Penalized Logistic Regression Analysis for Genetic Association Studies of Binary Phenotypes

Ying Yu[1,*], Siyuan Chen[1], and Brad McNeney[1]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada

[*]Corresponding Author: ying_yu_5@sfu.ca

Abstract

Increasingly, logistic regression methods for genetic association studies of binary phenotypes must be able to accommodate rare genetic variants. Rare variants represent sparse covariates that are predominately zeros, and this sparseness leads to maximum likelihood estimators (MLEs) of log-OR parameters that are biased away from their null value of zero. Different penalized-likelihood methods have been developed to mitigate sparse-data bias. We study penalized logistic regression using a class of log-F priors indexed by a shrinkage parameter m to shrink the biased MLE towards zero. For a given m, log-F-penalized logistic regression may be easily implemented using data augmentation and standard software. We propose a two-step approach to the analysis of a genetic association study: first, a set of approximately independent variants is used to estimate m; and second, the estimated m is used for log-F-penalized logistic regression analyses of all variants. Our estimate of m is the maximizer of a marginal likelihood obtained by integrating the latent log-ORs out of the joint distribution of the parameters and observed data. We consider two approximate approaches to maximizing the marginal likelihood: (i) a Monte Carlo EM algorithm and (ii) a Laplace approximation to each integral, followed by derivative-free optimization of the approximation. We evaluate the statistical properties of our proposed two-step method and compared its performance to other shrinkage methods by a simulation study.

Keywords: Rare genetic variants, Penalized logistic regression, log-F priors, Monte Carlo EM, Laplace approximation, Data augmentation

1 Introduction

Genome-wide association studies (GWAS) have successfully identified thousands of genetic variants associated with human traits and diseases [12]. Traditional GWAS focused on common variation in the genome with minor allele frequency (MAF) $\geq 5\%$, but such variants were found to explain only a small fraction of trait heritability [20], leading researchers to consider rare variants. Rare genetic variants, which occur at a low frequency in a population with MAF $< 1\%$, are thought to have larger effect sizes [24], and may explain some of the missing heritability of complex human diseases [19].

There has been recent interest in methods for estimation and testing of single rare variant effects on binary phenotypes [21, 29]. However, logistic regression analysis of rare variants involves sparse covariates that are predominately zeros, and this sparseness leads to finite-sample, or sparse-data bias [8] of the maximum likelihood estimator (MLE) of log-OR parameters. In particular, the log-OR estimator is biased away from the null value of zero. As far as we are aware, Firth logistic regression [5] is the most widely used method for mitigating sparse-data bias in genetic association studies. Firth logistic regression can be viewed as a penalized MLE for a likelihood penalized by the Jeffreys invariant prior [5, 11, 13].

A variety of alternative penalties have been proposed for logistic regression, offering more or less shrinkage than the Jeffreys prior [7]. Greenland and Mansournia develop penalized logistic regression based on a class

of log-F priors indexed by a shrinkage parameter m [9]. In this method, the log-OR parameters, β, are assumed to be independent and to have log-$F(m, m)$ distributions with density

$$f(\beta|m) = \frac{1}{B(\frac{m}{2}, \frac{m}{2})} \frac{\exp(\frac{m}{2}\beta)}{(1 + \exp(\beta))^m}, \tag{1}$$

where $B(\cdot, \cdot)$ is the beta function (see Figure 1 for plots of log-$F(1, 1)$ and log-$F(10, 10)$ density curves). Maximizing the posterior density is equivalent to maximizing a penalized likelihood obtained by multiplying the logistic regression likelihood by as many independent log-$F(m, m)$ distributions as there are explanatory variables. Note that the intercept term is not penalized.

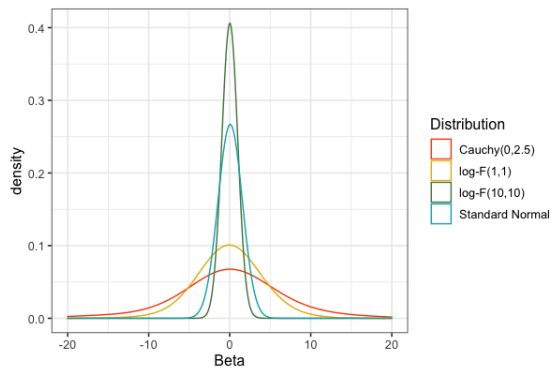

Figure 1: Comparison of log-F, standard normal and Cauchy distributions. The log-$F(m, m)$ density is symmetrically bell-shaped with a single peak at zero, and its variance decreases as increasing m. As $m \to \infty$, the distribution tends toward a point mass at zero.

Limited simulation studies have shown that, for fixed m, log-$F(m, m)$ penalized methods outperform other approaches for case-control data [7]. Compared to Firth's method, the log-F approach is more flexible, since we can change the amount of shrinkage by changing the value of m. Furthermore, greater shrinkage may reduce MSE [9]. However, there is little guidance on how best to select the value of m for a particular phenotype. As a shrinkage parameter, m controls the bias-variance trade-off, with the variance of the log-OR estimator decreasing and the bias increasing as m increases [9]. We follow the suggestion by Greenland and Mansourinia of using an empirical Bayes method to estimate m [9]. The context for our approach is that we are interested in fitting single-SNV logistic regressions over a genomic region, or over the entire genome. From this region we select K approximately independent SNVs and use these to estimate m. Our approach is based on a marginal likelihood for m obtained by integrating the latent log-OR parameters out of the joint distribution of the parameters and observed data. The integrals are challenging to evaluate analytically and so we consider two approximate approaches to maximizing the marginal likelihood. The first is a Monte Carlo EM algorithm that treats the latent log-OR parameters as missing data. The second approach is a combination of a Laplace approximation to the integral and derivative-free optimization of the resulting approximate marginal likelihood.

The main goal of this paper is to employ the log-F penalized logistic regression for analyzing rare variant associations in a two-step approach. First, we estimate the shrinkage parameter m based on a set of approximately independent variants selected from those available. Second, we perform penalized logistic regression for each variant in the study using log-$F(m, m)$ penalization with m estimated from step one. For a given m, the log-F penalized likelihood method can be conveniently implemented by fitting a standard logistic regression to an augmented dataset [9].

