100%. In the Bayesian interpretation, this would mean absolute certainty that a particular SNP is not associated with disease even though it is associated with disease (cf. Bickel, 2010a).

In the simulation studies above, we assumed for simplicity that there are 2000 effect-size groups associated with disease and the SNPs within a effect-size group are independent. In order to investigate whether the results are sensitive to the number of the disease-associated effect-size groups and to the correlation between SNPs within a effect-size group, we also performed additional simulation studies as before except that the number of the disease-associated effect-size group varied between 10 and 4000. For each of the disease-associated groups, we considered the extreme case that the SNPs within a effect-size group have the same Wald test statistics regarding the disease-associated group as a haplotype. The results are not shown here since they do not affect the performance of the estimators relative to each other.

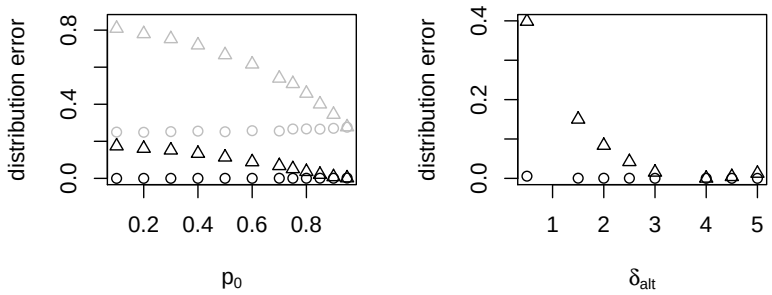


Figure 3: Left panel: the mean Kullback–Leibler divergence between $g_{\delta_{\text{alt}}}$ and its estimator versus $p_0$ for two true values of $\delta_{\text{alt}}$ equal to 1.01 (grey) and 3.11 (black) under PMM (circles) and under PNM (triangles) when each simulation study has multiple values of $\delta_{\text{alt}}$ (Section 4.1). Right panel: the mean Kullback–Leibler divergence between $g_{\delta_{\text{alt}}}$ and its estimator versus the true values of $\delta_{\text{alt}}$ under PMM (circles) and under PNM (triangles) when each simulation study has a single value of $\delta_{\text{alt}}$ (Section 4.2).

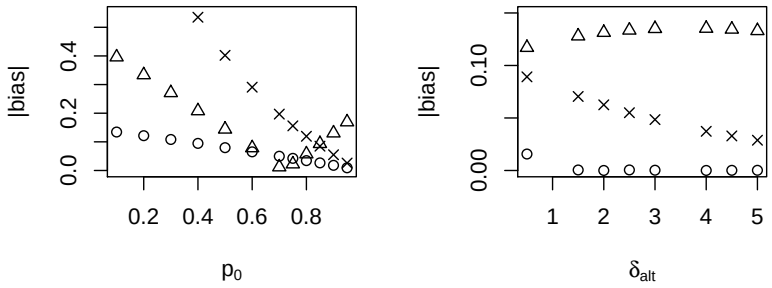


Figure 4: The same as Figure 3 except that the absolute values of the estimated bias of each estimator of $p_0$ replace the mean Kullback–Leibler divergence between $g_{\delta_{\text{alt}}}$ and its estimator for all three models. In both panels, the estimated bias of each estimator of $p_0$ is positive under SMM("x"), positive under PMM (circles) and negative under PNM (triangles).

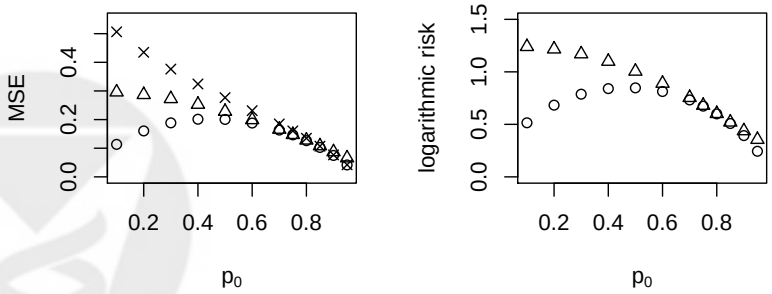


Figure 5: The approximate risk of $\widehat{a}_i$ versus $p_0$ when each simulation study has multiple values of $\delta_{\text{alt}}$ (Section 4.1) under three models, SMM ("x"), PMM (circles) and PNM (triangles). Left panel: the approximate mean square error (MSE) of $\widehat{a}_i$ versus $p_0$. Right panel: the approximate logarithmic risk of $\widehat{a}_i$ versus $p_0$. Under SMM, the approximate logarithmic risk of $\widehat{a}_i$ is infinite and is not displayed.

