Shrinkage Parameter Estimation in Penalized Logistic Regression Analysis of Case-Control Data

Ying Yu[1,*], Siyuan Chen[1,2], and Brad McNeney[1]

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

[2]Faculty of Medicine, BC Children's Hospital Research Institute, Vancouver, BC, Canada

[*]Corresponding Author: ying_yu_5@sfu.ca

Abstract

In genetic epidemiology, rare variant case-control studies aim to investigate the association between rare genetic variants and human diseases. Rare genetic variants lead to sparse covariates that are predominately zeros and this sparseness leads to estimators of log-OR parameters that are biased away from their null value of zero. Different penalized-likelihood methods have been developed to mitigate this sparse-data bias for case-control studies. In this research article, 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. We propose a maximum marginal likelihood method for estimating m, with the 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 combination of a Laplace approximation and derivative-free optimization of the marginal likelihood. We evaluate the statistical properties of the estimator through simulation studies and apply the methods to the analysis of genetic data from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

Keywords: Rare variant association studies; Penalized logistic regression; log-F priors; Monte Carlo EM; Laplace approximation

1 Introduction

In genetic epidemiology, inference of associations between disease status and a rare exposure is complicated by the finite-sample bias of the maximum likelihood estimator for logistic regression. Rare variant association studies, which aim to investigate the association between rare single nucleotide variants (SNVs) and human diseases, are prone to sparse data bias [7]. Sparse data bias for estimation of a log-OR parameter is a bias away from the null value of zero that arises when the corresponding covariate is predominately zero. A useful approach to reduce such bias is penalized likelihood, in which the log-OR parameters are viewed as latent variables from a prior distribution. The most widely used penalization is Firth's bias-reduction method, which corresponds to using the Jeffrey's prior distribution as the penalty [6, 11]. Here, we study penalization based on a class of log-F priors developed by Greenland and Mansournia [8]. The family of log-F distributions is indexed by a shrinkage parameter m, which controls the penalty strength. In the penalization of the likelihood, the log-OR parameters are assumed to be independent and to each follow a log-$F(m, m)$ distribution. The penalty term is therefore the product of independent log-$F(m, m)$ priors. For a given m, the log-F penalized likelihood method can be implemented by fitting a standard logistic regression to a dataset augmented with m pseudo-individuals per covariate [8].

Though penalization by log-F prior distributions is straightforward to implement, methods to select the shrinkage parameter m are limited. Greenland and Mansourinia suggested an empirical Bayes method to estimate m from data [8] but did not provide details. In this paper, we follow this suggestion and propose

an empirical Bayes method to estimate the shrinkage parameter m. The context for our approach is that we are interested in fitting single-SNV logistic regressions over a genomic region. From this region we assume a common log-F distribution for the log-OR parameters. We select K approximately independent SNVs from the region 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. We conduct a simulation study to evaluate the performance of our methods under a variety of data-generating scenarios, and apply the methods to real data from a genetic association study of Alzheimer's disease.

2 Models and Methods

2.1 Case-control Likelihood

Consider a case-control study of association between a rare SNV and disease status, in which the n_0 controls are indexed by $i = 1, ..., n_0$ and n_1 cases are indexed by $i = n_0 + 1, ..., n$ for $n = n_0 + n_1$. Let Y_i be a binary response indicating the disease status and X_i be the covariate data for subject i, and let β be the log-OR parameter of the model. Qin and Zhang [17] expressed the case-control likelihood in terms of a two-sample semi-parametric model as follows

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

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. It is more convenient to obtain the MLE of β by maximizing the profile likelihood [15]:

$$\begin{aligned} L(\alpha^*, \beta) &= f(\boldsymbol{X}|\alpha^*, \beta) \\ &= \prod_{i=1}^{n_0} \frac{1}{1 + \exp(\alpha^* + X_i\beta)} \prod_{i=n_0+1}^{n_0+n_1} \frac{\exp(\alpha^* + X_i\beta)}{1 + \exp(\alpha^* + X_i\beta)} \\ &= \prod_{i=1}^{n} \frac{\exp(Y_i(\alpha^* + X_i\beta))}{1 + \exp(\alpha^* + X_i\beta)}, \end{aligned} \tag{2}$$

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 [18]. The profile likelihood $L(\alpha^*, \beta)$ for case-control data is in 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 [15, 17].