The rest of the paper is organized as follows. We introduce our methods in Section 2 and conduct simulation studies to evaluate the performance of our methods under a variety of data-generating scenarios in Section 3. We end with concluding remarks and a discussion in Section 4.

2 Models and Methods

Though our methods apply for any binary phenotype, special attention is given to case-control data. We start by reviewing the likelihood from case-control data. We then derive a marginal likelihood for the shrinkage parameter m based on data from K approximately independent SNVs and discuss maximization of the marginal likelihood. We conclude by reviewing how log-F-penalized logistic regression for the second-stage of the analysis can be implemented by data augmentation.

2.1 Case-control Likelihood

The association between a single-nucleotide variant (SNV) and disease status is often estimated from case-control (i.e. retrospective) data, where covariates X_i are sampled conditional on disease status Y_i for each individual i. Let β denote the log-OR parameter, Qin and Zhang [23] expressed the case-control likelihood in terms of a two-sample semi-parametric model as follows

$$\begin{aligned} L(\beta, g) &= \prod_{i=1}^{n_0} P(X_i|Y_i = 0) \prod_{i=n_0+1}^{n_0+n_1} P(X_i|Y_i = 1) \\ &= \prod_{i=1}^{n_0} g(X_i) \prod_{i=n_0+1}^{n_0+n_1} c(\beta, g)\exp(X_i\beta)g(X_i), \end{aligned} \tag{2}$$

where $c(\beta, g)$ is a normalizing constant. The infinite-dimensional nuisance parameter g makes the case-control likelihood $L(\beta, g)$ difficult to derive and maximize to find the MLE of β. Therefore, we rewrite the case-control likelihood as a profile likelihood [22]:

$$\begin{aligned} L(\alpha^*, \beta) &= f(\boldsymbol{X}|\alpha^*, \beta) \\ &= \prod_{i=1}^{n_0} \frac{1}{1 + \exp(\alpha^* + X_i\beta)} \prod_{i=n_0+1}^{n_0+n_1} \frac{\exp(\alpha^* + X_i\beta)}{1 + \exp(\alpha^* + X_i\beta)} \\ &= \prod_{i=1}^{n} \frac{\exp(Y_i(\alpha^* + X_i\beta))}{1 + \exp(\alpha^* + X_i\beta)}, \end{aligned} \tag{3}$$

where $\alpha^* = \alpha + \log\left(\frac{n_1}{n_0}\right) - \log\left(\frac{P(D=1)}{P(D=0)}\right)$, α is the intercept term in the logistic regression model for $P(Y = 1|X)$, and $P(D = 1)$ and $P(D = 0)$ are the population probabilities of having and not having the disease, respectively [25]. The profile likelihood $L(\alpha^*, \beta)$ for case-control data is of the same form as a prospective likelihood. The MLE of β under the case-control sampling design can be obtained by maximizing $L(\alpha^*, \beta)$ as if the data were collected in a prospective study [22, 23].

2.2 Marginal likelihood for estimating m

Suppose we have K approximately independent SNVs, either from a particular genomic region or from across the genome. For each we specify a one-covariate logistic regression model. Let $\boldsymbol{X}$ denote a design matrix containing all K SNVs, and $\boldsymbol{X}_{.k}$, $k = 1, \ldots, K$, denote the genotype data on the kth SNV. Let $L(\alpha_k^*, \beta_k)$ denote the profile likelihood (3) for the kth log-OR parameter β_k. Here α_k^* is the intercept term from the kth profile likelihood, considered to be a nuisance parameter. Using the assumed independence of the SNVs and log-ORs, the joint penalized profile likelihood is the product:

$$L(\boldsymbol{\alpha^*}, \boldsymbol{\beta})f(\boldsymbol{\beta}|m) = \prod_{k=1}^{K} L(\alpha_k^*, \beta_k)f(\beta_k|m), \tag{4}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)^T$ and $\boldsymbol{\alpha^*} = (\alpha_1^*, ..., \alpha_K^*)$. The marginal likelihood for $(\boldsymbol{\alpha^*}, m)$ is obtained by integrating $\boldsymbol{\beta}$ out of the profile complete-data likelihood:

$$\begin{aligned} L(\boldsymbol{\alpha^*}, m) &= \int L(\boldsymbol{\alpha^*}, \boldsymbol{\beta}) f(\boldsymbol{\beta}|m) d\boldsymbol{\beta} \\ &= \int \prod_{k=1}^{K} L(\alpha_k^*, \beta_k) f(\beta_k|m) d\boldsymbol{\beta} \\ &= \prod_{k=1}^{K} \int L(\alpha_k^*, \beta_k) f(\beta_k|m) d\beta_k \\ &= \prod_{k=1}^{K} L(\alpha_k^*, m), \end{aligned} \tag{5}$$

where $L(\alpha_k^*, m)$ is the marginal likelihood contribution from the kth variant. We select the value of m that maximizes the marginal log-likelihood

$$l(\boldsymbol{\alpha^*}, m) = \sum_{k=1}^{K} \log[L(\alpha_k^*, m)] = \sum_{k=1}^{K} l(\alpha_k^*, m). \tag{6}$$

Preliminary analyses suggest that basing the marginal likelihood on weakly-associated SNVs leads to over-estimation of m. We therefore weight the contribution of each SNV to the marginal log-likelihood by a weight w_k intended to down-weight SNVs of small effect. Following the notion that common variants should tend to have weaker effects and rare variants stronger effects, we propose to set $\sqrt{w_k} = 1/\sqrt{MAF_k(1 - MAF_k)}$ so that w_k is inversely proportional to the MAF of kth SNV. With weighting, equation (6) becomes

$$l(\boldsymbol{\alpha^*}, m) = \sum_{k=1}^{K} w_k l(\alpha_k^*, m) \tag{7}$$

Maximization is done in two stages:

1. For fixed m, we maximize $l(\boldsymbol{\alpha^*}, m)$. The factorization of the likelihood when m is fixed implies that to maximize $l(\boldsymbol{\alpha^*}, m)$ we maximize each $l(\alpha_k^*, m)$ over α_k^*. Let $\alpha_k^*(m)$ be the value of α_k^* that maximizes $l(\alpha_k^*, m)$, $\hat{\boldsymbol{\alpha}}^*(m) = (\alpha_1^*(m), \ldots, \alpha_K^*(m))$, and $l(\hat{\boldsymbol{\alpha}}^*(m), m) = \sum_{k=1}^{K} l(\hat{\alpha_k^*}(m), m)$.