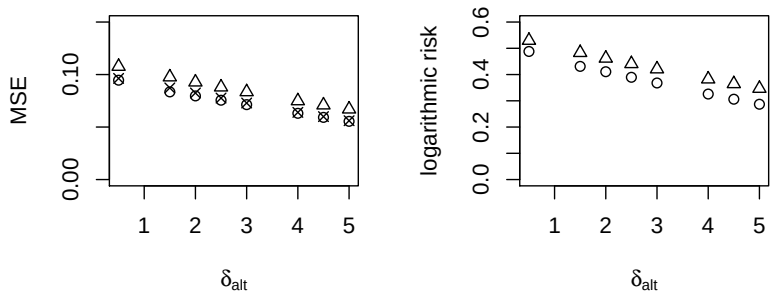


Figure 6: The same as Figure 5 except that the each simulation study has a single value of $\delta_{\text{alt}}$ (Section 4.2).

## 4.2 One value of $\delta_{\text{alt}}$ per study

In this set of simulation studies, the true value of $p_0$ is fixed at 0.90, the PMM estimate reported in Table 2. Since the total number of markers is $N$ = 335,169, the number of disease-associated SNPs is equal to 33,516. We performed 8 simulation studies, each with a different value of $\delta_{\text{alt}}$, ranging from 0.5 to 5. Similar to Section 4.1, we simulated 50 data sets for each simulation study, and the results were plotted versus the true values of $\delta_{\text{alt}}$ instead of the true values of $p_0$ in the right-hand panels of Figure 3 and 4. Figure 6 displays the approximate risk of $\widehat{a}_i$ versus $\delta_{\text{alt}}$ for each simulation study.

We also conducted the simulation studies with the true value of $p_0$ fixed at 0.77, the PNM estimate (Table 2). The results are not displayed here since the relative performances of the estimators associated with the three models are the same as that in the case of $p_0$ fixed at 0.90.

# 5 Discussion

As observed in Section 3, the SMM estimate of $f_1$ is much narrower than the true $\overline{f}_1$, which indicates that using the nonparametric method to estimate the density function for the disease-associated SNPs led to overfitting the CAD data. Such problems in estimating the mean density function propagate to the estimates of the true association indicator $a_i$ and

of the true null proportion $p_0$. While the parametric models may underfit the data, the simulation studies of Section 4.1 demonstrated that they are robust against misspecification in the form of multiple noncentrality parameters values.

Figure 4 indicates that SMM has a severe upward bias in estimating $p_0$, as previously observed in microarray studies using similar semiparametric models (Pawitan *et al.*, 2005). As Pawitan *et al.* (2005) explained, that bias reflects the fact that nonparametric entity estimators inadequately separate the features with small effect size from those with no effect. The biased estimate of $p_0$ will lead to a biased estimate of LFDR according to equation (4). Using the biased estimate of LFDR in turn results in failing to detect associations with SNPs of small odds ratios.

As a consequence of this bias, in both sets of simulation studies, the approximate logarithmic risk for the disease-associated SNPs under SMM is infinite as a result of the infinite loss for some of the disease-associated SNPs under SMM. Here, the infinite loss is due to the value of the estimated LFDR equalling 100% for a specific disease-associated SNP, which indicates that the LFDR estimator based on SMM does not have the Bayesian posterior probability interpretation that popularized the LFDR.

Also due to this bias in estimating $p_0$, the investigator cannot interpret a high LFDR estimate as evidence in favor of the null hypothesis. Both PMM and PNM can measure the strength of evidence for the hypothesis of non-association (Table 3), but the LFDR estimator based on SMM, similar to conventional p-values, cannot quantify the amount of information in the data supporting non-association.

Among the two parametric models, PMM has substantially better performance for estimating the true values of $\delta_{\text{alt}}$ and $p_0$ and has slightly better performance for estimating $a_i$ (Figures 3, 4, 5 and 6).

# Acknowledgements

Y.Y. thanks Corey Yanofsky for helpful discussions and for assistance with `R` syntax. The `Biobase` (Gentleman *et al.*, 2004) and `locfdr` (Efron, 2007b) packages of R (R Development Core Team, 2008) facilitated the computational work. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada, by the Canada Foundation for Innovation, by the Ministry of Research and Innovation of Ontario, and by the Faculty of Medicine of the University of Ottawa. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

# References

Allison, D. B., Gadbury, G. L., Heo, M., Fernádez, J. R., Lee, C., Prolla, T. A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, **38**(5), 1–20.

