Penalized Likelihood Methods for Sparse Datasets, with Applications to Genetic Epidemiology

by

Ying Yu

M.Sc., Simon Fraser University, 2019
B.Sc., The University of British Columbia, 2017

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2023

Declaration of Committee

Name:	Ying Yu
Degree:	**Doctor of Philosophy**
Thesis title:	**Penalized Likelihood Methods for Sparse Datasets, with Applications to Genetic Epidemiology**
Committee:	**Chair:** David Stenning Assistant Professor, Statistics and Actuarial Science
	Brad McNeney Supervisor Associate Professor, Statistics and Actuarial Science
	Lloyd Elliott Committee Member Assistant Professor, Statistics and Actuarial Science
	Jinko Graham Examiner Professor, Statistics and Actuarial Science
	Angelo Canty External Examiner Associate Professor, Mathematics and Statistics McMaster University

Abstract

Increasingly, logistic regression methods for genetic association studies of binary phenotypes must be able to accommodate data sparsity, which arises from unbalanced case-control ratios and/or rare exposures. Sparseness leads to maximum likelihood estimates (MLEs) of log odds-ratio parameters that are biased away from their null value of zero and tests with inflated type I errors. Different penalized-likelihood methods have been developed to mitigate sparse-data bias. We study penalized logistic and conditional regression using a class of log-F priors indexed by a shrinkage parameter m to shrink the biased MLE towards zero. The thesis is organized in three parts. First, we propose a two-step methodology for implementing log-F penalization for inference of regression parameters from logistic regression, with application to genome-wide association studies. In the first step we estimate the shrinkage parameter, and in the second step we use the penalized regression estimator to estimate single-variant associations across the genome. Next, we explore log-F penalization for inference of regression parameters from conditional logistic regression, with application to data from matched case-control and case-parent trio studies. In the first two projects we use simulation to study the statistical properties of our methods and make comparisons to methods that use Firth penalization. Finally, we apply log-F-penalized logistic regression to data from the UK Biobank, to investigate the method's feasibility for genome-wide, biobank-scale data. The complexity and size of biobank data present unique challenges, and we make modifications to our methodology to increase its flexibility and adaptability to such datasets.

Keywords: rare-variant analysis; penalized logistic regression; conditional logistic regression; sparse-data bias; empirical Bayes; UK Biobank

Dedication

It is not our abilities that show what we truly are.
It is our choices.

- J.K. Rowling, Harry Potter and the Chamber of Secrets

Acknowledgements

I would first like to express my sincere gratitude to my supervisor Dr. Brad McNeney for bringing me to the field of statistical genetics, and for his generous patience, support, and encouragement throughout my PhD studies at Simon Fraser University. I greatly appreciate the effort and insights he has contributed to this thesis.

I would also like to express my thanks to my examining committees, Dr. Jinko Graham, Dr. Lloyd Elliott, and Dr. Angelo Canty for taking their time to participate in my defense. My deep gratitude goes out to all the staff and faculty in the department of statistics and actuarial science for their great help and support. Furthermore, I feel incredibly fortunate to have encountered numerous exceptional individuals who have both inspired and guided me during my undergraduate years at The University of British Columbia. A special acknowledgment goes to Dr. Lang Wu for sparking my interest in statistics and providing encouragement on my path towards pursuing a PhD degree. I am also thankful to my colleagues in John Staples' lab at Vancouver General Hospital for having me worked for their research projects and seeing the potential in me.

In addition to my academic journey, my warmest thanks are reserved for my friends, neighbors, and fellow graduate students who have enriched my experience immeasurably. A special mention is deserved by my beloved dog, Cola, whose constant companionship brings a wealth of happiness to my life.

Perhaps most importantly of all, I would like to express my deepest appreciation to my parents, Dr. Huimin Yu and Mrs. Ying Xiao, for their endless love, support and encouragement over years.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

In this thesis we study penalized likelihood estimators for dealing with sparse data bias in inference of regression parameters from logistic and conditional logistic regression. Sparse data bias is a bias of the maximum likelihood estimator that arises when there are small numbers of cases or controls for some combination of the explanatory variables [16]. For example, in logistic regression analysis, when the case-control ratio is unbalanced or the exposure is rare [18], standard logistic likelihood-based inference can have an inflated Type I error [60] and biased maximum likelihood estimator (MLE) of the odds ratio (OR) [10]. Genome-wide association studies (GWAS) often encounter sparse data bias due to the rarity of certain genetic variants. This challenge can make it difficult to detect significant associations between genetic factors and diseases or traits of interest. A related issue arises in conditional logistic regression, which is commonly used to analyze binary outcomes in stratified or matched sampling designs with numerous small strata or matched sets [23]. In this situation, the conditional MLE of regression coefficients may also be biased away from zero [15, 17, 23]. Moreover, when separation occurs — meaning that one or more predictor variables perfectly predict the outcome — the MLE becomes infinite [1]. These concerns emphasize the importance of using appropriate statistical techniques and methods to mitigate the effects of sparse data bias in logistic and conditional logistic regression analysis. Such techniques may include penalized regression methods, Bayesian approaches, or alternative estimators that are less sensitive to sparse data.

Firth logistic regression is a widely used method for addressing sparse-data bias in logistic regression analysis [10]. Software tools such as EPACTS and REGENIE incorporate testing and effect estimation based on Firth logistic regression [10, 24]. Furthermore, Heinze and Puhr [23] developed point- and interval-estimators of ORs for *conditional* logistic regression following Firth's approach [10]. Firth's method employs penalized-likelihood inference, with the penalty term derived from the Jeffreys prior [27]. However, this prior is data-dependent and not considered a true subjective prior, as it relies on the observed data [13, 19].

Various alternative penalties have been proposed for logistic regression, offering different degrees of shrinkage compared to the Jeffreys prior [14]. Greenland and Mansournia introduced a penalized logistic regression approach based on a class of log-F priors, indexed by a shrinkage parameter m [19]. This parameter determines the degree of penalization applied to the regression coefficients, allowing for more flexible control over the bias-variance tradeoff. By selecting an appropriate shrinkage parameter, researchers can select the amount of regularization applied to the model. In our context, log-$F(m,m)$ penalization amounts to assuming that the log-OR parameter β of interest has a log-$F(m,m)$ distribution with density

$$f(\beta|m) = \frac{1}{B(\frac{m}{2},\frac{m}{2})}\frac{\exp(\frac{m}{2}\beta)}{(1+\exp(\beta))^{m}}, \tag{1.1}$$

where $B(\cdot,\cdot)$ is the beta function (see Figure 1.1 for plots of log-$F(1,1)$ and log-$F(10,10)$ density curves). In the log-F penalization approach, maximizing the posterior density is equivalent to maximizing a penalized likelihood obtained by multiplying the logistic regression likelihood by the log-$F(m,m)$ prior. In Chapter 3 we develop log-F-penalized *conditional* logistic regression and show that it is an extension of existing methods for the analysis of sparse matched pairs data.

