Incorporation of Heterogeneity in a Case-Control Study Through a Mixture Model

Subrata Paul^{*1} and Stephanie A. Santorico^{†2}

¹Institute for Behavioral Genetics, University of Colorado Boulder

^{1,2}Department of Mathematical and Statistical Sciences, University of Colorado
Denver

²Human Medical Genetics and Genomics Program, University of Colorado Denver

²Department of Biostatistics & Informatics, University of Colorado Denver

September 9, 2020

Abstract

Most common human diseases and complex traits are etiologically heterogeneous. Genome-wide Association Studies (GWAS) aim to discover common genetic variants that are associated with complex traits, typically without considering heterogeneity. Heterogeneity, as well as imprecise phenotyping, significantly reduces the power to find genetic variants associated with human diseases and complex traits. Disease subtyping through unsupervised clustering techniques such as latent class analysis can explain some of the heterogeneity; however, subtyping methods do not typically incorporate heterogeneity into the association framework. Here, we use a finite mixture model with logistic regression to incorporate heterogeneity into the association testing framework for a case-control study. In the proposed method, the disease outcome is modeled as a mixture of two binomial distributions. One of the component distributions refers to the subgroup of the population for which the genetic variant is not associated with the disease outcome and another component distribution corresponds to the subgroup for which the genetic variant is associated with the disease outcome. The mixing parameter corresponds to the proportion of the population for which the genetic variant is associated with the disease outcome. A simulation study of a trait with differing levels of prevalence, SNP minor allele frequency, and odds ratio was performed, and effect size estimates compared between the models with and without incorporating heterogeneity. The proposed mixture model yields lower bias of odds ratios while having comparable power compared to classical logistic regression.

Introduction

Most human diseases, such as diabetes, cardiovascular disease, and cancer, are complex traits. Complex traits, being the result of many genetic variants, their interactions, and interaction with

^{*}subrata.paul@colorado.edu

[†]stephanie.santorico@ucdenver.edu

environmental factors, are heterogeneous with respect to disease pathophysiology; however, it is not generally easy to determine the heterogeneity based on phenotypic symptoms. To illustrate how heterogeneity affects an association study, suppose, in a case-control study we want to test whether a SNP, S , is associated with a disease trait, T , using a sample of 1000 cases and 50,000 controls. Assume that the disease has two underlying subtypes with 100 cases corresponding to type 1 and the remainder corresponding to type 2. If the SNP S is highly associated with disease type 1 but not with type 2, we are likely to be underpowered to detect an association based on the heterogeneous mixture of cases. In such situations the effect sizes are underestimated in magnitude. This simple example of heterogeneity is similar to phenotypic misclassification. Phenotypic misclassification can occur in two ways: a case can be misclassified as a control or a control can be misclassified as a case. For example, the Prostate Specific Antigen (PSA) test is used to diagnose prostate cancer. Not all men who have prostate cancer show elevated PSA, which results in a case being misclassified as a control (Gordon and Finch, 2005). Lack of consideration of underlying subtypes leads to imprecise phenotyping. Prior work has shown that imprecise phenotyping significantly reduces the effect sizes for models of genetic association and contributes to the missing heritability problem (Sluis et al., 2010).

One way to address phenotypic heterogeneity is disease subtyping using unsupervised statistical techniques with phenotypic information such as symptoms or using one or multiple types of omics data. Existing approaches for disease subtyping fall roughly into three categories. Approaches in the first category use phenotypic data and are highly applicable to psychiatric disorders. Various clustering processes such as Latent Class Analysis (LCA) (Grados et al., 2008; Prust et al., 2011) and finite mixture models (Wessman et al., 2009) are used based on secondary phenotypic measures on disease cases. The second category of approaches utilize one type of genomic data, such as gene expression or epigenetic data, and are applicable to a wide variety of diseases (Bair and Tibshirani, 2004; Pyatnitskiy et al., 2014). These approaches utilize clustering algorithms and variable selection techniques. The third category of approaches uses multiple types of genomic data, such as gene expression, copy number variation, and DNA methylation. Various statistical tools are used, including models implemented in a Bayesian framework as well as clustering techniques. Examples of these techniques include Shen et al. (2012) and Kristensen et al. (2014). None of the techniques incorporate heterogeneity into an association testing framework; however, doing so, could improve effect size estimation and provide inferences on heterogeneity.

The finite mixture model is a classical statistical technique to model population heterogeneity. Since the work of Pearson (1894), the finite mixture model has been widely used as a flexible method of modeling heterogeneity in fields of study such as medicine, genetics, biology, astronomy, and social sciences. Many well-developed statistical techniques in these fields, e.g., latent class analysis, cluster analysis, discriminant analysis, image analysis, and survival analysis, are underpinned by finite mixture model theory. The finite mixture model is ideal in situations where covariates that characterize the distribution of the variable under consideration are missing because they were not observed or measured during the data collection process. Typically, the finite mixture model is used to model data that arises from a heterogeneous population. When the population can be divided into some number of sub-populations, and for each sub-population the data can be modeled by a parametric distribution, then their mixture gives the marginal distribution for the population. Here we implement a finite mixture of logistic regression models to incorporate heterogeneity into an association framework. Simulation studies are used to show that the effect size estimates have lower bias compared to classical logistic regression while maintaining comparable power.