2.2 Marginal likelihood for estimating m

Suppose we have K (approximately) independent covariates. For each we specify a one-covariate logistic regression model. Let $\boldsymbol{X}_{.k}$ $k = 1, \ldots, K$, denote the data on the k^{th} covariate and $\boldsymbol{X}$ denote a matrix containing all K. Let $L(\alpha_k^*, \beta_k)$ denote the profile likelihood (equation 2)for the kth log-OR parameter β_k. Here α_k^* is the intercept term from the k^{th} 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:

$$L(\boldsymbol{\alpha}^*, \boldsymbol{\beta}) f(\boldsymbol{\beta}|m) = \prod_{k=1}^{K} L(\alpha_k^*, \beta_k) f(\beta_k|m), \tag{3}$$

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:

$$\begin{aligned} L(\boldsymbol{\alpha}^*, m) &= \int L(\boldsymbol{\alpha}^*, \boldsymbol{\beta}) f(\boldsymbol{\beta}|m) d\boldsymbol{\beta} \\ &= \int \prod_{k=1}^{K} L(\alpha_k^*, \beta_k) f(\beta_k|m) d\boldsymbol{\beta} \\ &= \prod_{k=1}^{K} \int L(\alpha_k^*, \beta_k) f(\beta_k|m) d\beta_k. \\ &= \prod_{k=1}^{K} L(\alpha_k^*, m), \end{aligned} \tag{4}$$

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

$$l(\boldsymbol{\alpha}^*, m) = \sum_{k=1}^{K} \log[L(\alpha_k^*, m)] = \sum_{k=1}^{K} l(\alpha_k^*, m). \tag{5}$$

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 equation (4). We discuss two approaches. The first (Section 2.3) is a Monte Carlo EM algorithm [5], and the second (Section 2.4) 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

$$L(\alpha_k^*, \gamma_k, \beta_k) = \prod_{i=1}^{n} \frac{\exp(Y_i(\alpha^* + Z_i\gamma_k + X_i\beta))}{1 + \exp(\alpha^* + Z_i\gamma_k + X_i\beta)} \tag{6}$$

and the marginal log-likelihood for estimating m is

$$\begin{aligned} l(\boldsymbol{\alpha}^*, \boldsymbol{\gamma}, m) &= \sum_{k=1}^{K} l(\alpha_k^*, \gamma_k, m) \\ &= \sum_{k=1}^{K} \log \int L(\alpha_k^*, \gamma_k, \beta_k) f(\beta_k|m) d\beta_k. \end{aligned} \tag{7}$$

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 Monte Carlo EM Algorithm

To maximize $l(\alpha_k^*, m)$ in stage 1, we first consider an EM algorithm. In our setting, $\boldsymbol{X}_{.k}$ is the observed data and β_k is the unobserved latent variable. 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

$$\begin{aligned} Q\left(\alpha_k^* | \alpha_k^{*(p)}, m\right) &= \mathrm{E}_{\beta_k | \boldsymbol{X}_{.k}, \alpha_k^{*(p)}, m}(\log[f(\boldsymbol{X}_{.k}, \beta_k | \alpha_k^*, m)]) \\ &= \int \log\left[f(\boldsymbol{X}_{.k}, \beta_k | \alpha_k^*, m)\right] f(\beta_k | \boldsymbol{X}_{.k}, \alpha_k^{*(p)}, m) d\beta_k \\ &\propto \int \log[f(\boldsymbol{X}_{.k} | \alpha_k^*, \beta_k) f(\beta_k | m)] f(\boldsymbol{X}_{.k} | \alpha_k^{*(p)}, \beta_k) f(\beta_k | m) d\beta_k \end{aligned} \tag{8}$$

and the M-step is to set

$$\alpha_k^{*(p+1)} = \underset{\alpha_k^*}{\operatorname{argmax}}\, Q\left(\alpha_k^* | \alpha_k^{*(p)}, m\right). \tag{9}$$

The E-step (8) 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 [22]. The MC integration in the E-step is obtained by sampling from the prior distribution $f(\beta_k | m)$ [12, 22]. Based on a sample $\beta_{k1}, ..., \beta_{kN}$ from the distribution $f(\beta_k | m)$, the MC approximation to the integral is

$$\begin{aligned} Q\left(\alpha_k^* | \alpha_k^{*(p)}, m\right) &\approx Q_{MC}\left(\alpha_k^* | \alpha_k^{*(p)}, m\right) \\ &= \frac{1}{N} \sum_{j=1}^{N} \log[f(\boldsymbol{X}_{.k} | \alpha_k^*, \beta_{kj}) f(\beta_{kj} | m)] f(\boldsymbol{X}_{.k} | \alpha_k^{*(p)}, \beta_{kj}) \\ &= \frac{1}{N} \sum_{j=1}^{N} (\log[f(\boldsymbol{X}_{.k} | \alpha_k^*, \beta_{kj})] + \log[f(\beta_{kj} | m)]) f(\boldsymbol{X}_{.k} | \alpha_k^{*(p)}, \beta_{kj}). \end{aligned} \tag{10}$$