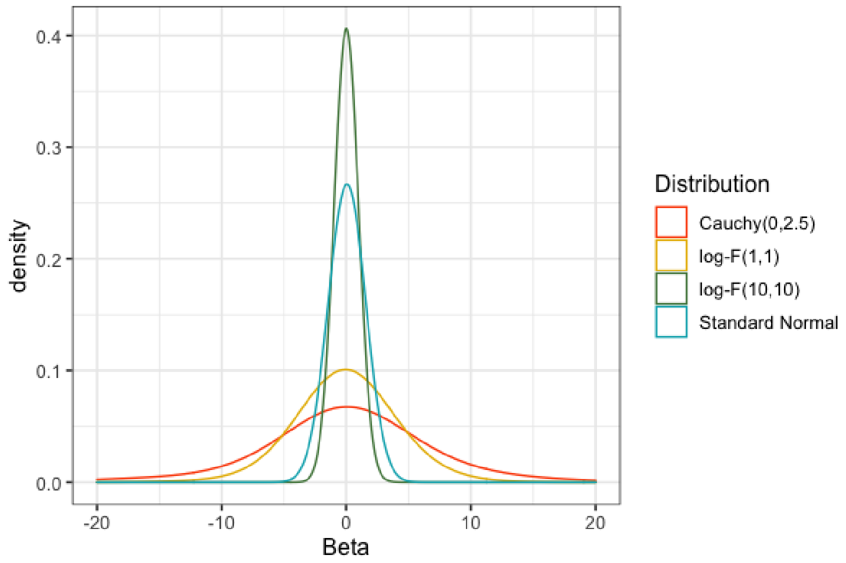


Figure 1.1: Comparison of log-F, standard normal and Cauchy distributions. The log-$F(m,m)$ density is symmetrically bell-shaped with a single peak at zero, and its variance decreases as increasing m. As $m \to \infty$, the distribution tends toward a point mass at zero.

Comparing log-F and Firth approaches is not straightforward due to their different penalization strategies. The log-F approach penalizes selectively, whereas the Jeffreys prior used in

Firth penalization is a function of the Fisher information matrix for all coefficients, including the intercept. However, some understanding can be derived by comparing approaches for matched pairs data and a single binary exposure. For matched pairs, the standard analysis employs conditional logistic regression, which eliminates intercept terms from the likelihood. It can be demonstrated that when the model only include a single binary exposure, Firth conditional logistic regression equals log-$F(1,1)$-penalized conditional logistic regression. This equivalency can be achieved through Haldane correction, as described by Greenland and Mansournia [19]. Appendix C.1 provides a detailed proof of the equivalence of these two methods under a matched case-control study design.

The log-F penalization approach exhibits several desirable properties compared to the Firth method. First, the log-F approach for logistic regression can be easily implemented through simple data augmentation, as detailed by Greenland and Mansournia [19]. Penalization can be implemented by adding pseudo-observations to the original dataset. Subsequently, analyzing the augmented data with standard logistic regression produces the desired penalized estimates and corresponding standard errors. In contrast, the Firth method requires a more complex iterative algorithm, which is computationally demanding. In Chapter 3 we develop an analogous data augmentation scheme to implement log-$F(m,m)$ penalization in conditional logistic regression for even values of m. Second, Greenland and Mansournia [19] argue against the Firth method from a Bayesian perspective, stating that the prior distribution should be data-independent.

The major contributions of this thesis can be summarized as follows:

1. Develop a standard pipeline for implementing the log-F-penalized logistic regression in GWAS of biobank data.
2. Extend the usage of the log-F method to conditional logistic regression that is commonly used in the matched case-control data from health and medical studies.
3. Advocate the log-F method as a more accessible alternative for researchers seeking to implement penalized logistic regression.

1.2 Organization of this Thesis

Chapter 2 introduces a two-step log-F-penalized logistic likelihood method for rare variant association studies. This chapter is a copy of the manuscript by Ying Yu, Siyuan Chen, Samantha J Jones, Rawnak Hoque, Olga Vishnyakova, Angela Brooks-Wilson, and Brad McNeney entitled Penalized Logistic Regression Analysis for Genetic Association Studies of Binary Phenotypes which was published in Human Heredity [56]. The method is to use a class of log-$F(m,m)$ priors for penalized logistic regression analysis with m viewed as a shrinkage parameter controlling the degree of penalization. The shrinkage parameter can be estimated using an Empirical Bayes approach. For a given genetic marker, the method assumes that the regression coefficient follows a log-F prior distribution, and a marginal likelihood for m is obtained by integrating the regression coefficient out of the joint distribution of the observed data and the regression coefficient. A composite likelihood is then constructed by taking a product of marginal likelihoods over multiple variants. Since the integrals cannot be solved analytically, the method relies on approximation techniques, such as Laplace approximation or the Monte Carlo EM algorithm, to the marginal likelihoods. The two-step procedure involves: estimating the shrinkage parameter m by maximizing the approximate composite likelihood, and implementing the log-F penalization approach with a simple data augmentation trick. The development of this method provides a standard pipeline for applying log-F-penalized logistic regression in GWAS.

Chapter 3 is an application study where we implement the log-F-penalized method proposed in Chapter 2 to the UK Biobank dataset. The UK Biobank dataset is a large-scale biobank containing genetic and health data on half a million participants [51], which makes it an ideal candidate for examining the feasibility of applying the log-F-penalized approach on a genome-wide scale. To better adapt the log-F-penalized method to real-world biobank data and improve its flexibility, we make three important updates: first, we incorporate polygenic effects estimated from a whole-genome regression into the logistic regression model in order to account for population structure and hidden relatedness; second, we introduce a frequency-specific parameter in log-F priors, allowing for different degrees of shrinkage for each variant; lastly, we explore the properties of shrinkage parameter selection based on phenotype characteristics, making the method better suited for various types of phenotypes. This application demonstrates the potential of the proposed method for identifying genetic associations within large-scale biobank data, while addressing some of the challenges faced in traditional GWAS approaches.

Chapter 4 presents an workflow to adapt log-F-penalization to conditional logistic regression. This chapter is a copy of the manuscript by Ying Yu, Jiying Wen, Jinko Graham and Brad McNeney entitled Log-F-penalized conditional logistic regression for sparse data. The standard conditional maximum likelihood estimator is known to be biased away from zero when the dataset is small or sparse. We develop point- and interval-estimation based on maximizing the conditional likelihood penalized by a log-F distribution. The estimators can be obtained from data augmentation and standard conditional logistic software. We compare the log-F-penalized approach to Firth's method in a simulation study. We also apply these methods to data from a study of the effects of maternal exposure to dietheylstibestrol on the risk of vaginal cancer in their daughters and a case-parent trio study of genetic risk factors for type 2 diabetes. This extension allows researchers to utilize the advantage of the log-F method, such as computational efficiency, in matched case-control studies.