1 Methods

1.1 Finite Mixture Model and EM algorithm

Let $Y_1, \dots, Y_n$ be a random sample of size n with observed values $y_1, \dots, y_n$. Random variable Y_i arises from a finite mixture distribution if the density function can be written in the form,

$$f(y_i; \boldsymbol{\theta}) = \sum_{k=1}^K \eta_k f_k(y_i; \boldsymbol{\xi}_k) \quad (1)$$

where $f_k(y_i; \boldsymbol{\xi}_k)$ are parametric density functions with unknown parameters $\boldsymbol{\xi}_k$, η_k satisfy $0 \leq \eta_k \leq 1$ for all k with $\sum_{k=1}^K \eta_k = 1$, and $\boldsymbol{\theta} = (\eta_1, \dots, \eta_K, \boldsymbol{\psi}^T)^T$ where the vector $\boldsymbol{\psi}$ contains all the parameters in $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_K$ assumed a priori to be distinct. The quantities η_k are called mixing proportions, and $f_k(y_i; \boldsymbol{\xi}_k)$ are component densities. The expectation maximization (EM) algorithm is an iterative process to approximate the maximum likelihood estimates of the model parameters that depends on unobserved variables. Each iteration of the EM algorithm consists of 2 steps: the expectation (E) step and the maximization (M) step. In the E-step a function is created for the expectation of the log-likelihood evaluated at the current estimates of the parameters, and in the M-step the parameters are estimated by maximizing the function from the E-step. Consider the mixture model (1). In the finite mixture model the observed data $\mathbf{y} = (y_1, \dots, y_n)^T$ is viewed as being incomplete as the vectors of component-labels, $\mathbf{z}_1, \dots, \mathbf{z}_n$, are missing, where $\mathbf{z}_i$ is a K -dimensional vector with $z_{ki} = 1$ or 0, according to whether y_i did or did not arise from the k th component. The vectors $\mathbf{z}_1, \dots, \mathbf{z}_n$ are taken to be the realized values of multinomial random variables, $\mathbf{Z}_1, \dots, \mathbf{Z}_n$. The log-likelihood of the complete data, $\mathbf{y}_c = (\mathbf{y}^T, \mathbf{z}^T)$, is given by

$$\ell_c(\boldsymbol{\theta}) = \sum_{k=1}^K \sum_{i=1}^n z_{ki} \{\log \eta_k + \log f_k(y_i; \boldsymbol{\xi}_k)\}. \quad (2)$$

The EM algorithm based on the complete data log-likelihood (2) is described in (Aitkin and Wilson, 1980) and also available in most finite mixture model textbooks such as McLachlan and Peel (2004). A general schematic of the EM algorithm is given in Figure 1. There are several different methods for choosing the initial values of the parameters as discussed in Karlis and Xekalaki (2003). The stopping criteria are generally based on the difference in likelihood between two consecutive iterations or the Aitken acceleration (McLachlan and Peel, 2004).

1.2 Bernoulli Mixture Model

Let Y_i and G_i be the disease status and minor allele count (0, 1, or 2) at a bi-allelic genetic locus under consideration for the i th observation where, $1 \leq i \leq n$. Under an additive model, the logistic regression model is given by,

$$\text{logit}(P(Y_i = 1|G_i)) = \mu + \beta G_i$$

where, μ and β are regression coefficients. Another form of the model equation is,

$$P(Y_i = y|G_i) = \frac{\exp((\mu + \beta G_i)y)}{1 + \exp(\mu + \beta G_i)}, \quad y \in \{0, 1\}.$$

Figure 1: EM algorithm for a finite mixture model.

The coefficient β is known as the effect size of the locus. Under the hypothesis that the genetic variant is associated with disease for only a part of the population, the disease outcome can be modeled as a finite mixture distribution. We model the distribution of Y_i as a mixture of two component distributions:

$$P(Y_i = y|G_i) = (1 - \alpha) \underbrace{\frac{\exp(\mu_1 y)}{1 + \exp(\mu_1)}}_{\text{Component 1}} + \alpha \underbrace{\frac{\exp((\mu_2 + \beta G_i)y)}{1 + \exp(\mu_2 + \beta G_i)}}_{\text{Component 2}}, \quad y \in \{0, 1\} \quad (3)$$

where, μ_1 , μ_2 , and β are regression coefficients and α is the mixing parameter that represents the proportion of the population for which the genetic variant is associated with disease. Component 1 and component 2 correspond to the partitions of the population at which the genetic variant has zero and non-zero effect, respectively. Model (3) is a mixture of a generalized linear model. Not all such models are identifiable. For models that lack identifiability, the estimation procedure may not be well-defined and asymptotic theory may not hold (McLachlan et al., 2019). Note that, for model (3), we can establish only one equation which has four unknown parameters:

$$P(Y_i = 1) = (1 - \alpha) \frac{\exp(\mu_1)}{1 + \exp(\mu_1)} + \alpha \frac{\exp(\mu_2 + \beta G_i)}{1 + \exp(\mu_2 + \beta G_i)}.$$

Hence model (3) is not identifiable since different values of the parameters μ_1 , μ_2 , α , and β result in the same probability distribution of Y . The identifiability problem can be resolved by considering ancestry clusters and modeling the number of diseased individuals in an ancestry cluster to be a binomial random variable. The R package "flexmix" (Grün and Leisch, 2007; Leisch, 2004) is readily available to fit a mixture of generalized linear regression such as the Bernoulli mixture model. In our main model, discussed in following section, we could not implement the package because of the unidentifiability of the model for our data.

1.3 Binomial Mixture Model

Identifiability of a mixture of binomial distributions has been well studied. Follmann and Lambert (1991) studied the mixture of binomial regression models, but their model only allows the intercept to follow a mixture distribution, keeping the coefficients of the covariates constant over components. Sufficient conditions for identifiability for a more general model is given in Grün and Leisch (2008). One of their three sufficient conditions is $N \geq 2K - 1$, where N is the number of trials of the binomial distribution and K is the number of components in the mixture distribution. The other two conditions relate to convergence. In the case of a binary disease outcome, the Bernoulli mixture model does not satisfy the first condition. In this section, we will introduce a binomial mixture model that is identifiable.