2. Maximize $l(\hat{\boldsymbol{\alpha}}^*(m), m)$ over m. To keep computations manageable, we restrict m to a grid of values, $m = 1, 2, ..., M$. One may optionally smooth the resulting $(m, l(\hat{\boldsymbol{\alpha}}^*(m), m))$ pairs and maximize this smoothed curve to obtain the estimate $\hat{m}$.

For a fixed value of m and k, the estimate $\hat{\alpha_k^*}(m)$ can be obtained by maximizing $l(\alpha_k^*, m)$ with respect to α_k^*. However, it is difficult to evaluate the integral $\int L(\alpha_k^*, \beta_k) f(\beta_k|m) d\beta_k$ in (5). We discuss two approximate approaches. The first (Section 2.3.1) is a Monte Carlo EM algorithm [4], and the second (Section 2.3.2) is a Laplace approximation to $L(\alpha_k^*, m)$ followed by derivative-free optimization of the approximation.
We conclude this sub-section by noting that it is possible to generalize the marginal likelihood approach for estimating m to incorporate non-genetic confounding variables, denoted Z. As confounders, Z will be correlated with the SNV covariates X_k, and such correlation may differ across SNVs. We therefore introduce coefficients γ_k for the confounding variables in the logistic regression on the kth SNV. Expanding the α_k^* component of the logistic model to $\alpha_k^* + Z\gamma_k$, the kth case-control profile likelihood is now

$$L(\alpha_k^*, \gamma_k, \beta_k) = \prod_{i=1}^{n} \frac{\exp(Y_i(\alpha_k^* + Z_i\gamma_k + X_{ik}\beta_k))}{1 + \exp(\alpha_k^* + Z_i\gamma_k + X_{ik}\beta_k)} \tag{8}$$

and the weighted marginal log-likelihood for estimating m is

$$\begin{aligned} l(\boldsymbol{\alpha}^*, \boldsymbol{\gamma}, m) &= \sum_{k=1}^{K} w_k l(\alpha_k^*, \gamma_k, m) \\ &= \sum_{k=1}^{K} w_k \log \int L(\alpha_k^*, \gamma_k, \beta_k) f(\beta_k|m) d\beta_k. \end{aligned} \tag{9}$$

For fixed m we maximize $l(\boldsymbol{\alpha}^*, \boldsymbol{\gamma}, m)$ by maximizing the component marginal likelihoods $l(\alpha_k^*, \gamma_k, m)$ over the nuisance parameters (α_k^*, γ_k). We then maximize the resulting expression over m to obtain $\hat{m}$. Though this generalization to include confounding variables is conceptually straightforward, we omit the confounders in what follows to keep the notation as simple as possible.

2.3 Maximization Approaches

2.3.1 Monte Carlo EM Algorithm

To maximize $l(\alpha_k^*, m)$, we first consider an EM algorithm, which treats β_k as the unobserved latent variable or missing data. For a fixed value of m and k, the EM algorithm is based on the profile complete-data log-likelihood $\log\left[f(\boldsymbol{X}_{.k}, \beta_k|\alpha_k^*, m)\right]$. At the $(p+1)^{th}$ iteration, the E-step is to determine

$$\begin{aligned} Q\left(\alpha_k^*|\alpha_k^{*(p)}, m\right) &= \mathrm{E}_{\beta_k|\boldsymbol{X}_{.k}, \alpha_k^{*(p)}, m}(\log[f(\boldsymbol{X}_{.k}, \beta_k|\alpha_k^*, m)]) \\ &= \int \log\left[f(\boldsymbol{X}_{.k}, \beta_k|\alpha_k^*, m)\right] f(\beta_k|\boldsymbol{X}_{.k}, \alpha_k^{*(p)}, m) d\beta_k \\ &\propto \int \log[f(\boldsymbol{X}_{.k}|\alpha_k^*, \beta_k) f(\beta_k|m)] f(\boldsymbol{X}_{.k}|\alpha^{*(p)}_k, \beta_k) f(\beta_k|m) d\beta_k \end{aligned} \tag{10}$$

and the M-step is to set

$$\alpha_k^{*(p+1)} = \underset{\alpha_k^*}{\operatorname{argmax}}\, Q\left(\alpha_k^*|\alpha_k^{*(p)}, m\right). \tag{11}$$

The E-step (10) is complicated by the fact that the integral cannot be solved analytically. We therefore approximate the integral numerically by Monte Carlo (MC); that is, we use a Monte Carlo EM (MCEM) algorithm [27]. The MC integration in the E-step is obtained by sampling from the prior distribution $f(\beta_k|m)$ [15, 27]. Based on a sample $\beta_{k1}, ..., \beta_{kN}$ from the distribution $f(\beta_k|m)$, the MC approximation to the integral is

$$\begin{aligned} Q\left(\alpha_k^*|\alpha_k^{*(p)}, m\right) &\approx Q_{MC}\left(\alpha_k^*|\alpha_k^{*(p)}, m\right) \\ &= \frac{1}{N}\sum_{j=1}^{N} \log[f(\boldsymbol{X}_{.k}|\alpha_k^*, \beta_{kj}) f(\beta_{kj}|m)] f(\boldsymbol{X}_{.k}|\alpha_k^{*(p)}, \beta_{kj}) \\ &= \frac{1}{N}\sum_{j=1}^{N} (\log[f(\boldsymbol{X}_{.k}|\alpha_k^*, \beta_{kj})] + \log[f(\beta_{kj}|m)]) f(\boldsymbol{X}_{.k}|\alpha_k^{*(p)}, \beta_{kj}). \end{aligned} \tag{12}$$

Note that $\log[f(\beta_{kj}|m)]$ is independent of the parameter α_k^*, so maximizing (12) in the M-step is equivalent to maximizing

$$\frac{1}{N}\sum_{j=1}^{N} \log[f(\boldsymbol{X}_{.k}|\alpha_k^*, \beta_{kj})] f(\boldsymbol{X}_{.k}|\alpha_k^{*(p)}, \beta_{kj}). \tag{13}$$

For a discussion of computational approaches to the M-step see Appendix A.