Chapter 2

Penalized Logistic Regression Analysis for Genetic Association Studies of Binary Phenotypes

2.1 Introduction

Standard likelihood-based inference of the association between a binary trait and genetic markers is susceptible to sparse data bias [18] when the case-control ratio is unbalanced and/or the genetic variant is rare. In particular, when data are sparse, hypothesis tests based on asymptotic distributions have inflated type I error [60] and the MLE of OR is biased away from zero [10].

The relevance of sparse data bias to genetic association analysis is highlighted by recent work on methods for genome-wide, phenome-wide association studies (PheWAS) of large biobanks. Despite the potential of multivariate methods that jointly analyze phenotypes (e.g., [63]), approaches for PheWAS of biobank-scale data typically reduce the problem to inferences of association between single nucleotide variants (SNVs) and traits, adjusted for population structure and relatedness among subjects *via* a linear mixed model (LMM) [6, 60] or WGR [44]. For valid testing of associations between rare binary phenotypes and/or SNVs, SAIGE [60], EPACTS [33] and REGENIE [44] implement an efficient saddle-point approximation (SPA) to the distribution of the score statistic that yields correct p-values. EPACTS and REGENIE also offer testing and effect estimation based on Firth logistic regression [10, 24], a maximum-penalized likelihood method that uses the Jeffreys prior [28] as the penalty. In addition to valid tests, the Firth logistic regression estimator of the odds-ratio is first-order unbiased. Reliable effect estimates are important for designing replication studies and polygenic risk scores, and for fine-mapping [36].

Greenland and Mansournia proposed an alternative penalized-likelihood approach in which each log-OR parameter follows a class of log-$F(m, m)$ priors, where m serves as the shrinkage parameter controlling the degree of shrinkage applied to the coefficient [19]. This method assumes that each covariate of interest is independent and follows a log-F prior distribution. The log-F-penalized logistic likelihood can be obtained by multiplying the logistic regression likelihood by independent log-F prior densities (equation (1.1)). The explanatory variables of the logistic regression may include other covariates such as age, sex, genetic principal components (PCs) or the predicted log-odds of being a case from a WGR. In general, we only penalize the SNV of interest but do not penalize other confounding covariates or the intercept, as suggested by Greenland and Mansournia [19].

Limited simulation studies have shown that, for fixed m, log-$F(m, m)$ penalized methods outperform other approaches for case-control data [14]. 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, and greater shrinkage may reduce MSE [19]. 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 [19]. We follow the suggestion by Greenland and Mansourinia of using an empirical Bayes method to estimate m [19].

Our interest is in fitting single-SNV logistic regressions over a genomic region, or over the entire genome. A motivating example is the Super Seniors study [21] that compared healthy "case" subjects aged 85 and older across Canada who had never been diagnosed with cancer, dementia, diabetes, cardiovascular disease or major lung disease to population-based middle-aged "controls" who were not selected based on health status. The genetic data for this study are described in detail in Section 2.4. After quality control, we have data available for 2,678,703 autosomal SNVs on 427 controls and 617 cases [21]. A preliminary genome-wide scan at a relatively liberal significance threshold of 5×10^{-5} found 57 SNVs associated with case-control status.

As in the Super Seniors data, the vast majority of SNVs have little or no effect, and a relatively small set have non-zero effects. The prior used for penalization is the distribution of log-ORs for SNVs with non-zero effects. We therefore propose to select K SNVs that show some evidence of having non-zero effects in a preliminary scan, e.g., the $K = 57$ SNVs from the preliminary scan of the Super Seniors data, and use these to estimate m. The intent is to learn about the distribution of non-zero log-ORs adaptively from the data [62].

The main goal of this chapter is to employ log-F penalized logistic regression for analyzing genetic variant associations in a two-step approach. First, we estimate the shrinkage parameter m based on a set of variants that show evidence of having non-zero effect in a preliminary scan. 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 [19]. In addition to estimates of SNV effects, confidence intervals and likelihood ratio tests follow from the penalized likelihood [24]. Corrections for multiple testing in GWAS/PheWAS applications would involve standard GWAS p-value thresholds, such as 5×10^{-8}.

2.2 Models and Methods

We start by reviewing the penalized likelihood for cohort data, followed by the likelihood for case-control data. We then introduce the penalized likelihood and derive a marginal likelihood for the shrinkage parameter m based on data from a single SNV. Taking products of marginal likelihoods from K SNVs yields a composite likelihood that we maximize to estimate m. We conclude by reviewing how log-F-penalized logistic regression for the second-stage of the analysis can be implemented by data augmentation.

2.2.1 Likelihood from Cohort Data

Inference of associations between a single-nucleotide variant (SNV) and disease status from cohort data is based on the conditional distribution of the binary response Y_i given the covariate X_i that encodes the SNV. For a sample of n independent subjects let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ denote the vector of response variables and $\boldsymbol{X} = (X_1, \ldots, X_n)$ denote the vector of genetic covarariates. The likelihood is

$$L(\alpha, \beta) = P(\boldsymbol{Y}|\boldsymbol{X}, \alpha, \beta) = \prod_{i=1}^{n} \frac{\exp(Y_i(\alpha + X_i\beta))}{1 + \exp(\alpha + X_i\beta)}, \tag{2.1}$$

where α is an intercept term and β is the log-OR of interest.

2.2.2 Likelihood from Case-control Data

The association between a single-nucleotide variant (SNV) and disease status can also be estimated from case-control (i.e. retrospective) data, in which covariates X_i are sampled conditional on disease status Y_i for each individual i. Suppose there are n_0 controls indexed $i = 1, ..., n_0$ and n_1 cases indexed $i = n_0 + 1, ..., n$, where $n = n_0 + n_1$ denoting the sample size of the study. Qin and Zhang [46] expressed the case-control likelihood in terms of a two-sample semi-parametric model as follows

$$\begin{aligned} L(\beta, g) = P(\boldsymbol{X}|\boldsymbol{Y}, \beta, g) &= \prod_{i=1}^{n_0} P(X_i|Y_i = 0, g) \prod_{i=n_0+1}^{n_0+n_1} P(X_i|Y_i = 1, \beta, g) \\ &= \prod_{i=1}^{n_0} g(X_i) \prod_{i=n_0+1}^{n_0+n_1} c(\beta, g)\exp(X_i\beta) g(X_i), \end{aligned} \tag{2.2}$$

where $c(\beta, g)$ is a normalizing constant and $g(X)$ is the distribution of the covariates in controls, considered to be a nuisance parameter. The potentially infinite-dimensional distribution 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 [45]:

$$\begin{aligned} L(\alpha^*, \beta) &= \prod_{i=1}^{n_0} \frac{1}{1 + \exp(\alpha^* + X_i\beta)} \prod_{i=n_0+1}^{n_0+n_1} \frac{\exp(\alpha^* + X_i\beta)}{1 + \exp(\alpha^* + X_i\beta)} \\ &= \prod_{i=1}^{n} \frac{\exp(Y_i(\alpha^* + X_i\beta))}{1 + \exp(\alpha^* + X_i\beta)}, \end{aligned} \tag{2.3}$$

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 [48]. The profile likelihood $L(\alpha^*, \beta)$ for case-control data is of the same form as the 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 [45, 46]. In what follows we write the likelihood as in equation (2.3) with the understanding that $\alpha^* = \alpha$ for cohort data.