Note that $\log[f(\beta_{kj} | m)]$ is independent of the parameter α_k^*, so maximizing (10) in the M-step is equivalent to maximizing

$$\frac{1}{N} \sum_{j=1}^{N} \log[f(\boldsymbol{X}_{.k} | \alpha_k^*, \beta_{kj})] f(\boldsymbol{X}_{.k} | \alpha_k^{*(p)}, \beta_{kj}). \tag{11}$$

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

2.4 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 [20]. 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

$$\tilde{L}(\alpha_k^*, m) = L(\alpha_k^*, \hat{\beta}_k) f(\hat{\beta}_k | m) \sqrt{\frac{2\pi}{c_p(\alpha_k^*)}}. \tag{12}$$

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 [3]. Figure 1 shows the quality of the LA for one simulated dataset generated under $m = 4$.

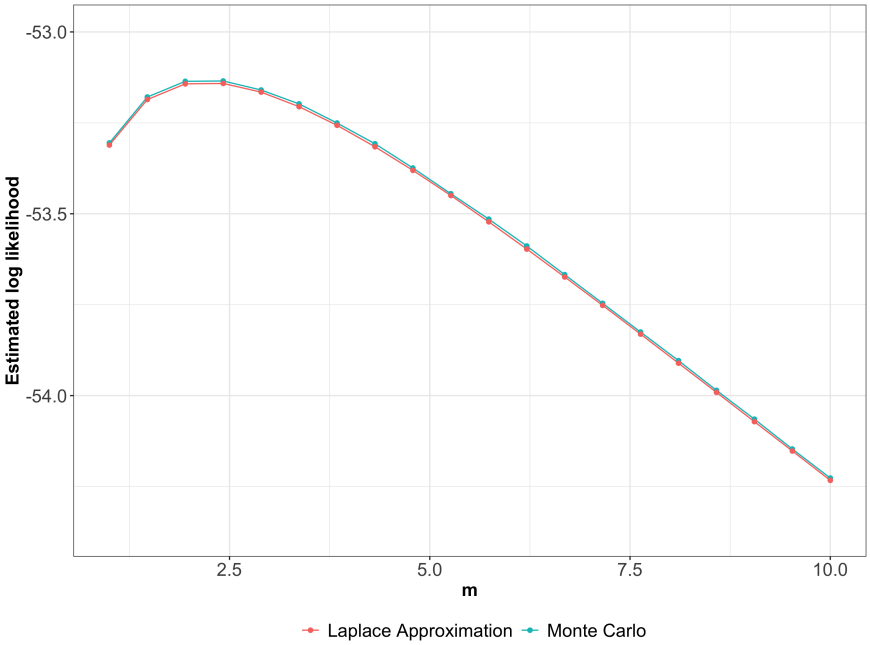

Figure 1: Natural logarithms of estimates of the marginal likelihood $L(\alpha_k^*, m)$ for one simulated dataset generated under $m = 4$. Estimates are by LA and Monte Carlo. Log-likelihood estimates are plotted over the grid $m = (1, 1.5, ..., 10)$ with α_k^* = -3

The approximate marginal likelihod $\tilde{L}(\alpha_k^*, m)$ may be maximized over α^* using standard derivative-free optimization methods, such as a golden section search or the Nelder-Mead algorithm.

3 Results: Simulated data

Data simulation methods are discussed in Appendix B. Results are shown in Tables 1-3. The results from simulations under $m = 4$ (Table 1) suggest that the MCEM- and LA-based marginal likelihood estimators of m are both slightly upwardly biased and that this bias decreases as the number of variants K increases. The SD of the distribution of the LA estimator is smaller than that of the MCEM estimator when the covariates are continuous, and the SDs of the two estimators are comparable when the covariates are SNVs.

The results from simulations under $m = 10$ (Table 2) also suggest an upward bias of both estimators, but suggest that the upward bias of the LA-based estimator is greater than that of the MCEM-based estimator. Also, the bias of the LA-based estimator does not appear to improve with K when the covariates are SNVs. When the true m is 10, the SD of the distribution of the LA estimator appears *larger* than that of the MCEM estimator, in contrast to the results when $m = 4$.