2.3.2 Maximization of a Laplace Approximation

An alternative to the EM algorithm is to make an analytic approximation, $\tilde{L}(\alpha^*, m)$, to $L(\alpha^*, m) = \int L(\alpha_k^*, \beta_k) f(\beta_k|m) d\beta_k$ and maximize this approximation. We considered Laplace approximation because it

is widely used for approximating marginal likelihoods [26]. The Laplace approximation of an integral is the integral of an unnormalized Gaussian density matched to the integrand on its mode and curvature at the mode. Letting $\hat{\beta}_k$ denote the mode of $L(\alpha_k^*, \beta_k) f(\beta_k|m)$ and $c_p(\alpha_k^*)$ minus its second derivative at $\hat{\beta}_k$, the Laplace approximation to $L(\alpha_k^*, m)$ is

$$\tilde{L}(\alpha_k^*, m) = L(\alpha_k^*, \hat{\beta}_k) f(\hat{\beta}_k|m) \sqrt{\frac{2\pi}{c_p(\alpha_k^*)}}. \tag{14}$$

Each $\hat{\beta}_k$ is the root of the derivative equation $\partial log(L(\alpha_k^*, \beta_k) f(\beta_k|m))/\partial \beta_k = 0$; this can be shown to be a global maximum of $L(\alpha_k^*, \beta_k) f(\beta_k|m)$. An expression for $c_p(\alpha_k^*)$ is given in Appendix A of [2]. Figure 2 shows the quality of the LA for one simulated dataset generated under $m = 4$. The approximate marginal likelihood $\tilde{L}(\alpha_k^*, m)$ may be maximized over α^* using standard derivative-free optimization methods, such as a golden section search or the Nelder-Mead algorithm.

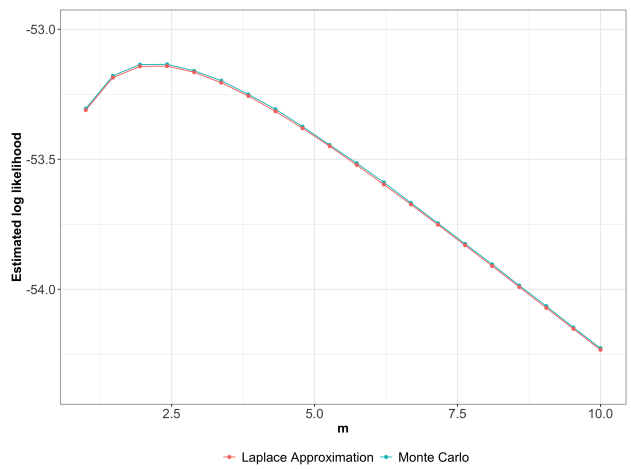

Figure 2: Natural logarithms of estimates of the marginal likelihood $L(\alpha_k^*, m)$ for one simulated dataset generated under $m = 4$. Estimates are obtained by LA and Monte Carlo. Log-likelihood estimates are plotted over the grid $m = (1, 1.5, ..., 10)$ with $\alpha_k^* = -3$.

2.4 Implementing log-F Penalization by Data Augmentation

Penalization by a log-$F(m, m)$ prior can be achieved by standard GLM through data augmentation suggested by Greenland and Mansournia [9]. Here, we provide some details. For simplicity, we start with a case-control study of the relation of a single covariate to the response. The penalized logistic likelihood by a log-$F(m, m)$ prior would be:

$$\begin{aligned} l^*(\beta) &= \prod_{i=1}^{n} \frac{\exp(Y_i X_i \beta)}{1 + \exp(X_i \beta)} \times \frac{\exp(\frac{m}{2}\beta)}{(1 + \exp(\beta))^m} \\ &= \prod_{i=1}^{n} \frac{\exp(Y_i X_i \beta)}{1 + \exp(X_i \beta)} \times \left[\frac{\exp(X_i \beta)}{1 + \exp(X_i \beta)}\right]^{\frac{m}{2}} \left[\frac{1}{1 + \exp(X_i \beta)}\right]^{\frac{m}{2}}, \end{aligned} \tag{15}$$

where $X_i = 1$. Thus, the penalized likelihood $l^*(\beta)$ is equivalent to the unpenalized likelihood obtained by adding m pseudo-observations to the response with covariate of one, in which $m/2$ are successes and $m/2$ are failures (even if m is an odd number). In the multivariate case, the penalty term is $\prod_k \frac{\exp(\frac{m}{2}\beta_k)}{(1+\exp(\beta_k))^m}$, where product ranges over multiple covariates. This implies, for each covariate, we add m pseudo-observations with $m/2$ successes and $m/2$ failures to the response, and a single row to the design matrix consisting all zeros except for a one indicating the index of the covariate of interest [9].

In our analyses (see Section 3), we analyze one SNV at a time using the log-F penalized logistic regression, adjusting for other confounding variables. The data augmentation approach is illustrated in Figure 3. Let

X denote the allele count of a SNV and Z_j, $j = 1, ..., p$, denote other confounding variables for adjustment. In the augmented dataset, the response is a two-column matrix with the number of successes and failures as the two columns. The m pseudo-observations are split into $m/2$ successes and $m/2$ failures. We only penalize the coefficient associated with the SNV, so we add a single row to the design matrix consisting all zeros except for a one indicating the SNV covariate. Analyzing the augmented dataset with standard logistic regression yields the penalized MLE and the corresponding standard errors.

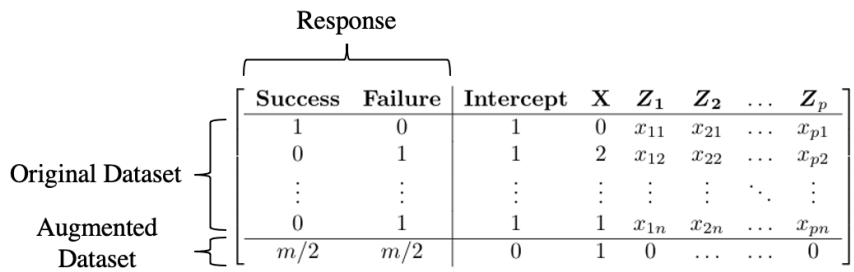

Figure 3: Illustration of data augmentation in the implementation of log-$F(m, m)$ penalization.