2.2.3 Penalized and Marginal Likelihoods

The penalized likelihood is obtained by multiplying the likelihood by a log-$F(m, m)$ distribution (equation (1.1)):

$$L_p(\alpha^*, \beta, m) = L(\alpha^*, \beta) f(\beta|m). \tag{2.4}$$

Integrating out the latent log-OR β gives a marginal likelihood of α and m:

$$L(\alpha^*, m) = \int L_p(\alpha^*, \beta, m) d\beta = \int L(\alpha^*, \beta) f(\beta|m) d\beta. \tag{2.5}$$

In the above likelihood, the smoothing parameter m is the parameter of interest, while the intercept α^* is a nuisance parameter.

We expect very little information about m in data from a single marker, because this represents a single realization of β from the log-$F(m, m)$ prior. In fact, empirical experiments (not shown) suggest a monotone, completely uninformative likelihood roughly 60-70 percent of the time. We therefore consider combining information across markers.

2.2.4 Composite Likelihood for Estimating m with K Markers

Suppose there are K SNVs available for estimating m (see subsection 2.2.4). For each SNV we specify a one-covariate logistic regression model. Let $\boldsymbol{X}$ denote a design matrix containing all K SNVs, and $\boldsymbol{X}_{\cdot k}$, $k = 1, \ldots, K$, denote the genotype data on the kth SNV. Let $L_p(\alpha_k^*, \beta_k)$ denote the likelihood (2.3) for the kth log-OR parameter β_k. Here α_k^* is the intercept term from the kth likelihood, considered to be a nuisance parameter.

A composite likelihood [26, 37, 53] for $\boldsymbol{\alpha}^{*=(\alpha_1^*,\dots,\alpha_K^*)^T}$ and m is the weighted product

$$L_{CL}(\boldsymbol{\alpha}^{*,m)=\prod_{k=1}^{K} L(\alpha_k^*,m)^{w_k}.} \tag{2.6}$$

The corresponding composite log-likelihood is

$$l_{CL}(\boldsymbol{\alpha}^{*,m)=\sum_{k=1}^{K} w_k l(\alpha_k^*,m),} \tag{2.7}$$

where $l(\alpha_k^*, m)$ is the marginal log-likelihood contribution from the kth variant obtained by integrating β_k out of the joint distribution of observed data and the parameter. Our estimate of m is the value that maximizes the composite log-likelihood equation (2.7). Following the notion that common variants should tend to have weaker effects and rare variants should tend to have stronger effects, we set $\sqrt{w_k} = 1/\sqrt{MAF_k(1 - MAF_k)}$ so that w_k is inversely proportional to the MAF of the kth SNV [55]. The idea is to up-weight rarer variants of potentially greater effects and down-weight more common SNVs that may have smaller effects.

Maximization is done in two stages:

1. For fixed m, we maximize $l_{CL}(\boldsymbol{\alpha}^{*,m)}$.The form of the composite likelihood when m is fixed, as a sum of terms involving only a single parameter, implies that to maximize $l_{CL}(\boldsymbol{\alpha}^{*,m)}$ we maximize each $l(\alpha_k^*, m)$ over α_k^*. Let $\alpha_k^*(m)$ be the value of α_k^* that maximizes $l(\alpha_k^*, m)$, $\hat{\boldsymbol{\alpha}}^{*(m)=(\hat{\alpha}_1^*(m),\dots,\hat{\alpha}_K^*(m))}$, and $l_{CL}(\hat{\boldsymbol{\alpha}}^{*(m),m)=\sum_{k=1}^{K} w_k l(\hat{\alpha}_k^*(m),m)}$.
2. Maximize $l_{CL}(\boldsymbol{\alpha}^*(\hat{m}), m)$ over m. To keep computations manageable, we restrict m to a grid of values, $m = 1, 2, ..., M$. One may optionally smooth the resulting $(m, l_{CL}(\hat{\boldsymbol{\alpha}}^{*(m),m))}$ pairs and maximize this smoothed curve to obtain the estimate $\hat{m}$. Smoothing is preferred if one seeks a more precise estimate of m, but the level of precision in m does not substantially impact the resulting log-OR based on our experiments.

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 (2.5). We discuss two approximate approaches. The first (Section 2.2.5) is a Monte Carlo EM algorithm [9], and the second (Section 2.2.5) is a Laplace approximation to $L(\alpha_k^*, m)$ followed by derivative-free optimization of the approximation.

Selecting variants for the composite likelihood

Using variants with no effect in the composite likelihood 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 (results not shown).

To avoid over-shrinkage we select SNVs with large marginal effects (i.e., small p-values) from a genome-wide scan, similar to the SNV-selection process used by FaST-LMM-Select [40]. For example, we can conduct a preliminary GWAS on all markers, or a thinned set of markers, and choose the SNVs with p-values below a multiple-testing-corrected threshold (refer this as Level 0 of Step 1). We then use the chosen SNVs to estimate m (Level 1 of Step 1).

Adjustment for confounding variables and offsets

We conclude this subsection by noting that it is possible to generalize the marginal likelihood approach for estimating m to incorporate non-genetic confounding variables, denoted Z, and known constants in the linear predictor, or "offset" terms, denoted b. 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. Offset terms can be used to include estimated polygenic effects in the logistic regression [44]. Expanding the α_k^* component of the logistic model to $\alpha_k^* + Z\gamma_k + b$, the kth likelihood is now

$$L(\alpha_k^*, \gamma_k, \beta_k) = \prod_{i=1}^{n} \frac{\exp(Y_i(\alpha_k^* + Z_i\gamma_k + b_i + X_{ik}\beta_k))}{1 + \exp(\alpha_k^* + Z_i\gamma_k + b_i + X_{ik}\beta_k)} \tag{2.8}$$

and the composite log-likelihood for estimating m is

$$\begin{aligned} l_{CL}(\boldsymbol{\alpha}^{*,\gamma,m}) &= \sum_{k=1}^{K} w_k l(\alpha_k^*, \gamma_k, m) \\ &= \sum_{k=1}^{K} w_k \log \int L(\alpha_k^*, \gamma_k, \beta_k) f(\beta_k|m) d\beta_k. \end{aligned} \tag{2.9}$$

For fixed m we maximize $l_{CL}(\boldsymbol{\alpha}^{*,\gamma,m})$ 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 the generalization to include confounding variables and offsets is conceptually straightforward, we omit it in what follows to keep the notation as simple as possible.