1.3.1 Model Definition

In a classical GWAS setting, ancestry covariates are often included in the association model to control for population structure. Here, we propose to use ancestry clusters and define the response variable as the number of cases within an ancestry cluster to solve the identifiability problem. Table 1 shows the data structure after collapsing case-control data for each unique combination of ancestry cluster and genotype. For each ancestry cluster, we have up to three possible genotypes. An instance in the data corresponds to a genotypic value (coded as minor allele count) at an

ancestry cluster. At an instance, let N be the number of observations (Bernoulli trials) and Y be the number of cases (number of successes). As long as all values of N are greater than or equal to 3, we have met the first condition for identifiability of a two-component ($K = 2$) binomial mixture model.

AC	G	N	Y
1	0	5	1
1	1	6	2
1	2	3	2
2	0	10	1
2	1	5	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 1: Data structure exemplar for a mixture of binomial distributions. Column AC and G represent the ancestry cluster and genotype, respectively. Column N gives the total number of observations in a ancestry cluster (AC) with minor allele count (G), which is the number of trials for the binomial distribution, and Y represents the number of cases within ancestry cluster AC with minor allele count G.

Assume m ancestry clusters, $i = 1, 2, \dots, m$, and within cluster i there are m_i genotype values observed with $m_i \in \{1, 2, 3\}$. Let N_{ij} , Y_{ij} , and G_{ij} represent the total number of observations, number of cases, and genotype corresponding to the i th ancestry cluster and the j th genotype value. The two-component mixture of binomial distributions is given by

$$Y_{ij} \sim (1 - \alpha) \text{Binomial}(N_{ij}, \pi_{1ij}) + \alpha \text{Binomial}(N_{ij}, \pi_{2ij}), \quad (4)$$

$$i = 1, \dots, m, \quad j = 1, \dots, m_i,$$

where the probabilities π_{1ij} and π_{2ij} are from logistic regression models,

$$\begin{aligned} \text{logit}(\pi_{1ij}) &= \mu_1, \\ \text{logit}(\pi_{2ij}) &= \mu_2 + \beta G_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, m_i. \end{aligned}$$

The first component, say C_1 , in the mixture model (4) corresponds to the part of the population for which the locus is not associated with disease and the second component, say C_2 , corresponds to the associated group. We use an expectation maximization (EM) algorithm to estimate the model parameters.

1.3.2 Parameter Estimation

The conditional densities of Y_{ij} given components C_1 and C_2 are,

$$f(Y_{ij}|C_k) = \binom{N_{ij}}{Y_{ij}} \pi_{kij}^{Y_{ij}} (1 - \pi_{kij})^{N_{ij} - Y_{ij}}, \quad \text{for } k = 1, 2.$$

The missing data is a two-dimensional latent class vector $\mathbf{z}_{ij} = (z_{1ij}, z_{2ij})$, $i = 1, \dots, m$ and $j = 1, \dots, m_i$, defined as,

$$z_{kij} = \begin{cases} 1 & \text{if } ij\text{th observation comes from the } k\text{th class} \\ 0 & \text{otherwise} \end{cases}$$

for $k = 1, 2$. The component-label vectors $\mathbf{z}_{ij}$ are taken to be realizations of the Bernoulli random vectors $\mathbf{Z}_{ij}$ for $i = 1, \dots, m$, $j = 1, \dots, m_i$ with density function,

$$f(\mathbf{z}_{ij}) = (1 - \alpha)^{z_{1ij}} \alpha^{z_{2ij}}.$$

The set of parameters to be estimated is $\boldsymbol{\theta} = \{\alpha, \beta, \mu_1, \mu_2\}$. The complete data likelihood is given by,

$$L(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}) = \prod_{i=1}^m \prod_{j=1}^{m_i} f(Y_{ij}|C_1)^{z_{1ij}} f(Y_{ij}|C_2)^{z_{2ij}} (1 - \alpha)^{z_{1ij}} \alpha^{z_{2ij}}.$$

where $\mathbf{Y} = (Y_{ij})$ and $\mathbf{Z} = (z_{kij})$, $k = 1, 2$, $i = 1, \dots, m$, and $j = 1, \dots, m_i$.

The log likelihood is

$$\begin{aligned} \ell(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}) &= \sum_{i=1}^m \sum_{j=1}^{m_i} [z_{1ij} \log f(Y_{ij}|C_1) + z_{2ij} \log f(Y_{ij}|C_2)] \\ &\quad + \sum_{i=1}^m \sum_{j=1}^{m_i} [z_{1ij} \log(1 - \alpha) + z_{2ij} \log \alpha]. \end{aligned} \quad (5)$$

In the expectation step of the EM algorithm we find the conditional expectation $E[\mathbf{Z}_{kij}|\mathbf{Y}, \boldsymbol{\theta}]$.

$$E[\mathbf{Z}_{2ij}|\mathbf{Y}, \boldsymbol{\theta}] = \frac{\alpha f(Y_{ij}|C_2)}{(1 - \alpha)f(Y_{ij}|C_1) + \alpha f(Y_{ij}|C_2)}, \quad \text{and } E[\mathbf{Z}_{1ij}|\mathbf{Y}, \boldsymbol{\theta}] = 1 - E[\mathbf{Z}_{2ij}|\mathbf{Y}, \boldsymbol{\theta}].$$