We also simulated data in the presence of population stratification; see Appendix B for details. Let Z denote a binary indicator of one of the two population strata. We create population-disease and population-SNV associations as follows. To create population-disease association we took the population-stratum log-OR, γ, to be either 1 or a random value simulated from a log-F distribution. To induce population-SNV association we selected SNVs with absolute allele frequency differences, Δ, of 0.3 or greater. Comparing results when we ignore or include the fixed effect of the population confounding variable (Table 3 panels (b) and (c), respectively) we see that ignoring population structure increases the upward bias of the estimator of m. Such bias is a consequence of the attenuating effect of ignoring confounding variables. In particular,

ignoring confounding variables leads to attenuated estimates of log-OR parameters of interest, which would suggest a distribution of log-OR estimates that is overly concentrated about zero, and consequently a larger value of m (recall that m is a precision parameter, with larger values indicating smaller variance).

Results for random population effects drawn from the log-$F(m, m)$ distribution (see Table 3 (a)) are qualitatively similar, though less pronounced because the population-disease association is generally weaker under random effects than under the fixed effect size in our simulations.

Continuous covariates						
Number of variants	K=10		K=30		K=50	
Methods	MCEM	LA	MCEM	LA	MCEM	LA
Mean	5.124	5.191	4.415	4.644	4.161	4.387
SD	3.569	2.215	2.192	1.145	1.160	0.726
95% CI	(4.425, 5.824)	(4.757, 5.625)	(3.985, 4.844)	(4.420, 4.868)	(3.934, 4.388)	(4.245, 4.529)
SNP covariates						
Number of variants	K=10		K=30		K=50	
Methods	MCEM	LA	MCEM	LA	MCEM	LA
Mean	5.568	5.451	4.895	5.176	4.694	4.783
SD	3.457	2.477	1.300	1.408	1.244	1.127
95% CI	(4.890, 6.245)	(4.966, 5.936)	(4.640, 5.150)	(4.900, 5.452)	(4.450, 4.938)	(4.562, 5.004)

Table 1: Simulations results for 100 datasets on continuous and SNP covariates. Mean and standard deviation are obtained after smoothing. True $m = 4$.

Continuous covariates						
Number of variants	K=10		K=30		K=50	
Methods	MCEM	LA	MCEM	LA	MCEM	LA
Mean	11.230	12.801	10.970	11.872	10.514	11.514
SD	4.279	5.262	2.789	3.101	2.596	2.696
95% CI	(10.392, 12.069)	(11.770, 13.832)	(10.423, 11.516)	(11.264, 12.480)	(10.005, 11.023)	(10.986, 12.042)
SNP covariates						
Number of variants	K=10		K=30		K=50	
Methods	MCEM	LA	MCEM	LA	MCEM	LA
Mean	11.840	14.393	11.549	14.601	10.716	14.053
SD	4.315	5.215	3.569	4.529	2.694	3.741
95% CI	(10.994, 12.686)	(13.371, 15.415)	(10.850, 12.248)	(13.713, 15.489)	(10.188, 11.244)	(13.320, 14.786)

Table 2: Simulations results for 100 datasets on continuous and SNP covariates. Mean and standard deviation are obtained after smoothing. True $m = 10$.

(a)						
Number of variants	K=10		K=30		K=50	
Methods	MCEM	LA	MCEM	LA	MCEM	LA
Mean	6.966	5.316	6.239	4.701	5.791	4.813
SD	4.410	2.315	2.823	1.483	2.155	1.218
95% CI	(6.101, 7.830)	(4.862, 5.770)	(5.686, 6.792)	(4.410, 4.992)	(5.368, 6.213)	(4.574, 5.052)
(b)						
Number of variants	K=10		K=30		K=50	
Methods	MCEM	LA	MCEM	LA	MCEM	LA
Mean	6.487	5.756	5.556	5.024	5.334	4.614
SD	3.422	2.417	2.052	1.548	1.883	1.030
95% CI	(5.817, 7.158)	(5.282, 6.230)	(5.154, 5.958)	(4.721, 5.327)	(4.964, 5.703)	(4.412, 4.816)
(c)						
Number of variants	K=10		K=30		K=50	
Methods	MCEM	LA	MCEM	LA	MCEM	LA
Mean	8.590	7.536	7.743	7.168	7.222	6.998
SD	3.906	2.216	2.790	1.826	1.855	1.602
95% CI	(7.824, 9.355)	(7.102, 7.970)	(7.197, 8.291)	(6.810, 7.526)	(6.852, 7.591)	(6.684, 7.312)

Table 3: Simulations results for 100 datasets on SNP covariates. Mean and standard deviation are obtained after smoothing. True $m = 4$, $\Delta \geq 0.3$. (a) With adjustment for population strata; $\gamma \sim$ log-$F(4, 4)$; (b) With adjustment for population; $\gamma = 1$; (c) Same simulated data as in (b) but without adjustment for population.