3 A Simulation Study

The empirical performance of the methods introduced in Section 2 was evaluated in a comprehensive simulation study. Besides the standard MLE, we compared the performance of our proposed two-step log-F-penalized method (LogF) to the performance of the following methods:

- Firth logistic regression (FLR) was first proposed by Firth [5], where the logistic likelihood is penalized by $|I(\beta)|^{1/2}$ with $I(\beta) = -E\left[\frac{\partial^2}{\partial\beta^2} l(\beta)\right]$ defined as the Fisher information. FLR is implemented in the `R` function `logistf` of the package `logistf` [10].
- Penalization by Cauchy priors (CP) was proposed by [6]. The input predictors are rescaled to have a mean of 0 and a standard deviation of 0.5. All predictors are penalized by a Cauchy prior with center 0 and scale 2.5, whereas the intercept is penalized by a weaker Cauchy prior with center 0 and scale 10. CP is implemented in the the `R` function `bayesglm` of the package `arm` [6].

In each simulation scenario, we generated 1000 data sets. For each dataset we estimated the shrinkage parameter m and then applied the estimated value of m to estimate the log-ORs. Therefore, the log-OR parameter of each variant will be estimated 1000 times. We examined the properties of all methods based on simulation results averaged across all variants. Confidence intervals were evaluated with regard to coverage and power. For logistic regression, profile penalized likelihood (PPL) confidence intervals estimated by solving $2[l^*(\hat{\beta}) - l^*(\beta)] = \chi^2_{1-\alpha}$ for β have been shown to have good empirical properties [11]. Thus, PPL confidence intervals were used for LF and FL, whereas Wald-type confidence intervals were used for MLE and CP.

3.1 Data Generation

We set the magnitude of each $|\beta_k| = \frac{\log 5}{2}|\log_{10} \text{MAF}_k|$ [28] for $k = 1, ..., K$, such that rare SNVs have larger effect sizes and common SNVs have smaller effect sizes. Here we use $K = 50$ in all scenarios if not stated otherwise. We took into account the effects of mixed signs, multiplying β_k by -1 for some k, in which are 50% positive and 50% negative. To mimic real situations we sampled MAFs from a site frequency spectrum of a real sample of SNVs. We used data from 1000 genomes project [3], restricting to European subjects. Data from the 1000 genomes project was downloaded using the Data Slicer (https://www.internationalgenome.org/data-slicer/). We selected a 1 million base pair region on chromosome 6. This process lead to genotypes on 8725 SNVs of 503 individuals (SNVs with MAF = 0 have been removed). MAFs were sampled from the empirical site frequency spectrum of this genomic region.

In the simulation studies, we set sample size n to 500 and 1000. For $n = 500$, the maximum OR = 5 ($|\beta_k| = 1.61$) for SNV with MAF = 0.01 and minimum OR = 1.31 ($|\beta_k| = 0.27$) for SNV with MAF = 0.47; for $n = 1000$, the maximum OR = 6.37 ($|\beta_k| = 1.85$) for SNV with MAF = 0.005 and minimum OR = 1.31 ($|\beta_k| = 0.27$) for SNV with MAF = 0.47. For each scenario, 1000 balanced case-control datasets were simulated and analyzed by the methods described in Section 2 and the competing methods mentioned above.

The case-control data was simulated according to a retrospective sampling design, where covariates are simulated based on disease status. For a given SNV, let X denote the allele count (i.e. 0, 1 or 2) of the SNV and β be the corresponding log-OR parameter. Following [23], the conditional density function for X in the controls and cases are

$$\begin{aligned} &P(X = x|Y = 0) = g(x) \text{ and} \\ &P(X = x|Y = 1) = h(x) = c(\beta, g)\exp(x\beta)g(x). \end{aligned} \tag{16}$$

We assume that the distribution of X in controls, $g(x)$, is Binomial$(2, p)$, where p is the MAF of the SNV. Then the distribution of X in cases, $h(x)$, is proportional to

$$g_z(x)\exp(x\beta) = \begin{cases} (1-p)^2 & x = 0 \\ 2p(1-p)\exp(\beta) & x = 1 \\ p^2\exp(2\beta) & x = 2 \end{cases}, \tag{17}$$

which has normalizing constant $(1-p)^2 + 2p(1-p)\exp(\beta) + p^2\exp(2\beta)$. Then X in cases can be obtained by sampling from (17). We simulate the genotype data of each SNV independently using the above approach.

We also simulated data in the presence of population stratification. We create population-disease and population-SNV associations as follows. To create population-disease association we introduced a population main effect on disease risk by taking population-stratum log-OR, γ, to be 1. To create population-SNV association we selected different SNV MAFs in different populations. Let Z denote a binary indicator of one of the two population strata. The respective frequencies in controls of the two populations are f_0 and f_1, respectively. Then the distribution of Z in controls is $P(Z = z|Y = 0) = f_z$, and the distribution of Z in cases is $P(Z = z|Y = 1) \propto f_z\exp(z\gamma)$ [30]. In our studies, we set $f_0 = f_1 = 0.5$. Now suppose that the MAF for a given SNV differs by sub-population, with p_z denoting the MAF in population z. Let $g_z(x)$ denote the distribution of X in controls of population z, i.e., $P(X = x|Z = z, Y = 0) = g_z(x) \sim$ Binomial$(2, p_z)$. The joint distribution of X and Z in controls is then $P(X = x, Z = z|Y = 0) = f_z g_z(x)$. If logit$[P(Y = 1|Z = z, X = x)] = \alpha + z\gamma + x\beta$, the joint distribution of X and Z in cases is $P(X = x, Z = z|Y = 1) \propto f_z g_z(x)\exp(z\gamma + x\beta)$ [23]. We then have

$$P(X = x|Z = z, Y = 1) = \frac{P(X = x, Z = z|Y = 1)}{P(Z = z|Y = 1)} \propto \frac{f_z g_z(x)\exp(z\gamma + x\beta)}{f_z\exp(z\gamma)} = g_z(x)\exp(x\beta). \tag{18}$$