In the maximization step, the log-likelihood (5) is maximized for α at:

$$\alpha = \frac{\sum_{i=1}^m \sum_{j=1}^{m_i} z_{2ij}}{\sum_{i=1}^m m_i}.$$

To maximize (5) for μ_1, μ_2 , and β , note that the second term does not depend on these parameters. We can simplify the first term as follows:

$$\sum_{i=1}^m \sum_{j=1}^{m_i} z_{1ij} \log \binom{N_{ij}}{Y_{ij}} \frac{\exp(\mu_1 Y_{ij})}{(1 + \exp(\mu_1))^{N_{ij}}} + \sum_{i=1}^m \sum_{j=1}^{m_i} z_{2ij} \log \binom{N_{ij}}{Y_{ij}} \frac{\exp[(\mu_2 + \beta G_{ij}) Y_{ij}]}{(1 + \exp(\mu_2 + \beta G_{ij}))^{N_{ij}}}. \quad (6)$$

Maximizing (6) is equivalent to maximizing the two terms separately, which can be achieved by fitting separate logistic regression models with weights of z_{1ij} , and z_{2ij} , respectively.

1.3.3 Hypothesis Test

We are interested in whether 1) there is phenotypic heterogeneity, and 2) under the assumption of phenotypic heterogeneity, a genetic variant is associated with disease status. Consider testing the null hypothesis: $H_0 : \alpha = 1$ and $\beta = 0$. Under this null hypothesis, the mixture model (4) reduces to the logistic regression, $\text{logit}(\pi_{1ij}) = \mu_1$. The corresponding alternative hypothesis is $H_a : \alpha \neq 1$ and/or $\beta \neq 0$. Under this alternative hypothesis we cannot conclude association as there can be a scenario where heterogeneity ($\alpha \neq 1$) exists in terms of a genetic variant, which is not associated

($\beta = 0$), due to a difference in disease prevalence between two groups in the population ($\mu_1 \neq \mu_2$). A more sensible hypothesis can be formed under the assumption that the population prevalence is equal in the two groups, which is a reasonable assumption for a polygenic disease and a genetic variant whose effect size is not extreme. Under the assumption that, $\mu_1 = \mu_2 = \mu$ the binomial mixture model becomes:

$$Y_{ij} \sim (1 - \alpha) \text{Binomial}(N_{ij}, \pi_{1ij}) + \alpha \text{Binomial}(N_{ij}, \pi_{2ij}) \quad (7)$$

where the probabilities π_{1ij} and π_{2ij} are from logistic regression models

$$\begin{aligned} \text{logit}(\pi_{1ij}) &= \mu \\ \text{logit}(\pi_{2ij}) &= \mu + \beta G_{ij}. \end{aligned}$$

Now we can test the null hypothesis, $H_0 : \alpha = 1$ or $\beta = 0$, against the alternative, $H_a : \alpha \neq 1$ and $\beta \neq 0$. Under this null hypothesis, the mixture (7) reduces to a logistic regression: $\text{logit}(\pi_{1ij}) = \mu$. Under the alternative hypothesis, we are now able to conclude association of the genetic variant with the trait.

The likelihood ratio test statistic is given by,

$$\lambda(\mathbf{Y}) = \frac{\sup_{\Theta_0} L(\boldsymbol{\theta}|\mathbf{Y})}{\sup_{\Theta} L(\boldsymbol{\theta}|\mathbf{Y})}$$

where $\Theta = [0, 1] \times \mathbb{R}^3$ is the full parameter space and Θ_0 is a subset corresponding to the null hypothesis. Under some regularity conditions, $-2 \log \lambda$ follows a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the full and null model Casella and Berger (2002). The regularity condition violated in model (7) is that the true parameter value, $\boldsymbol{\theta}_0$, is an interior point in the parameter space. Self and Liang (1987) provides the asymptotic distribution of $-2 \log \lambda$ under such non-standard conditions. According to Self and Liang (1987) the LRTS has an asymptotic $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$ distribution. A Q-Q plot of the LRTS is given in Figure 6.

1.4 Simulation

1.4.1 Simulation Model

Assume that we want to simulate N_1 cases and N_2 controls from a population that is divided into two subgroups S_1 and S_2 of proportions $1 - \alpha$ and α , respectively. The subgroup S_2 represents the group of individuals for which the genetic variant is associated with disease status, and S_1 represents the non-associated group. The subgroups are considered at the population level, so they contain both cases and controls. Let the effect size in the association group be β . Let p_1 , p_2 , K_1 , and K_2 be the minor allele frequencies (maf) and disease prevalences for S_1 and S_2 , respectively. The number of cases N_{11} drawn from S_1 follows a Binomial $(N_1, 1 - \alpha)$ and so we have $N_{12} = N_1 - N_{11}$ cases from S_2 . Similarly, we have $N_{21} \sim \text{Binomial}(N_2, 1 - \alpha)$ and $N_{22} = N_2 - N_{21}$ controls from S_1 and S_2 , respectively.

Observations were randomized to one of the m ancestry clusters. We expect that the allele frequencies will differ among the clusters from the overall allele frequencies p_1 and p_2 . Let F_{ST} be the fixation index. Between-population variation in allele proportions at a biallelic locus is often

modeled by the Beta distribution with parameters λp and $\lambda(1 - p)$, where $\lambda = F_{ST}^{-1} - 1$ (Balding and Nichols, 1995). We will denote the maf at the i th ancestry cluster by $\tilde{p}_i$.