4 Results: ADNI data

Alzheimer's Disease (AD) is the most common cause of dementia (a general form of memory loss) and loss of cognitive abilities. The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal multicenter study designed to identify significant genetic variants for the early detection and tracking of AD [1]. We illustrate our methodology by applying it to a dataset obtained from the first phase of ADNI study (abbreviated as ADNI-1). More information about the ADNI-1 study design are available on the ADNI website (adni.loni.ucla.edu). After quality control, filtering and imputation, the data comprise information on 490 SNVs across 33 genes obtained from a total of 632 subjects, of which 179 are cognitively normal (CN), 144 have Alzheimer's Disease (AD) and 309 are in late mild cognitive impairment (LMCI) stage.

In this study we are interested in identifying SNVs that are associated with AD. The association between SNVs and the AD phenotype is estimated by penalized logistic regression with a log-$F(m, m)$ prior, adjusting for confounding variables, such as age, gender, Apolipoprotein E (APOE) genotype and the first 10 multidimensional scaling components (MDS). The APOE genotype is a genetic risk factor for AD, which is located in chromosome 19 and contains three different alleles: e2, e3, e4 [10]. Individuals carrying the e4 allele have a higher risk of developing AD than those carrying the more common e3 allele, whereas the e2 allele decreases the risk [13]. The first 10 MDS components are used to correct for ancestry and population stratification. The logistic regression model for this single-SNP association test is

$$\text{Y} = \text{single-SNP} + \text{Age} + \text{Gender} + \text{APOE genotypes} + 10\ \text{MDS}.$$

We include $n = 323$ (179 CN and 144 AD) subjects from a dataset prepared for Greenlaw et al. [9]. We restrict our analyses to a genomic region spanning the NEDD9 gene on chromosome 6, and we apply our methods on 69 SNPs over this region to estimate the value of m.

Figure 2 displays 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 11 and 10). The monotone increasing likelihood curves suggest a large value of m, indicating that the variance of the log-OR distribution is small; i.e., that the distribution of log-ORs is highly concentrated about zero.

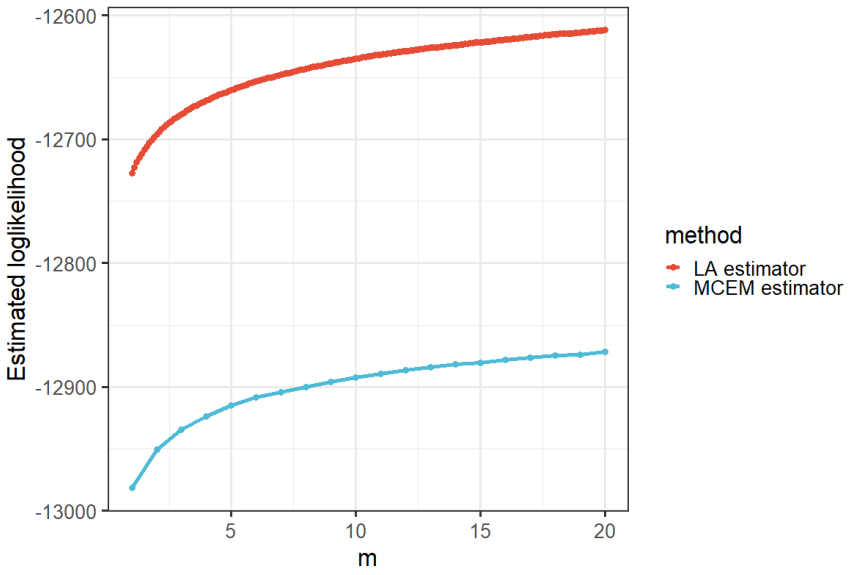

Figure 2: The estimated profile log-likelihood of m constructed using the ADNI-1 case-control data.

5 Discussion

We have proposed a method to select the shrinkage parameter m in penalized logistic regression analysis for case-control studies with log-$F(m, m)$ prior. Information about m accrues with the number, K, of independent genetic markers 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. The methods are applied to simulated data and to a real dataset from the ADNI database. Our simulation studies suggest an upward bias of both estimator, and the bias appears to improve with the number of genetic variants included in the model. These two methods are comparable when m is small, but the MCEM-based estimator outperforms the LA-based estimator when estimating large m in terms of both bias and variance. The EM algorithm and Monte Carlo integration are time-consuming, so the LA-based approach is a computationally cheaper alternative choice. Adjustment for confounding variables is required to avoid further upward bias. It should be noted that we simplified the confounding variable that caused by population stratification to a binary indicator, though typically principal components or MDS components would be used to correct for population stratification in genetic association studies; investigation of the properties of our approach when PC adjustment is used will be considered in future work.