To summarize, we first assign population status for each subject using

$$\begin{aligned} &P(Z = z|Y = 0) = f_z, \\ &P(Z = z|Y = 1) \propto f_z\exp(z\gamma) = \begin{cases} f_0 & z = 0 \\ f_1\exp(\gamma) & z = 1 \end{cases} \end{aligned} \tag{19}$$

Then we simulate each SNV conditional on population status by sampling from

$$\begin{aligned} &P(X = x|Z = z, Y = 0) = g_z(x) \sim \text{Binomial}(2, p_z), \\ &P(X = x|Z = z, Y = 1) \propto g_z(x)\exp(x\beta) = \begin{cases} (1-p_z)^2 & x = 0 \\ 2p_z(1-p_z)\exp(\beta) & x = 1 \\ p_z^2\exp(2\beta) & x = 2 \end{cases} \end{aligned} \tag{20}$$

MAFs, p_z, for different populations were obtained from 1000 genomes project [3]. Here we consider two populations: Caucasian (CEU) and Yoruba (YRI) subjects. Pooled-MAFs, $(p_0+p_1)/2$, were used to generate β_k's. The maximum OR = 6.44 ($|\beta_k| = 1.86$) for SNV with pooled-MAF = 0.0048 and minimum OR = 1.36 ($|\beta_k| = 0.31$) for SNV with pooled-MAF = 0.41.

3.2 Results

In our setup, the variants can be either deleterious or protective (i.e. β_k is either positive or negative), so we define the bias of estimates as the signed bias, $E[\text{sign}(\beta_k)(\hat{\beta}_k - \beta_k)]$; positive values indicate bias away from zero, while negative bias indicates bias towards zero. Table 1 shows the estimated bias, SD and MSE of log-OR estimators averaged across all 50 SNVs when $n = 500$ or 1000. The MLE has considerably larger bias, SD and MSE than the other methods in both scenarios because of the sparse-data bias. FLR and CP essentially eliminate bias, while the LogF approaches over-correct and have negative bias. However, the LogF approaches outperform the other methods with respect to standard errors and MSE, which is consistent with results of [9]. The LogF-MCEM estimator and LogF-LA estimator perform similarly, being associated with the smallest and second smallest MSE throughout both scenarios. Over 1000 simulations, the average of estimated m obtained by MCEM and LA are 2.93 (SD = 0.77) and 5.38 (SD = 1.42) respectively for $n = 500$, and 2.79 (SD = 0.47) and 3.36 (SD = 0.44) respectively for $n = 1000$. LA tends to return a larger value of m than MCEM does, leading to larger shrinkage and bias but smaller variance. Simulation results under the presence of population stratification are shown in Table 3. Introducing the confounding variable leads to a slight decrease in m. Over 1000 simulations, the mean estimated m obtained by MCEM and LA are 2.40 (SD = 0.68) and 4.24 (SD = 1.55) respectively for $n = 500$, and 2.38 (SD = 0.52) and 2.91 (SD = 0.59) respectively for $n = 1000$. In these simulations, the LogF estimators achieve the smallest MSE with biases that are comparable to the biases of FLR and CP estimates.

Results on the coverage and power of two-sided nominal 95% CIs averaged over all coefficients can be found in Tables 2 and 4. With the exception of the LogF-LA when $n = 500$, coverage is within simulation error of the nominal 95%. The LogF-LA CIs are slightly anti-conservative because of the downward bias and small SD of the estiamtor. The Wald-type CIs have slightly higher coverage probability than CIs based on PPL. In comparison, the power of LogF-MCEM, LogF-LA and FLR methods are generally higher than the power of MLE and CP. In particular, FLR have the highest power to exclude zero in all scenarios followed by LogF-MCEM and logF-LA.

Table 1: Simulation results for bias, SD and MSE of log-OR coefficients, averaged over all 50 SNVs.

n	Method	Bias($\times$100)	SD($\times$100)	MSE ($\times$100)
500	**MLE**	40	102	588
	CP	0	31	17
	FLR	-1	29	14
	LogF-MCEM	-8	25	12
	LogF-LA	-13	23	13
1000	**MLE**	7	40	89
	CP	0	21	8
	FLR	-0	21	7
	LogF-MCEM	-3	19	6
	LogF-LA	-4	18	6

Table 2: Simulation results for coverage and power of CIs of log-OR coefficients with population stratification ($\gamma = 1$), averaged over all 50 SNVs. * Coverage probability of two-sided nominal 95% confidence intervals for log-OR coefficient. † Relative frequency of confidence intervals for log-OR coefficients excluding 0. § PPL, profile penalized likelihood.

n	Method	Method for CI	Coverage*($\times$1000)	Power†($\times$1000)
500	**MLE**	Wald	954	739
	CP		952	750
	FLR	PPL§	950	772
	LogF-MCEM		947	758
	LogF-LA		934	743
1000	**MLE**	Wald	953	942
	CP		952	944
	FLR	PPL§	950	950
	LogF-MCEM		950	946
	LogF-LA		949	944

Table 3: Simulation results for bias, SD and MSE of log-OR coefficients with adjustment for population stratification ($\gamma = 1$), averaged over all 50 SNVs.

n	Method	Bias($\times$100)	SD($\times$100)	MSE($\times$100)
500	**MLE**	39	106	571
	CP	0	34	19
	FLR	-1	32	15
	LogF-MCEM	-6	29	13
	LogF-LA	-12	26	14
1000	**MLE**	12	51	155
	CP	0	24	10
	FLR	-0	23	9
	LogF-MCEM	-3	22	8
	LogF-LA	-5	21	8

Table 4: Simulation results for coverage and power of CIs of log-OR coefficients with population stratification ($\gamma = 1$), averaged over all 50 SNVs. * Coverage probability of two-sided nominal 95% confidence intervals for log-OR coefficient. † Relative frequency of confidence intervals for log-OR coefficients excluding 0. § PPL, profile penalized likelihood confidence interval.

n	Method	Method for CI	Coverage*($\times$1000)	Power†($\times$1000)
500	**MLE**	Wald	957	790
	CP		954	798
	FLR	PPL§	953	820
	LogF-MCEM		951	808
	LogF-LA		940	795
1000	**MLE**	Wald	955	947
	CP		954	951
	FLR	PPL§	952	958
	LogF-MCEM		951	954
	LogF-LA		950	952

4 Discussion

We have proposed a method for single rare variant analysis with binary phenotypes by logistic regression penalized by log-F priors. Our approach consists of two steps. First, information about m accrues with the number, K, of independent genetic markers are used for estimation. The value of m is the maximizer of a marginal likelihood obtained by integrating the random effects out of the joint distribution of the genetic data and random effects. Our maximization algorithm contains two approximate approaches: (1) a hybrid of an EM algorithm and brute-force maximization of Monte Carlo estimates of the marginal likelihood; and (2) a combination of a Laplace approximation and derivative-free optimization of the marginal likelihood. Second, log-F penalties are conveniently implemented with standard logistic regression by translating each coefficient penalty into a pseudo-data record [9].