The disease status does not depend on genotype in the non-associated group, so the genotype distribution of cases and controls are obtained assuming Hardy-Weinberg equilibrium (HWE). For the cases and controls from S_2 , the genotype distribution can be obtained using Bayes theorem for $g \in \{0, 1, 2\}$ as

$$P(G = g|Y = 1, S_2) = \frac{P(Y = 1|G = g, S_2)P(G = g|S_2)}{K_2}$$

$$P(G = g|Y = 0, S_2) = \frac{P(Y = 0|G = g, S_2)P(G = g|S_2)}{1 - K_2}$$

where $P(Y = y|G = g, S_2)$ is obtained from a logistic regression model as,

$$P(Y = y|G = g, S_2) = \frac{\exp((\mu_2 + \beta g)y)}{1 + \exp(\mu_2 + \beta g)}.$$

The constant coefficient μ_2 , specific to the i th ancestry cluster, satisfies the following equation:

$$K_2 = \frac{\exp \mu_2}{1 + \exp \mu_{2i}}(1 - \tilde{p}_i)^2 + \frac{\exp(\mu_2 + \beta)}{1 + \exp(\mu_2 + \beta)}2\tilde{p}_i(1 - \tilde{p}_i) + \frac{\exp(\mu_2 + 2\beta)}{1 + \exp(\mu_2 + 2\beta)}\tilde{p}_i^2.$$

1.4.2 Simulation Parameters

Diseases with four different levels of prevalence were simulated: 0.01, 0.02, 0.05, and 0.1. The overall population prevalence is decomposed into prevalences for the two sub-groups, S_1 , and S_2 as,

$$K = (1 - \alpha)K_1 + \alpha K_2 \quad (8)$$

For a complex disease, one single genetic variant is not expected to have a large effect on disease prevalence. If a variant is deleterious, we can expect a higher prevalence in the S_2 group than S_1 . On the other hand, if a genetic variant is beneficial, we can expect a lower prevalence in S_2 compared to S_1 . For simulation, we set K_2 to be 5% higher or lower than K and calculate K_1 accordingly from (8).

We simulate genetic variants with maf from 5% to 50% with an increment of 5%. Since it is highly unlikely that allele frequency at a particular locus is driven by the fact that the locus is associated with a complex trait, we assume that the allele frequency in S_1 and S_2 are equal.

In our simulation, we consider the mixing parameter α to vary from 0.1 to 1 with an increment of 0.1. The effect size β is taken to fall in a range of -1 to 1 in increments of 0.1. Theoretically, the fixation index, F_{ST} varies from 0 to 1, but it is rare to have F_{ST} more than 0.3. In most human populations, such as in Europeans, the fixation index is reported to be about 0.01 (Tian et al., 2009) which is used in our simulation. For each point in the parameter space consisting of K, p, α , and β , we performed 1,000 simulations with 5,000 cases and 20,000 controls.

2 Results

2.1 Use of Binomial Mixture to Meet Identifiability Conditions

We have discussed the identifiability issue of the Bernoulli mixture model (3) that motivated us to develop the binomial mixture model (7). Figure 2 shows the impact of non-identifiability of

the estimates of the mixing proportion based on simulations with 2,000 cases and 8,000 controls. If we use the Bernoulli mixture model in equation (3), almost all observations are assigned to the association group, S_2 , whereas the mean of the mixing parameter estimates are comparatively closer to the truth for the binomial mixture model in equation (7). For example, for $\alpha = 0.6, 0.7, 0.8$, and 0.9 , the mean estimates are $0.67, 0.72, 0.77$, and 0.85 , respectively. The estimates are upward biased and lack precision. For example, the standard deviation of $\hat{\alpha}$ at $\alpha = 0.7$ is 0.29 . As shown in Figure 3, the standard error of the mixing proportion estimate decreases as the magnitude of the effect size increases. For a genetic variant with moderate to high effect ($\beta \geq 0.5$), $SE(\hat{\alpha}) < 0.05$ when mixing proportion $\alpha \geq 0.7$. For a fixed effect size, a higher mixing proportion is estimated with more precision, i.e., smaller $SE(\hat{\alpha})$.

Figure 2: Estimation of the mixing parameter using a Bernoulli mixture model and a binomial mixture model. The x and y axes show the mixing parameter, α , and corresponding estimates, $\hat{\alpha}$, respectively. The estimates correspond to simulations with all possible combinations of maf and β . The * on the boxplots represents the mean of $\hat{\alpha}$.

2.2 OR Estimates are More Precise for Higher Values of OR and Mixing Proportion

Figure 4 shows the standard error of the odds ratio estimates for $\alpha = 0.1, 0.2, \dots, 0.5$ against the effect size, β . The standard error of the odds ratio estimates decreases with an increase in mixing proportion and OR. With a small mixing proportion, the effective number of cases, i.e., the cases in the associated group, is small. For example, with $\alpha = 0.1$ the effective number of cases is only 500. Such a small sample size does not provide precise estimates of effect sizes. In the extreme case of $\alpha = 0.1$ the standard error varies between 0.1 and 0.2. Given this lack of precision, only results for simulations with mixing proportion higher than 0.50 will be provided.

For a protective genetic variant ($OR < 1$), the standard error of the OR estimate decreases with a decrease in OR. For a deleterious genetic variant ($OR > 1$), the standard error of the OR estimate decreases from low to moderate effect sizes and increases slightly for higher effect sizes (Figure 4). For a fixed OR, the standard error of the OR estimate is smaller for a higher level of α .

Figure 3: Standard error of mixing proportion estimates $\hat{\alpha}$. Results are based on simulations with disease prevalence $K = 0.02$ and combined over different levels of allele frequencies. The horizontal axis represents the effect size used in simulations, and the vertical axis represents the standard error of $\hat{\alpha}$ calculated from 1,000 replicates.