2.2.5 Maximization Approaches

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 iterates between taking the conditional expectation of the complete-data log-likelihood given the observed data and the current parameter estimates, and maximizing this conditional expectation. The conditional distribution of β_k given the observed data is a posterior distribution that is proportional to the likelihood $L(\alpha^*, \beta_k)$ times the prior $f(\beta_k|m)$. Thus, at the $(p+1)^{th}$ iteration, the E-step is to determine

$$Q\left(\alpha_k^*|\alpha_k^{*(p)}, m\right) \propto \int \log[L(\alpha_k^*, \beta_k)f(\beta_k|m)]L(\alpha_k^{*(p)}, \beta_k)f(\beta_k|m)d\beta_k \tag{2.10}$$

and the M-step is to set

$$\alpha_k^{*(p+1)} = \underset{\alpha_k^*}{\operatorname{argmax}}\, jQ\left(\alpha_k^*|\alpha_k^{*(p)}, m\right). \tag{2.11}$$

The E-step (2.10) 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 [54]. The MC integration in the E-step is obtained by sampling from the prior distribution $f(\beta_k|m)$ [54, 38]. Based on a sample $\beta_{k1}, ..., \beta_{kN}$ from the distribution $f(\beta_k|m)$, the MC approximation to the integral is

$$\begin{aligned} Q\left(\alpha_k^*|\alpha_k^{*(p)}, m\right) &\approx Q_{MC}\left(\alpha_k^*|\alpha_k^{*(p)}, m\right) \\ &= \frac{1}{N}\sum_{j=1}^{N} \log[L(\alpha_k^*, \beta_{kj})f(\beta_{kj}|m)]L(\alpha_k^{*(p)}, \beta_{kj}) \\ &= \frac{1}{N}\sum_{j=1}^{N} (\log[L(\alpha_k^*, \beta_{kj})] + \log[f(\beta_{kj}|m)])L(\alpha_k^{*(p)}, \beta_{kj}). \end{aligned} \tag{2.12}$$

Note that $\log[f(\beta_{kj}|m)]$ is independent of the parameter α_k^*, so maximizing (2.12) in the M-step is equivalent to maximizing

$$\frac{1}{N}\sum_{j=1}^{N} \log[L(\alpha_k^*, \beta_{kj})]L(\alpha_k^{*(p)}, \beta_{kj}). \tag{2.13}$$

For a discussion of computational approaches to the M-step see the online Supplementary Material.

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 [52]. 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

$$\tilde{L}(\alpha_k^*, m) = L(\alpha_k^*, \hat{\beta}_k) f(\hat{\beta}_k|m) \sqrt{\frac{2\pi}{c_p(\alpha_k^*)}}. \tag{2.14}$$

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 [7]. 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.

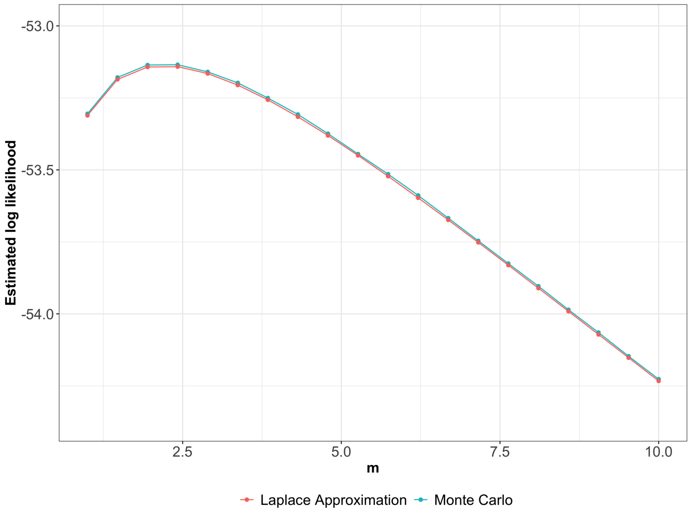


Figure 2.1: Natural logarithms of estimates of the marginal likelihood $L(\alpha_k^*, m)$ for one simulated dataset generated under $m = 4$. Estimates are obtained by LA and Monte Carlo. Log-likelihood estimates are plotted over the grid $m = (1, 1.5, ..., 10)$ with $\alpha_k^* = -3$.

2.2.6 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 [19]. Here, we provide some details. The logistic likelihood penalized by a log-$F(m, m)$ prior (equation (2.4)) is:

$$\begin{aligned} L_P(\alpha^*, \beta) &= \prod_{i=1}^{n} \frac{\exp(Y_i(\alpha^* + X_i\beta))}{1 + \exp(\alpha^* + X_i\beta)} \times \frac{\exp(\frac{m}{2}\beta)}{(1 + \exp(\beta))^m} \\ &= \prod_{i=1}^{n} \frac{\exp(Y_i(\alpha^* + X_i\beta))}{1 + \exp(\alpha^* + X_i\beta)} \times \left[\frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}\right]^{\frac{m}{2}} \left[\frac{1}{1 + \exp(X_i\beta)}\right]^{\frac{m}{2}}, \end{aligned} \tag{2.15}$$

$$\begin{array}{lcc|cccccc}
 & \multicolumn{2}{c|}{\overbrace{\hphantom{\textbf{Success}\quad\textbf{Failure}}}^{\textbf{Response}}} & & & & & & \\
 & \textbf{Success} & \textbf{Failure} & \textbf{Intercept} & \mathbf{X} & \boldsymbol{Z_1} & \boldsymbol{Z_2} & \cdots & \boldsymbol{Z_p} \\
\hline
\text{Original Dataset} & 1 & 0 & 1 & 0 & x_{11} & x_{21} & \cdots & x_{p1} \\
 & 0 & 1 & 1 & 2 & x_{12} & x_{22} & \cdots & x_{p2} \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 & 0 & 1 & 1 & 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \\
\hline
\text{Augmented Dataset} & m/2 & m/2 & 0 & 1 & 0 & \cdots & \cdots & 0
\end{array}$$

Figure 2.2: Illustration of data augmentation in the implementation of log-$F(m,m)$ penalization.

where $X_i = 1$. Thus, the penalized likelihood $L_p(\alpha^*, \beta)$ is equivalent to an unpenalized likelihood obtained by adding m pseudo-observations to the response with no intercept and covariate one, in which $m/2$ are successes and $m/2$ are failures (even if m is an odd number). In our analyses (see Section 2.3), 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 of all zeros except for a one indicating the SNV covariate. Analyzing the augmented dataset with standard logistic regression yields the penalized MLE and its standard errors, as well as penalized likelihood ratio tests and penalized-likelihood-ratio-based confidence intervals. We conclude by noting that, for fixed m, the influence of the m pseudo-observations on the fitted logistic regression diminishes as the sample size increases. In other words, for any m, the extent of penalization decreases with sample size.