Our simulation studies suggest that the proposed log-F-penalized approach is competitive with other methods, and it is slightly better in terms of MSE. It requires extra computation time up front to select the shrinkage parameter m, but once selected, log-F-penalized logistic regression is faster than Firth logistic regression. For example, in our experiments, Firth logistic regression takes approximately three times longer than a log-F-penalized regression with $m = 2$ (results not shown). Moreover, the running time for selecting m can be largely reduced by parallelizing the computation of the marginal likelihood for a given value of m. It is possible to either split the computation of K marginal likelihoods, or the computation of the marginal likelihood for each value of m in the grid search, over different cores.

Our method for estimating m uses multiple variants chosen from a genomic region or from across the genome. Using variants with no effect in this first stage of the analysis leads to large estimates of m, which correspond to strong shrinkage toward zero. Over-shrinkage biases the log-F-penalized estimator towards zero, and reduces power in the second stage of analysis. In the extreme, the use of weakly-associated variants in the first stage can lead to a monotone marginal likelihood in m. For example, we applied our first-stage method to data from the first phase of the Alzheimer's Disease Neuroimaging Initiative (ADNI) study (available at adni.loni.ucla.edu) [1]. We selected 69 approximate independent SNVs over a particular region to estimate the value of m. Figure 4 shows the estimated profile log-likelihood for m over the grid of values $m = 1, 2, ..., 20$ for the MCEM estimator, and $m = 1, 1.1, ..., 20$ for the LA estimator. The log-likelihood curves are of similar shape, though shifted because the MCEM approach estimates the likelihood up to a constant (compare equations (12) and (13)). The likelihood is monotone increasing, leading to a large estimate of m and consequently over-shrinkage of parameter estimates. This is because the SNVs used to estimate m are weakly associated to the phenotype (i.e. the distribution of log-ORs associated with SNVs is highly concentrated about zero). To avoid over-shrinkage we can select SNVs with large marginal effects (i.e., small p-values) from a genome-wide scan for the purpose of estimating the shrinkage parameter m, similar to the SNV-selection process used by FaST-LMM-Select [17]. Investigation of the properties of such an approach is ongoing work.

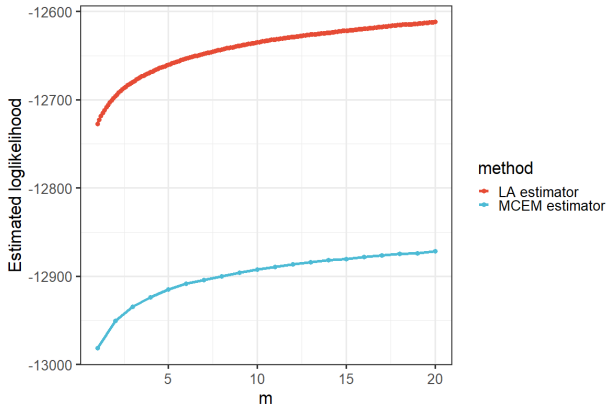

Figure 4: The estimated profile log-likelihood of m constructed using the ADNI-1 case-control data.

It should be noted that we used a simplified, binary confounding variable to represent population stratification, though typically principal components or multidimensional scaling (MDS) components would be used. However, it is straightforward to expand the set of confounding variables in our approach. Another extension of interest is to use log-F penalization for a SNP covariate of interest in a model that uses linear mixed models (LMMs) to correct for confounding due to population structure and genetic relatedness [14, 16]. LMMs can be viewed as either regression with correlated errors, using a kinship matrix derived from anonymous SNPs to model correlations, or as a "whole-genome" regression that includes these anonymous SNPs in the regression model with random polygenic effects. It should be straightforward to extend either formulation to include log-F penalization of the SNP of interest through data augmentation. Investigation of the properties of our approach in conjunction with LMMs is an area for future work.

In practice, identifying rare genetic causes of common diseases can improve diagnostic and treatment strategies for patients. Recent studies have found that patients with low genetic risk scores (GRS) are more likely to carry rare pathogenic variants [18]. Although GRS are currently based on common variants, our method might be of use in extending GRS methods to include low-frequency or even rare variants of large effect sizes.

Rare variants are thought to have larger effect sizes than common variants, and so it is of interest to let the distribution of effect sizes vary with variant frequency. That is, we might allow the prior distribution to be indexed by a variant-frequency-specific parameter instead of a common parameter for all variants. The

idea could be as simple as multiplying the global shrinkage parameter m by a frequency-specific parameter α_k; i.e., a variant in frequency bin k could have prior distribution log-$F(\alpha_k m, \alpha_k m)$. We can choose the α_k values such that the distribution of common variants has a smaller variance for common variants and a larger variance for rare variants. We also note that standardizing the genetic covariate data has the effect of making the coefficients of rare variants more comparable to the coefficients of common variants, which is more consistent with the assumption of a common log-F prior distribution for all variants.