2.3 Binomial Mixture Model Yields Smaller Bias Compared to Logistic Regression

For all simulations with mixing proportions from 0.60 to 0.90, for all levels of maf, disease prevalence, and positive β , the odds ratio estimates using the logistic regression and the binomial mixture have downward bias. The bias is upward for negative β . The binomial mixture model reduces the odds ratio bias compared to the logistic regression. Table 2 shows the percentage of bias reduced using the binomial mixture model compared to logistic regression for the simulations with $K = 0.02$. The percentages are calculated as $100 \times (\text{Bias}_{\text{Logistic}} - \text{Bias}_{\text{Binomial Mixture}}) / \text{Bias}_{\text{Logistic}}$. For simulations with $K = 0.02$, the bias reduction by binomial mixture model varies from 1% to 410%. Large bias reduction ($> 100\%$) corresponds to small effects where both models have very small bias (Figure 5). Small bias reduction occurs for a high effect genetic variant where both of the models have high bias. For a complex trait, most of the associated genetic variants have moderate effect sizes ($0.3 \leq \beta \leq 0.7$). For a genetic variant with moderate effect, the binomial mixture reduces

Figure 4: Standard error of odds ratio estimates. The horizontal and vertical axes have the OR and SE of estimated OR, respectively. The left and right panes show results for $\alpha \leq 0.5$ and $\alpha > 0.5$, respectively.

bias in the range of 2.32% to 40.76%. We see similar results for protective loci ($\beta < 0$) and for all levels of disease prevalence considered. A similar tables for $K=0.01, 0.05, 0.10$ are contained in Supplementary Materials. A paired-sample t-test was performed on the bias of the odds ratio based on simulations with $K = 0.02$, mixing proportion 0.60 to 0.90, and all levels of maf considered. In the presence of heterogeneity ($0 < \alpha < 1$), this hypothesis test concludes that the binomial mixture model yields significantly less bias than logistic regression (p-value $< 2.2 \times 10^{-16}$). The performance of the binomial mixture model in terms of bias reduction is graphically shown in Figure 5. The figure shows the results from 1,000 simulation using $K = 0.02$, maf = 0.15, and mixing proportion, $\alpha = 0.70$.

Table 2: Percentage of bias reduced by the binomial mixture model compared to the logistic regression for simulations with $K = 0.02$.¹

Deleterious Loci ($\beta > 0$)										
α	β									
	Small Effect		Moderate Effect						High Effect	
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.60	144.14	44.45	17.26	9.27	5.39	3.54	2.32	1.65	1.15	0.84
0.70	178.91	51.82	19.70	10.35	6.01	3.73	2.44	1.72	1.13	0.80
0.80	241.19	63.19	25.75	12.96	7.48	4.34	2.87	1.78	1.21	0.83
0.90	409.57	104.76	40.76	20.24	11.23	7.10	4.01	2.73	1.79	1.08
Protective Loci ($\beta < 0$)										
α	β									
	Small Effect		Moderate Effect						High Effect	
	-0.1	-0.2	-0.3	-0.4	-0.5	-0.6	-0.7	-0.8	-0.9	-1
0.60	124.99	47.66	21.66	13.63	8.38	6.02	4.64	3.73	3.15	2.63
0.70	167.08	52.40	23.15	13.81	8.75	6.27	4.58	3.48	2.87	2.59
0.80	215.75	65.21	28.56	16.64	10.75	7.15	5.16	4.10	3.10	2.61
0.90	371.06	104.77	47.59	24.54	15.54	9.82	7.20	5.45	3.92	3.34

¹ Results are based on simulations with $K = 0.02$. Percentages are calculated as 100 times the ratio of bias difference to the bias using the logistic regression.

Figure 5: Comparison of the binomial mixture model and the logistic regression in terms of bias in odds ratio estimation. A value closer to the black curve yields smaller bias. The results are based on 1,000 samples with 5,000 cases and 20,000 controls, each considering disease prevalence of 2%, maf of 0.15 and proportion of associated group 70%.

2.4 Establishing the Null Distribution of The Hypothesis Test

There are 1,240,000 simulated sets of data corresponding to the null hypothesis $H_0 : \alpha = 0$ or $\beta = 0$ over different levels of K and maf. The likelihood ratio test statistic (LRTS) did not exceed the critical value of 32.491245 corresponding to a genome-wide significance level (5×10^{-8}) for the distribution of $1/2\chi_1^2 + 1/2\chi_2^2$ for any set of data. The Q-Q and density plots in Figure 6 shows that the distribution of the LRTS has the same shape as the theoretical distribution.

Figure 6: Under the null hypothesis, the likelihood ratio test statistic has an asymptotic $1/2\chi_1^2 + 1/2\chi_2^2$ distribution. The top panel of the figures show density plots of the LRTS and the bottom panel shows Q-Q plots. Three columns (from left to right) represent maf of 0.10, 0.25, and 0.40. For each allele frequency, the results are based on 31,000 simulated samples with $K = 0.02$, and either $\beta = 0$ or $\alpha = 0$.

2.5 Power Comparison of the Binomial Mixture Model with Logistic Regression

While the classical hypothesis test in logistic regression framework pertains to association, the hypothesis test in the binomial mixture model provides information about heterogeneity. To compare the two methods, we compared the compound hypothesis of the proposed method with the simple hypothesis of logistic regression. Figure 7 shows the power curves for disease prevalence, $K = 0.02$. Similar figures for $K = 0.01, 0.05$ and 0.1 are given in the supplementary material. When both the allele frequency and mixing proportion are small, none of the models have high power to detect association because of small effective sample size. Power increases with allele frequency and mixing proportion. When the mixing proportion is close to one, the models have comparable power; otherwise, the logistic regression is slightly more powerful in detecting association.

3 Discussion

The goal of the study was to incorporate heterogeneity into an association model for a case-control analysis. Finite mixture models are commonly used to model heterogeneity. A finite mixture model for a binary outcome is not identifiable, so is not directly applicable to a case-control study. Utilizing ancestry clusters, we produce an identifiable binomial mixture model. We fit a binomial mixture model to a special form of data. An instance in the data corresponds to a genotype in an ancestry cluster. For identifiability each instance has to have at least 3 observations. For real data analysis, the matched control samples can be selected so that this condition is met.