2.3 A Simulation Study

The empirical performance of the methods introduced in Section 2.2 was evaluated in a simulation study. The proposed two-step log-F-penalized method (LogF) was compared with the standard MLE and the following methods:

- Firth logistic regression (FLR) was first proposed by Firth [10], 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` [22].

- Penalization by Cauchy priors (CP) was proposed by [13]. 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 `R` function `bayesglm` of the package `arm` [13].

All simulations were performed using `R` (Version 4.1.2) [47] on the Compute Canada cluster Cedar. We restricted m to a grid of values between 1 and 10, and we used parallel processing that splits the computation of the composite likelihood for each $m \in [1, 10]$ over different cores. Each node on the cluster has at least 32 CPU cores and we allocated 10G to each core. For detailed description of its nodes' characteristics please refer to `https://docs.computecanada.ca/wiki/Cedar#Node_characteristics`.

We set the sample size to 500, 1000 and 1500, and 100 data sets were generated in each scenario. For each data set, we first estimated m based on a set of SNVs which show non-zero effects in a preliminary scan (Step 1), and then implemented the log-F penalized likelihood method to test single-variant association for each SNV by the data augmentation approach (Step 2). For the MLE and CP approaches we used Wald tests for SNV effects and Wald confidence intervals for the SNV coefficient. For FLR and the LogF approaches tests we used likelihood-ratio tests (LRTs). For a penalized log likelihood $l_P(\alpha, \beta)$, the likelihood ratio statistic [24] is

$$T = 2[l_P(\hat{\alpha}, \hat{\beta}) - l_P(\hat{\alpha}_0, 0)] \tag{2.16}$$

where $(\hat{\alpha}, \hat{\beta})$ is the maximum of the penalized likelihood function and $\hat{\alpha}_0$ is the maximum of the penalized likelihood when $\beta = 0$. The p-value is computed from the χ^2_1 distribution. For penalized logistic method, profile penalized likelihood (PPL) confidence intervals have shown to have better empirical properties than standard Wald-based confidence intervals [24]. A PPL confidence interval can be obtained by inverting the LRT, i.e., by finding all values of $\hat{\beta}_0$ such that $2[l_P(\hat{\alpha}, \hat{\beta}) - l_P(\hat{\alpha}_0, \hat{\beta}_0)] \leq \chi^2_{1,1-\alpha}$ gives a $100(1-\alpha)\%$ confidence interval for β.

2.3.1 Data Generation

To keep computations manageable, we simulate preliminary datasets of 50 causal and 950 null SNVs. Null SNVs were used to assess the Type I error performance and the power was estimated using the set of causal SNVs. The data was simulated according to a case-control sampling design, where covariates are simulated based on disease status. For a given SNV, let X_j denote the allele count (i.e. 0, 1 or 2) of SNV$_j$ and β_j be the corresponding log-OR parameter. Following [46], the conditional density function for X_j in the controls and cases are

$$\begin{aligned} &P(X_j = x|Y = 0) = g(x) \;\text{ and} \\ &P(X_j = x|Y = 1) = h(x) = c(\beta_j, g)\exp(x\beta_j)g(x). \end{aligned} \tag{2.17}$$

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

$$g_z(x)\exp(x\beta_j) = \begin{cases} (1-p)^2 & x = 0 \\ 2p(1-p)\exp(\beta_j) & x = 1 \\ p^2\exp(2\beta_j) & x = 2 \end{cases}, \tag{2.18}$$

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

We 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 $P(Z = z|Y = 1) \propto f_z\exp(z\gamma)$ [59]. 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_j in controls of population z, i.e., $P(X_j = x|Z = z, Y = 0) = g_z(x) \sim$ Binomial$(2, p_z)$. The joint distribution of X_j and Z in controls is then $P(X_j = x, Z = z|Y = 0) = f_z g_z(x)$. If $\text{logit}[P(Y = 1|Z = z, X_j = x)] = \alpha + z\gamma + x\beta_j$, the joint distribution of X and Z in cases is $P(X_j = x, Z = z|Y = 1) \propto f_z g_z(x)\exp(z\gamma + x\beta_j)$ [46]. We then have

$$P(X_j = x|Z = z, Y = 1) = \frac{P(X_j = x, Z = z|Y = 1)}{P(Z = z|Y = 1)} \propto \frac{f_z g_z(x)\exp(z\gamma + x\beta_j)}{f_z\exp(z\gamma)} = g_z(x)\exp(x\beta_j). \tag{2.19}$$

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

$$\begin{aligned} &P(Z = z|Y = 0) = f_z, \\ &P(Z = z|Y = 1) \propto f_z\exp(z\gamma) = \begin{cases} f_0 & z = 0 \\ f_1\exp(\gamma) & z = 1 \end{cases} \end{aligned} \tag{2.20}$$

Then using (2.18), we simulated the genotype data of each SNV$_j$, for $j = 1, ..., 1000$, conditional on population status by sampling from

$$P(X_j = x|Z = z, Y = 0) = g_z(x) \sim \text{Binomial}(2, p_z),$$

$$P(X_j = x|Z = z, Y = 1) \propto g_z(x)\exp(x\beta_j) = \begin{cases} (1 - p_z)^2 & x = 0 \\ 2p_z(1 - p_z)\exp(\beta_j) & x = 1 \\ p_z^2 \exp(2\beta_j) & x = 2 \end{cases} \tag{2.21}$$

MAFs, p_z, for different populations were obtained from 1000 Genomes Project [8]. Here we consider two populations: Caucasian (CEU) and Yoruba (YRI) subjects, and we sampled MAFs of SNVs from a 1 million base-pair region on Chromosome 6 (SNVs with MAF = 0 have been removed). Data from the 1000 Genomes Project was downloaded using the Data Slicer (`https://www.internationalgenome.org/data-slicer/`). The effect sizes of causal SNVs were assumed to be a decreasing function of MAF, which allows rare SNVs to have larger effect sizes and common SNVs to have smaller effect sizes. We set the magnitude of each $\beta_j = \frac{\log 5}{2}|\log_{10}\text{MAF}_j|$ [55], where MAF$_j$ is the pooled-MAFs $(p_0 + p_1)/2$ of the SNV$_j$. We took into account the effects of mixed signs, multiplying β_j by -1 for some j, in which are 50% positive and 50% negative. This process gives the maximum OR = 6.44 ($|\beta_j| = 1.86$) for SNVs with pooled-MAF = 0.0048 and the minimum OR = 1.40 ($|\beta_j| = 0.4$) for SNVs with pooled-MAF = 0.38 (Supplementary Figure 2 B).