The method we proposed is based on summarizing the joint effects of multiple variants within a genomic region to estimate the shrinkage parameter m. Regions can be defined by a specific gene or moving windows across the genome. The SKAT testing approach of Wu *et al.* [23] also employs a moving-window and a random-SNV-effects model. Our estimation approach consists of two steps. For each region, we first determine m based on variants effects on a phenotype while adjusting for other covariates, and then perform single-variant association test with log-$F(m, m)$ prior through data augmentation suggested by Greenland and Mansournia [8]. The augmented dataset is constructed by adding one pseudo-observation with $\frac{m}{2}$ success and $\frac{m}{2}$ failures to the response, and a single row to the design matrix consisting all zeros except for a one indicating the index of the variant [8]. Analyzing the augmented dataset with standard logistic regression gives the penalized estimates and corresponding standard errors. Analyzing multiple regions and variants requires adjustments for multiple comparisons, for example with the Bonferroni correction and FDR control. Such region-based analysis can also be applied to any meaningful set of variants.

Numerous studies have showed that shrinkage is an effective way to solve issues that arise in datasets where sampling variability is large [2, 16, 21, 14]. Performance of shrinkage methods strongly depends on the choice of the shrinkage parameters. A classical approach is to tune the parameter by minimizing some measure of mean squared errors (MSE) using cross validation or some information criterion. A recent study, however, noted that in small or sparse datasets, the optimized values of parameter obtained by such tuning procedures are negatively correlated with the optimal amount of shrinkage and suffer from large variability which leads to large MSE [19]. In this study, Šinkovec et.al suggest that pre-specifying the degree of shrinkage is beneficial to give accurate coefficients and predictions under non-ideal datasets with rare outcomes or sparse covariates [19]. Their findings give support to a two-step penalized regression approach that includes our method in the first step to determine the shrinkage parameter m before applying individual association tests. The log-F priors used in our study fall in the class of zero-centered informative priors condisered by Šinkovec et.al [19].

Funding: This work was supported, in part, by a Discovery Grant to Brad McNeney from the Natural Sciences and Engineering Research Council of Canada (NSERC).

6 Appendix

6.1 Appendix A: Computational considerations

We use the weighted logistic regression approach to maximize equation (11) 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 (11) 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 3 and associate with each observation in this augmented dataset a weight and an offset. The weight for each observation in the j^{th} replicate of the dataset is the weight $f(\boldsymbol{X}_{.k}|\alpha_k^{*(p)}, \beta_{kj})$ from the weighted average in equation (2.15). The offsets account for known quantities in the logistic model. In particular, the linear prediction in the logistic model for observation i in the j^{th} 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 (11). Thus the term $x_{ik}\beta_{kj}$ is a known offset.

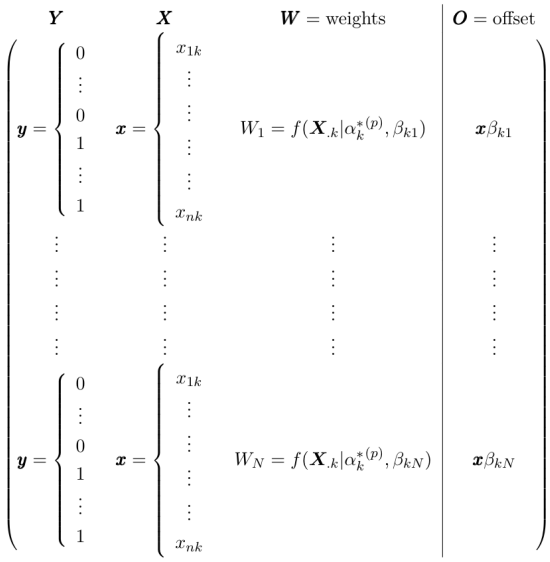

Figure 3: $\boldsymbol{Y}_{(Nn\times 1)}$ is a vector containing N replicates of $\boldsymbol{y}$ and $\boldsymbol{X}_{(Nn\times 1)}$ is a vector containing N replicates of $\boldsymbol{x}$. $\boldsymbol{W}$ stands for the weights for each Monte Carlo replicate such that $W_j = f(\boldsymbol{X}_{.k}|\alpha_k^{*(p)}, \beta_{kj})$ and the offset term $\boldsymbol{O} = \{\boldsymbol{x}\beta_{kj}\}_{j=1}^{N}$.

By constructing the augmented dataset in Figure 3, maximizing (11) 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.