Figure 7: Power curves for hypothesis tests using the proposed mixture model and logistic regression based on samples with disease prevalence, $K = 0.02$. Three rows corresponds to maf of 0.05, 0.25, and 0.50 and three columns corresponds to mixing proportions of 20%, 50% and 90%.

Compared to logistic regression, the binomial mixture model yields lower odds ratio bias, especially for moderate effect sizes. Performance of the proposed model, in terms of bias reduction, was investigated using simulated data for different levels of disease prevalence, allele frequency of the genetic variant, and proportion of the associated group. In addition to bias reduction, the simulation study shows that the binomial mixture model does not suffer much from power loss compared to logistic regression. A likelihood ratio test was performed to assess the hypotheses: $H_0 : \alpha = 0$ or $\beta = 0$ versus $H_a : \alpha \neq 0$ and $\beta \neq 0$. Under the null hypothesis, the LRTS follows a $1/2\chi_1^2 + 1/2\chi_2^2$ distribution. Using simulated data, we validated the null distribution of the LRTS. We compared power of the compound hypothesis test that pertain to association and heterogeneity with a simple hypothesis test in classical logistic regression that tests for only association. Though the binomial mixture model does not gain power in detecting association, it indicates the presence of heterogeneity and reduces bias.

A common genetic variant only explains a small proportion of the total variability for a complex trait. A precise estimate of the associated group is challenging. Though the estimate of mixing proportion is seen to have high standard error when combined over all simulation scenarios, it is more precise when the proportion of associated group is high, and the genetic variant has higher effect on disease outcome. In the case of a small proportion of the associated group ($\alpha < 0.5$), a larger sample should reduce the variances in estimating the mixing parameter, α and effect size, β . A precise estimate of the mixing parameter is not expected since subtypes of a complex disease are generally unattainable based on association of a single genetic variant with the disease. Still, even a noisy estimate and indication of heterogeneity may provide important information about the disease architecture.

A more accurate estimate of odds ratios using the binomial mixture could benefit other analyses that uses GWAS summary level data, such as polygenic risk scores (PRS). We expect that the predictive power of PRS would improve using more unbiased estimates of the effect sizes.

The model can be extended to use detailed phenotypes instead of case-control status, and in that way the model would be able to use the underlying structure of a trait. Another possible extension of the binomial mixture model is a multi-SNP model that uses a set of genetic variants from a specific pathway or a gene. Such extensions could lead us to a more precise estimate of the proportion of the associated group. A precise estimate of the mixing proportion should be influential to the drug discovery process since it indicates the proportion of the target population for a specific drug.

Data Availability Statement: All the codes that were used to simulate the data are available at <https://github.com/spaul-genetics/BMM>. The github repository also includes all the simulation results.

References

- Aitkin, M. and Wilson, G. T. (1980). Mixture models, outliers, and the em algorithm. *Technometrics*, 22(3):325–331.
- Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*, 2(4):e108.

- Balding, D. J. and Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1-2):3–12.
- Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.
- Follmann, D. A. and Lambert, D. (1991). Identifiability of finite mixtures of logistic regression models. *Journal of Statistical Planning and Inference*, 27(3):375–381.
- Gordon, D. and Finch, S. J. (2005). Factors affecting statistical power in the detection of genetic association. *Journal of Clinical Investigation*, 115(6):1408–1418.
- Grados, M. A., Mathews, C. A., for Genetics, T. S. A. I. C., et al. (2008). Latent class analysis of gilles de la tourette syndrome using comorbidities: clinical and genetic implications. *Biological psychiatry*, 64(3):219–225.
- Grün, B. and Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, 51(11):5247–5252.
- Grün, B. and Leisch, F. (2008). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of classification*, 25(2):225–247.
- Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3-4):577–590.
- Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Volla, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299–313.
- Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8):1–18.
- McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, 6:355–378.
- McLachlan, G. J. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.
- Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110.
- Prust, M., Wang, J., Morizono, H., Messing, A., Brenner, M., Gordon, E., Hartka, T., Sokohl, A., Schiffmann, R., Gordish-Dressman, H., et al. (2011). Gfap mutations, age at onset, and clinical subtypes in alexander disease. *Neurology*, 77(13):1287–1294.
- Pyatnitskiy, M., Mazo, I., Shkrob, M., Schwartz, E., and Kotelnikova, E. (2014). Clustering gene expression regulators: new approach to disease subtyping. *PLoS One*, 9(1):e84955.
- Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

- Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., Ladanyi, M., and Sander, C. (2012). Integrative subtype discovery in glioblastoma using icluster. *PloS one*, 7(4):e35236.
- Sluis, S., Verhage, M., Posthuma, D., and Dolan, C. V. (2010). Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. *PloS one*, 5(11):e13929.
- Tian, C., Kosoy, R., Nassir, R., Lee, A., Villoslada, P., Klareskog, L., Hammarström, L., Garchon, H.-J., Pulver, A. E., Ransom, M., et al. (2009). European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse european ethnic groups. *Molecular Medicine*, 15(11-12):371–383.
- Wessman, J., Paunio, T., Tuulio-Henriksson, A., Koivisto, M., Partonen, T., Suvisaari, J., Turunen, J. A., Wedenoja, J., Hennah, W., Pietiläinen, O. P., et al. (2009). Mixture model clustering of phenotype features reveals evidence for association of dtnbp1 to a specific subtype of schizophrenia. *Biological psychiatry*, 66(11):990–996.