2.3.2 Results

We first evaluated the performance of the two different log-F methods described above. Over 100 simulation replicates, the mean estimates of m obtained by MCEM and LA are 4.77 (SD = 1.27) and 4.76 (SD = 1.18) respectively for $n = 500$, and are 3.88 (SD = 1.56) and 3.83 (SD = 1.33) respectively for $n = 1000$. The scatterplots (Figure 2.3) show good agreement between the two methods. Figure 2.4 compares the LA- and MCEM-based likelihood curves of m for the first 20 simulated data sets. These likelihoods were plotted with m of grid values from 1 to 10 on the x-axis, and each was smoothed by a smoothing spline. The likelihood curves are of similar shape, though shifted because the MCEM approach estimates the likelihood up to a constant (compare equations (2.12) and (2.13)). The compute time of LogF and FLR is given in Table 2.1. We see that LA is 160× and 300× faster than MCEM in elasped time for Step 1 when analyzing 1000 SNVs of sample size 500 and 1000, respectively. Although MCEM is computationally more expensive than LA, the accuracy of its approximation can be controlled by the number of Monte Carlo replicates, whereas the accuracy of LA cannot be controlled. We used $N = 1000$ Monte Carlo replicates in the MCEM throughout, which gives reasonably good accuracy. The agreement of the MCEM and LA approaches for smaller sample sizes validates the accuracy of LA. MCEM results are not available for the largest sample size of $n = 1500$, because our current implementation fails due to numerical underflow. As expected, once m is selected, LogF is computationally efficient as only a simple data augmentation approach is used in Step 2. Combining Step 1 (with LA) and Step 2, along with the preliminary scan, which is of the same order of computation time as Step 2, the combined computation time of the LogF approach is roughly half that of FLR.

We further examined the accuracy of effect sizes from LogF-MCEM, LogF-LA, FLR and CP. All variants were binned based on the pooled-MAF in five bins: $(0\%, 1\%)$, $[1\%, 5\%)$, $[5\%, 10\%)$, $[10\%, 25\%)$ and $[25\%, 50\%]$, and there were 51, 128, 213, 401, and 207 SNVs in each bin. The causal variants can be either deleterious or protective (i.e. β_j is either positive or negative), so we define the bias of effect size estimates as the signed bias, $E[\text{sign}(\beta_j)(\hat{\beta}_j - \beta_j)]$; positive values indicate bias away from zero, while negative values indicate bias towards zero. We also evaluated the SD of effect size estimates as the standard deviation of $\hat{\beta}_j$ across 100 simulation replicates, and the mean squared error (MSE) as the sum of squared bias and squared SD. MAF-binned results are shown in Table 2.2 and Figure 2.5. In the Figure, results for the MLE obscure those for the other methods and are not shown. We find that for variants of MAF 1% or greater, all methods are comparable. However, for rare variants of MAF < 1%, the SD of LogF is much smaller than other methods. In addition, the signed bias of the LogF is more concentrated around zero compared with other methods, though this tendency about zero is counteracted by some extreme negative signed biases that suggest over-shrinkage in some cases. The MAFs of the three SNVs that lead to these extreme negative signed biases (Figure 2.5) are 0.0048, 0.0072, and 0.0074, respectively. We note that 0.0048 was the smallest MAF in our simulated datasets. Combining bias and SD results in a much smaller MSE for the LogF than other methods. Comparing the results under samples sizes of 500, 1000, and 1500, one can see that penalization makes less of an impact as the sample size increases.

Through simulations, we also investigated the Type 1 error and power of the test of SNV effects from the different approaches (Figure 2.6). Although all the methods provide good control of Type 1 error, we found that LogF approaches result in a relatively smaller false positive rate. All the methods had similar power, with slightly less power from LogF approaches. We believe that the increased power of the FLR and Cauchy approaches can be partly attributed to their bias away from zero for rare variants.

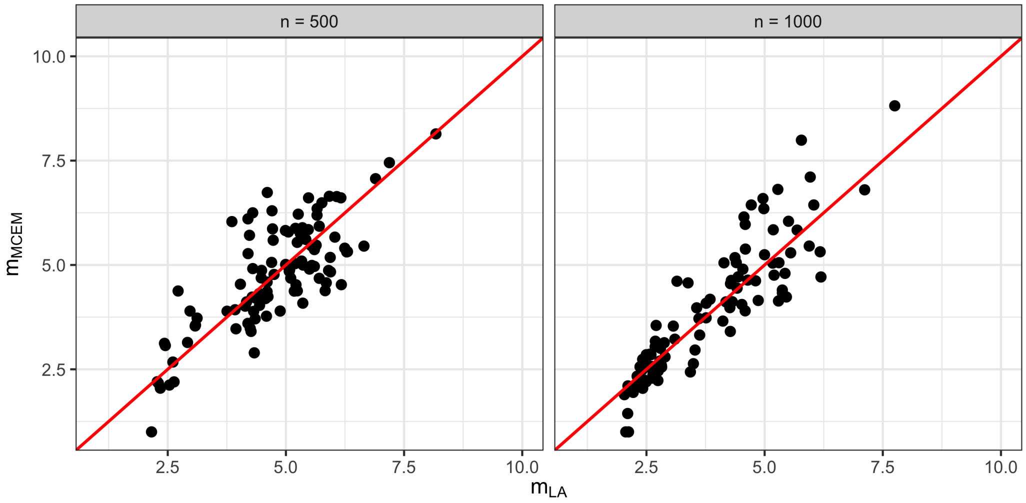


Figure 2.3: Scatterplot comparing the estimated values of m using the two methods over 100 simulation replicates. Values estimated by LA are on x-axis, and values estimated by MCEM are on y-axis. Red line is $y = x$.

Method	**Step**	**Elapsed time (s)**					
		Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
				$\boldsymbol{n = 500}$			
LogF	**1**	0.35	0.49	0.54	0.74	0.58	13.22
LogF	**2**	3.53	3.66	3.69	3.78	3.92	4.31
FLR	**NA**	10.48	10.94	11.09	11.18	11.36	12.36
				$\boldsymbol{n = 1000}$			
LogF	**1**	1.02	1.51	1.68	1.79	1.82	3.73
LogF	**2**	4.38	4.46	4.55	4.59	4.66	5.17
FLR	**NA**	17.04	17.60	17.79	17.85	18.00	19.48
				$\boldsymbol{n = 1500}$			
LogF	**1**	2.42	2.66	2.75	2.78	2.89	3.73
LogF	**2**	5.21	5.30	5.37	5.39	5.39	5.94
FLR	**NA**	23.45	23.96	24.35	24.33	24.61	25.35

Table 2.1: Computation time (elapsed time in seconds) of LogF and FLR when analyzing 1000 SNVs with sample size 500, 1000 and 1500 using 100 simulated data sets. No results are available for LogF-MCEM when $n = 1500$ due to numerical underflow.