Aubert, J., Bar-Hen, A., Daudin, J., and Robin, S. (2004). Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics*, **5**.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley and Sons, New York.

Bickel, D. R. (2010a). Estimating the null distribution to adjust observed confidence levels for genome-scale screening. *Biometrics, DOI: 10.1111/j.1541-0420.2010.01491.x*.

Bickel, D. R. (2010b). Minimum description length methods of medium-scale simultaneous inference. *Technical Report, Ottawa Institute of Systems Biology, arXiv:1009.5981*.

Bickel, D. R. (2010c). Statistical inference optimized with respect to the observed sample for single or multiple comparisons. *Technical Report, Ottawa Institute of Systems Biology, arXiv:1010.0694*.

Bickel, D. R. (2010d). The strength of statistical evidence for composite hypotheses: Inference to the best explanation. *Technical Report, Ottawa Institute of Systems Biology, COBRA Preprint Series, Article 71, available at biostats.bepress.com/cobra/ps/art71*.

Blume, J. D. (2002). Likelihood methods for measuring statistical evidence. *Statistics in Medicine*, **21**(17), 2563–2599.

Bukszár, J., McClay, J. L., and van den Oord, E. J. C. G. (2009). Estimating the posterior probability that genome-wide association findings are true or false. *Bioinformatics*, **25**(14), 1807–1813.

Edwards, A. W. F. (1969). Statistical methods in scientific inference. *Nature*, **222**(5200), 1233–1237.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, **99**(465), 96–104.

Efron, B. (2007a). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, **102**(477), 93–103.

Efron, B. (2007b). Size, power and false discovery rates. *Annals of Statistics*, **35**, 1351–1377.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**(456), 1151–1160.

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.

Göing, H. H. H., Terwilliger, J. D., and Blangero, J. (2001). Large upward bias in estimation of locus-specific effects from genomewide scans. *American Journal of Human Genetics*, **69**(6), 1357–1369.

Greenwood, C. M. T., Rangrej, J., and Sun, L. (2007). Optimal selection of markers for validation or replication from genome-wide association studies. *Genetic epidemiology*, **31**(5), 395–407.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. The MIT Press, London.

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge University Press, Cambridge.

He, Y., Pan, W., and Lin, J. (2006). Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data. *Computational Statistics and Data Analysis*, **51**(2), 641–658.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, **6**(2), 95–108.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression*. John Wiley and Sons, New York.

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, **437**(7063), 1299–1320.

Kullback, S. (1968). *Information Theory and Statistics*. Dover, New York.

Liao, J. G., Lin, Y., Selvanayagam, Z. E., and Shih, W. J. (2004). A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics*, **20**(16), 2694–2701.

McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R., Hinds, D. A., Pennacchio, L. A., Tybjaerg-Hansen, A., Folsom, A. R., Boerwinkle, E., Hobbs, H. H., and Cohen, J. C. (2007). A common allele on chromosome 9 associated with coronary heart disease. *Science*, **316**(5830), 1488–1491.

Montana, G. (2006). Statistical methods in genetics. *Briefings in Bioinformatics*, **7**(3), 297–308.

Muralidharan, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *Annals of Applied Statistics*, **4**, 422–438.

Pan, W., Lin, J., and Le, C. T. (2003). A mixture model approach to detecting differentially expressed genes with microarray data. *Functional and Integrative Genomics*, **3**(3), 117–124.

Pawitan, Y., Murthy, K., Michiels, S., and Ploner, A. (2005). Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics*, **21**(20), 3865–3872.

Pounds, S. and Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, **19**(10), 1236–1242.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**(3), 559–575.

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

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. CRC Press, New York.

Sabatti, C., Service, S., and Freimer, N. (2003). False discovery rate in linkage and association genome screens for complex disorders. *Genetics*, **164**(2), 829–833.

Shieh, G. (2005). On power and sample size calculations for Wald tests in generalized linear models. *Journal of Statistical Planning and Inference*, **128**(1), 43–59.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **64**(3), 479–498.

Wacholder, S., Chanock, S., Garcia-Closas, M., Ghormli, L. E., and Rothman, N. (2004). Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, **96**(6), 434–442.

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *American Journal of Human Genetics*, **81**(2), 208–227.

Wei, Y., Wen, S., Chen, P., Wang, C., and Hsiao, C. K. (2010). A simple Bayesian mixture model with a hybrid procedure for genome-wide association studies. *European Journal of Human Genetics*, **18**(8), 942–947.

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**(7145), 661–678.

Ziegler, A., I.R., K., and Thompson, J. (2008). Biostatistical aspects of genome-wide association studies. *Biometrical Journal*, **50**(1), 8–28.