A Supplementary Materials

In the supplementary materials we provide additional tables and figures. In the results section we provided a table that describes the percentage of bias reduction using the Binomial Mixture Model based on simulations with disease prevalence, $K = 0.02$. Here we provide similar tables for $K = 0.01, 0.05$ and 0.10 . In the tables, α represents the mixing proportion, i.e., the proportion of the associated group. The parameter β represents the effect size of the genetic variant.

A.1 Supplementary Tables

Table A1: Percentage of bias reduced by the binomial mixture model compared to the logistic regression for simulations with $K = 0.01$. Results are based on simulations with $K = 0.01$. Percentages are calculated as 100 times the ratio of bias difference to the bias using the logistic regression.

Deleterious Loci ($\beta > 0$)										
α	β									
	Small Effect		Moderate Effect					High Effect		
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.60	141.16	44.07	17.38	9.24	5.26	3.37	2.24	1.56	1.18	0.85
0.70	182.23	51.11	20.84	10.22	6.07	3.54	2.39	1.67	1.12	0.81
0.80	246.64	64.83	25.64	12.92	7.40	4.38	2.88	1.87	1.21	0.85
0.90	401.28	105.66	41.08	19.73	11.58	6.67	4.21	2.58	1.73	1.08
Protective Loci ($\beta < 0$)										
α	β									
	Small Effect		Moderate Effect					High Effect		
	-0.1	-0.2	-0.3	-0.4	-0.5	-0.6	-0.7	-0.8	-0.9	-1
0.60	121.19	46.55	21.25	12.73	8.32	6.19	4.73	3.84	3.33	2.70
0.70	153.33	51.98	24.26	13.81	8.97	6.11	4.50	3.63	3.04	2.58
0.80	212.17	64.29	29.45	16.29	10.24	7.27	5.18	3.96	3.14	2.51
0.90	362.88	102.77	46.33	25.09	15.30	10.50	7.13	5.12	4.11	3.14

Table A2: Percentage of bias reduced by the binomial mixture model compared to the logistic regression for simulations with $K = 0.05$. Results are based on simulations with $K = 0.05$. Percentages are calculated as 100 times the ratio of bias difference to the bias using the logistic regression.

Deleterious Loci ($\beta > 0$)										
α	β									
	Small Effect		Moderate Effect					High Effect		
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.60	148.94	46.40	18.39	9.34	5.51	3.54	2.27	1.61	1.11	0.87
0.70	190.76	53.33	19.86	10.45	5.86	3.83	2.50	1.68	1.16	0.84
0.80	253.26	62.67	25.76	13.27	7.41	4.47	2.84	1.91	1.32	0.86
0.90	395.29	105.11	39.60	20.45	11.66	7.05	4.29	2.86	1.77	1.18

Protective Loci ($\beta < 0$)										
α	β									
	Small Effect		Moderate Effect					High Effect		
	-0.1	-0.2	-0.3	-0.4	-0.5	-0.6	-0.7	-0.8	-0.9	-1
0.60	125.40	44.91	21.00	12.11	8.10	5.96	4.54	3.68	2.96	2.57
0.70	154.28	49.70	23.85	13.61	8.89	6.21	4.64	3.50	2.83	2.45
0.80	211.16	65.68	28.67	15.99	9.82	6.92	4.95	3.69	2.88	2.60
0.90	375.45	106.83	46.75	24.52	14.41	9.74	6.96	5.11	3.92	3.03

Table A3: Percentage of bias reduced by the binomial mixture model compared to the logistic regression for simulations with $K = 0.10$. Results are based on simulations with $K = 0.10$. Percentages are calculated as 100 times the ratio of bias difference to the bias using the logistic regression.

Deleterious Loci ($\beta > 0$)										
α	β									
	Small Effect		Moderate Effect					High Effect		
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.60	245552.35	45.71	17.96	8.99	5.66	3.47	2.41	1.61	1.21	0.84
0.70	179.78	51.58	19.66	10.68	6.17	3.63	2.51	1.79	1.23	0.89
0.80	486.94	64.05	24.84	12.93	7.77	4.73	3.11	1.90	1.33	0.89
0.90	522.51	107.44	42.22	19.91	11.86	7.50	4.64	2.76	1.82	1.31

Protective Loci ($\beta < 0$)										
α	β									
	Small Effect		Moderate Effect					High Effect		
	-0.1	-0.2	-0.3	-0.4	-0.5	-0.6	-0.7	-0.8	-0.9	-1
0.60	125.32	45.44	21.84	12.73	8.40	5.89	4.43	3.43	3.03	2.43
0.70	157.69	51.10	24.21	13.79	8.65	5.92	4.47	3.42	2.86	2.37
0.80	216.48	65.05	29.46	16.48	9.63	6.67	4.52	3.67	2.89	2.45
0.90	380.55	103.07	46.88	25.44	15.10	9.89	6.99	5.12	3.59	2.89

Figure A1: Power curves for hypothesis tests using the proposed mixture model and logistic regression based on samples with disease prevalence, $K = 0.01$. Three rows corresponds to maf of 0.05, 0.25, and 0.50 and three columns corresponds to mixing proportions of 20%, 50% and 90%

Figure A2: Power curves for hypothesis tests using the proposed mixture model and logistic regression based on samples with disease prevalence, $K = 0.05$. Three rows corresponds to maf of 0.05, 0.25, and 0.50 and three columns corresponds to mixing proportions of 20%, 50% and 90%

Figure A3: Power curves for hypothesis tests using the proposed mixture model and logistic regression based on samples with disease prevalence, $K = 0.10$. Three rows corresponds to maf of 0.05, 0.25, and 0.50 and three columns corresponds to mixing proportions of 20%, 50% and 90%