6.2 Appendix B: Simulation methods

We apply our methods to simulated data. We simulated both continuous and SNP covariates. For a given m, K independent random covariate effects are sampled from a log-$F(m,m)$ distribution, by sampling first from a $F(m,m)$ distribution and then taking logs. We consider $K = 10$, 30 and 50.

Following [24], the conditional density function for the covariate in the controls and cases are

$$P(X = x|Y = 0) = g(x) \text{ and} \tag{13}$$
$$P(X = x|Y = 1) = h(x) = c(\beta, g)\exp(x\beta)g(x). \tag{14}$$

To generate continuous covariates we take $g(x)$ to be the standard normal density $g(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \propto e^{-x^2/2}$ [24]. For a given β, the conditional density function for the covariate in cases is

$$P(X = x|Y = 1) = h(x) \propto \exp(x\beta)g(x) \propto \exp(-\frac{1}{2}(x-\beta)^2). \tag{15}$$

Thus, h(x) is the normal density function with mean β and standard deviation 1. The covariates in the control group are sampled from the standard normal distribution $N(0,1)$ and the covariates in the case group are sampled from the normal distribution $N(\beta,1)$. To generate a SNV with minor allele frequency (MAF) p, we suppose a Binomial$(2,p)$ distribution in controls. The SNV distribution $h(x)$ in cases is then proportional to

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

which has normalizing constant $(1-p)^2 + 2p(1-p)\exp(\beta) + p^2\exp(2\beta)$.

To mimic real situations we sampled MAFs from a site frequency spectrum (SFS) of a real sample of SNVs (see Figure 4). We used data from 1000 genomes project [4], restricting to European subjects (CEU, TSI, FIN, GBR, and IBS). 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 spanning the gene NEDD9 on chromosome 6. This process lead to genotypes on 8725 SNVs of 503 individuals (SNVs with zero MAF have been removed). MAFs were sampled from the kernel density estimation of this empirical SFS by inverse-CDF method. Preliminary simulation experiments suggested that variants with low MAF have very little information about m, and so we restrict sampling to common variants with MAF $\geq 5\%$.

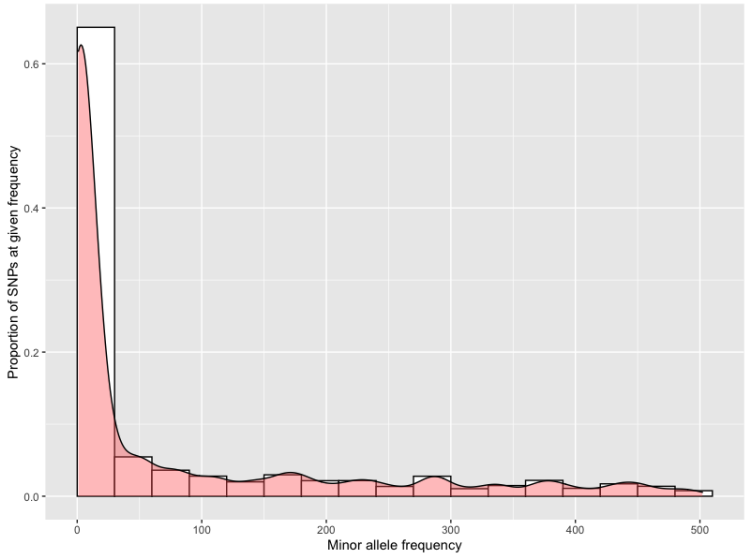

Figure 4: The (folded) site frequency spectrum of 503 European individuals.

For some of the simulations, we also included population substructure as a confounding variable. We create population-SNV association by selecting different SNV allele frequencies in different populations, and population-disease correlation by a population main effect on disease risk. We first simulate population status, and then simulate SNVs conditional on population status. Let population Z take values 0 or 1 with frequency f_0 and f_1, respectively in the general population, which we assume to be the respective frequencies in controls of the two populations as well. Let γ denote the log-OR for the population, the distribution of Z in controls, $P(Z = z|D = 0)$, is Bernoulli(f_z), and the distribution of Z in cases $P(Z = z|D = 1) \propto f_z \exp(z\gamma)$ [24].