5 Appendix: Computational considerations for MCEM

We use the weighted logistic regression approach to maximize equation (13) over α_k^*. This equation is a weighted average of logistic regression likelihoods, with weights given by the density values $f(\boldsymbol{X}_{.k}|\alpha_k^{*(p)}, \beta_{kj})$. Each likelihood is itself a sum over the n subjects in the dataset. Our approach is to write equation (13) as a weighted likelihood comprised of $N \times n$ observations and use standard logistic regression software to maximize over α_k^*. One way to do this is to "stack" the response vector and covariates N times over as illustrated in Figure **??** and associate with each observation in this augmented dataset a weight and an offset. The weight for each observation in the jth replicate of the dataset is the weight $f(\boldsymbol{X}_{.k}|\alpha_k^{*(p)}, \beta_{kj})$ from the weighted average in equation (13). The offsets account for known quantities in the logistic model. In particular, the linear prediction in the logistic model for observation i in the jth replicate of the dataset is $\alpha_k^* + x_{ik}\beta_{kj}$, where β_{kj} is drawn from the log-$F(m, m)$ distribution and is considered fixed in equation (13). Thus the term $x_{ik}\beta_{kj}$ is a known offset. By constructing the augmented dataset in Figure **??**, maximizing (13) over α_k^* is equivalent to estimating the intercept of a logistic regression and we can use standard logistic regression software, such as `glm()` in `R`, to do this.

$$
\begin{array}{cccc|c}
\boldsymbol{Y} & \boldsymbol{X} & \boldsymbol{W} = \text{weights} & \boldsymbol{O} = \text{offset} \\
\left(\begin{array}{c} \boldsymbol{y} = \left\{\begin{array}{c} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{array}\right. \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \boldsymbol{y} = \left\{\begin{array}{c} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{array}\right. \end{array}\right. &
\begin{array}{c} \boldsymbol{x} = \left\{\begin{array}{c} x_{1k} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ x_{nk} \end{array}\right. \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \boldsymbol{x} = \left\{\begin{array}{c} x_{1k} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ x_{nk} \end{array}\right. \end{array} &
\begin{array}{c} W_1 = f(\boldsymbol{X}_{.k}|\alpha_k^{*(p)}, \beta_{k1}) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ W_N = f(\boldsymbol{X}_{.k}|\alpha_k^{*(p)}, \beta_{kN}) \end{array} &
\begin{array}{c} \boldsymbol{x}\beta_{k1} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \boldsymbol{x}\beta_{kN} \end{array}
\end{array}
$$

Figure 5: $\boldsymbol{Y}_{(Nn\times 1)}$ is a vector containing N replicates of $\boldsymbol{y}$ and $\boldsymbol{X}_{(Nn\times 1)}$ is a vector containing N replicates of $\boldsymbol{x}$. $\boldsymbol{W}$ stands for the weights for each Monte Carlo replicate such that $W_j = f(\boldsymbol{X}_{.k}|\alpha_k^{*(p)}, \beta_{kj})$ and the offset term $\boldsymbol{O} = \{\boldsymbol{x}\beta_{kj}\}_{j=1}^N$.

References

[1] ADNI procedures manual. Retrieved from https://adni.loni.usc.edu/wp-content/uploads/2010/09/ADNI_GeneralProceduresManual.pdf in (February, 2019), (March, 2006). University of California, San Diego.

[2] Siyuan Chen. Approximate marginal likelihoods for shrinkage parameter estimation in penalized logistic regression analysis of case-control data. Master's thesis, Simon Fraser University, 2020.

[3] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[4] Ludwig Fahrmeir and Gerhard Tutz. *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media, 2013.

[5] David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

[6] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, Yu-Sung Su, et al. A weakly informative default prior distribution for logistic and other regression models. *Annals of applied Statistics*, 2(4):1360–1383, 2008.

[7] Jinko Graham, Brad McNeney, and Robert W. Platt. Small sample methods. In Norman Breslow, Oernulf Borgan, Nilanjan Chatterjee, Mitchell H. Gail, Alastair Scott, and Christopher John Wild, editors, *Handbook of Statistical Methods for Case-Control Studies*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, chapter 9, pages 134–162. Chapman and Hall/CRC Press, Boca Raton, Florida, 2018.

[8] Sander Greenland. Prior data for non-normal priors. *Statistics in medicine*, 26(19):3578–3590, 2007.

[9] Sander Greenland and Mohammad Ali Mansournia. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in medicine*, 34(23):3133–3143, 2015.

[10] Georg Heinze, Meinhard Ploner, Daniela Dunkler, and Harry Southworth. logistf: Firth's bias reduced logistic regression. *R package version*, 1, 2013.

[11] Georg Heinze and Michael Schemper. A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16):2409–2419, 2002.

[12] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[13] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[14] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010.

[15] Richard A Levine and George Casella. Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439, 2001.

[16] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835, 2011.

[17] Jennifer Listgarten, Christoph Lippert, Carl M Kadie, Robert I Davidson, Eleazar Eskin, and David Heckerman. Improved linear mixed models for genome-wide association studies. *Nature methods*, 9(6):525–526, 2012.

[18] Tianyuan Lu, Sirui Zhou, Haoyu Wu, Vincenzo Forgetta, Celia MT Greenwood, and J Brent Richards. Individuals with common diseases but with a low polygenic risk score could be prioritized for rare variant screening. *Genetics in Medicine*, 23(3):508–515, 2021.

[19] Brendan Maher. Personal genomes: The case of the missing heritability. *Nature News*, 456(7218):18–21, 2008.

[20] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, Oct 2009.

[21] J. Mbatchou, L. Barnard, J. Backman, A. Marcketta, J. A. Kosmicki, A. Ziyatdinov, C. Benner, C. O'Dushlaine, M. Barber, B. Boutkov, L. Habegger, M. Ferreira, A. Baras, J. Reid, G. Abecasis, E. Maxwell, and J. Marchini. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*, May 2021.

[22] Ross L Prentice and Ronald Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

[23] Jing Qin and Biao Zhang. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84(3):609–618, 1997.

[24] Nicholas J Schork, Sarah S Murray, Kelly A Frazer, and Eric J Topol. Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, 19(3):212–219, 2009.

[25] Alastair J Scott and CJ Wild. Maximum likelihood for generalised case-control studies. *Journal of Statistical Planning and Inference*, 96(1):3–27, 2001.

[26] Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[27] Greg CG Wei and Martin A Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704, 1990.

[28] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[29] Xiaowei Zhan, Youna Hu, Bingshan Li, Goncalo R. Abecasis, and Dajiang J. Liu. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics*, 32(9):1423–1426, 02 2016.

[30] Biao Zhang. Bias-corrected maximum semiparametric likelihood estimation under logistic regression models based on case–control data. *Journal of statistical planning and inference*, 136(1):108–124, 2006.