Now suppose that the allele frequency for a given SNV X differs by sub-population, denoted as p_z in population z. Let $g_z(x)$ denote the distribution of X in population z, i.e., $P(X = x|Z = z, D = 0) = g_z(x) \sim$ Binomial $(2, p_z)$. The joint distribution of X and Z in controls is then $P(X = x, Z = z|D = 0) = f_z g_z(x)$. Let β denote the log-OR for the SNV. By assuming a joint disease risk model with logit$[P(D = 1|Z = z, X = x)] = \alpha + z\gamma + x\beta$, the joint distribution of X and Z in cases is $P(X = x, Z = z|D = 1) \propto f_z g_z(x) \exp(z\gamma + x\beta)$ [24]. We then have

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

Here we consider two populations: Caucasian (CEU) and Yoruba (YRI) subjects [4]. The joint (folded) SFS is based on the data from 1000 genomes project, summed across a set of approximate 2000 common SNPs genotyped in these two populations. Small correlation in the allele frequencies of these two populations is illustrated in Figure 5 that most of the SNPs are not failing along the diagonal of the SFS. In our simulation, we sampled allele frequencies from this empirical joint SFS.

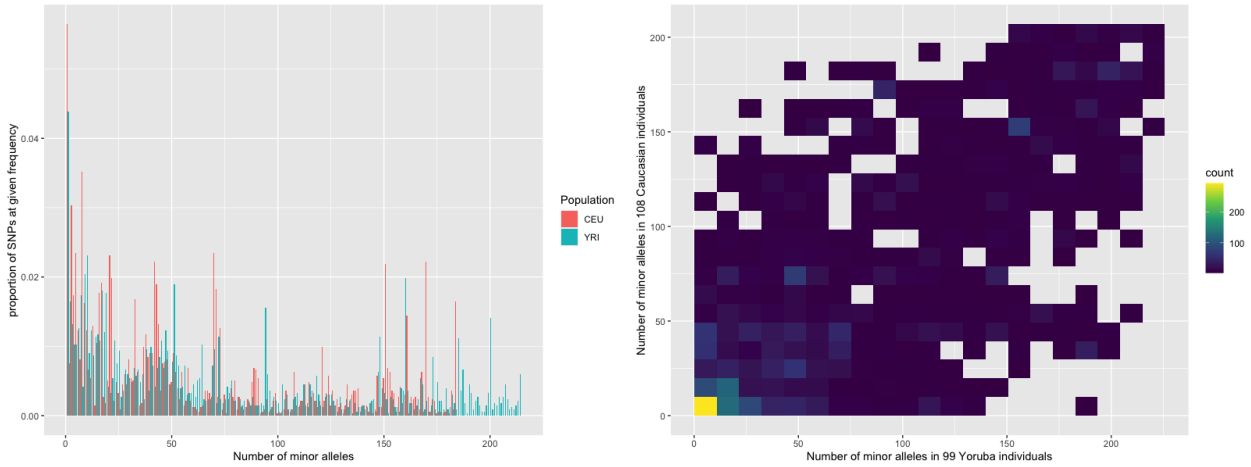

Figure 5: Information in the site frequency spectrum of Caucasian and Yoruba individuals.

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] Rok Blagus and Jelle J Goeman. Mean squared error of ridge estimators in logistic regression. *Statistica Neerlandica*, 74(2):159–191, 2020.

[3] 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.

[4] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

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

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

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

[8] 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.

[9] Keelin Greenlaw, Elena Szefer, Jinko Graham, Mary Lesperance, Farouk S Nathoo, and Alzheimer's Disease Neuroimaging Initiative. A bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics*, 33(16):2513–2522, 2017.

[10] Mary N Haan and Elizabeth R Mayeda. Apolipoprotein e genotype and cardiovascular diseases in the elderly. *Current cardiovascular risk reports*, 4(5):361–368, 2010.

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

[13] Chia-Chen Liu, Takahisa Kanekiyo, Huaxi Xu, and Guojun Bu. Apolipoprotein e and alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology*, 9(2):106, 2013.

[14] Menelaos Pavlou, Gareth Ambler, Shaun Seaman, Maria De Iorio, and Rumana Z Omar. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in medicine*, 35(7):1159–1177, 2016.

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

[16] Rainer Puhr, Georg Heinze, Mariana Nold, Lara Lusa, and Angelika Geroldinger. Firth's logistic regression with rare events: accurate effect estimates and predictions? *Statistics in medicine*, 36(14):2302–2317, 2017.

[17] 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.

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

[19] Hana Šinkovec, Georg Heinze, Rok Blagus, and Angelika Geroldinger. To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. *arXiv preprint arXiv:2101.11230*, 2021.

[20] 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.

[21] Maarten van Smeden, Karel GM Moons, Joris AH de Groot, Gary S Collins, Douglas G Altman, Marinus JC Eijkemans, and Johannes B Reitsma. Sample size for binary logistic prediction models: beyond events per variable criteria. *Statistical methods in medical research*, 28(8):2455–2474, 2019.

[22] 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.

[23] 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.

[24] 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.