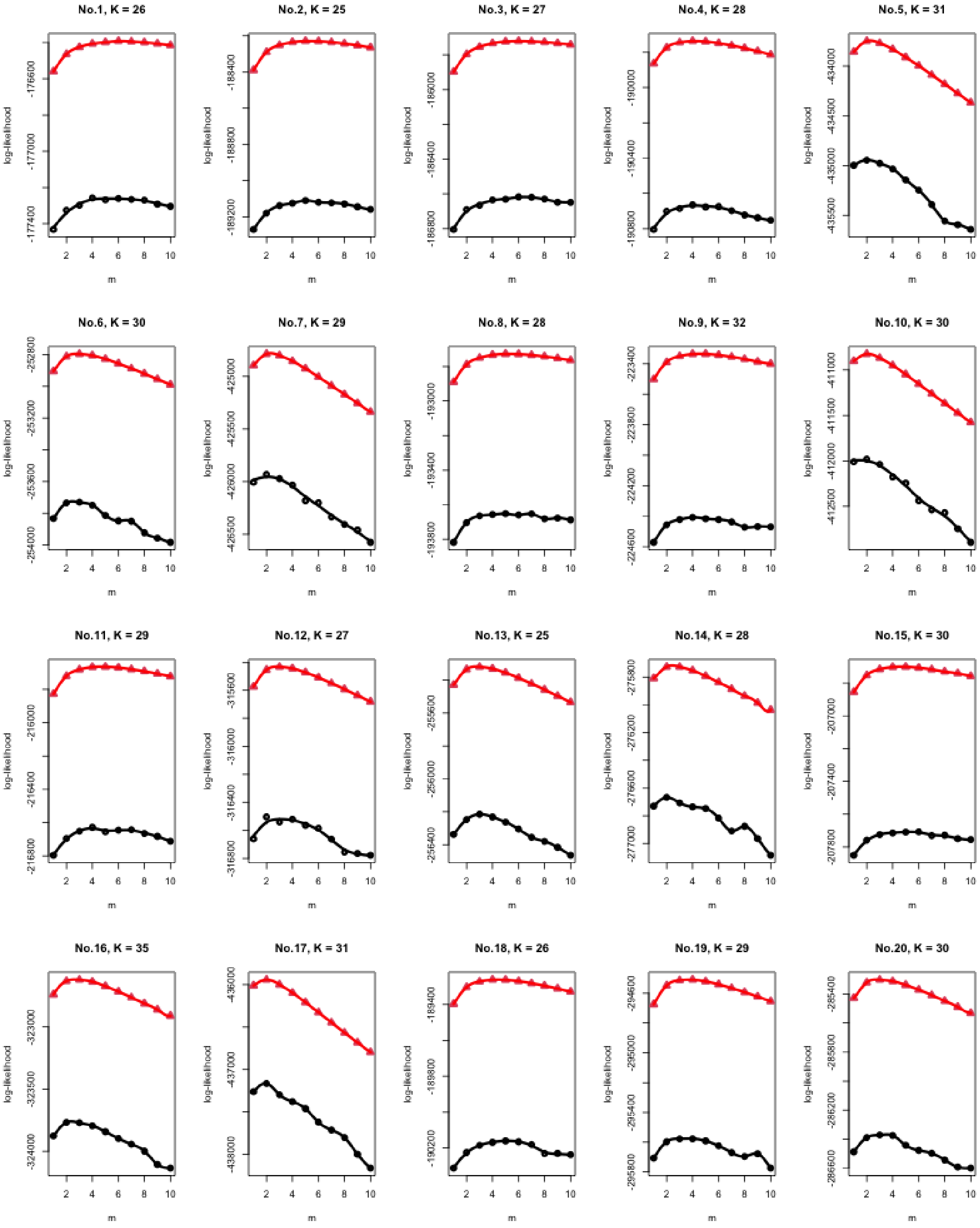


Figure 2.4: Comparison of profile log-likelihood curves obtained by the two different methods described in text, for the first 20 simulated data sets ($n = 1000$). In each case the likelihood curve was generated based on K SNVs selected in a preliminary genome-wide scan, and was smoothed by smoothing spline. The red line connecting triangles is based on LA, whereas the black line connecting dots corresponds to MCEM.

Estimate	**Method**	**MAF**				
		(0%, 1%)	**[1%, 5%)**	**[5%, 10%)**	**[10%, 25%)**	**[25%, 50%]**
		n = **500**				
Bias (×1000)	**MLE**	69	-4	3	1	0
	CP	-8	-4	3	0	-0
	FLR	-6	-4	3	1	0
	LogF-MCEM	-41	-10	0	-0	-0
	LogF-LA	-40	-10	0	-0	-0
SD (×1000)	**MLE**	4832	487	267	187	147
	CP	909	405	263	184	145
	FLR	881	404	261	184	145
	LogF-MCEM	476	339	244	179	143
	LogF-LA	473	339	244	179	143
MSE (×1000)	**MLE**	27275	376	74	36	22
	CP	852	178	71	35	21
	FLR	799	177	71	35	22
	LogF-MCEM	259	122	62	33	21
	LogF-LA	256	122	62	33	21
Coverage* (×1000)	**MLE**	992	957	952	950	950
	CP	990	960	953	951	951
	FLR	967	951	950	950	950
	LogF-MCEM	983	965	958	952	952
	LogF-LA	984	965	958	952	952
		n = **1000**				
Bias (×1000)	**MLE**	51	4	-0	-0	-1
	CP	7	3	-1	-0	-1
	FLR	8	3	-1	-0	-1
	LogF-MCEM	-14	0	-2	-1	-1
	LogF-LA	-15	0	-2	-1	-1
SD (×1000)	**MLE**	1542	290	187	130	104
	CP	663	283	185	129	103
	FLR	661	282	185	129	103
	LogF-MCEM	491	263	180	128	103
	LogF-LA	489	263	180	128	103
MSE (×1000)	**MLE**	3590	92	36	17	11
	CP	462	87	36	17	11
	FLR	456	87	35	17	11
	LogF-MCEM	257	74	34	17	11
	LogF-LA	254	74	34	17	11
Coverage* (×1000)	**MLE**	976	953	951	951	946
	CP	974	954	951	952	946
	FLR	955	951	950	951	946
	LogF-MCEM	974	956	952	952	946
	LogF-LA	975	956	952	952	946
		n = **1500**				
Bias (×1000)	**MLE**	27	2	1	-0	-1
	CP	0	2	1	-0	-1
	FLR	-0	2	0	-0	-1
	LogF-LA	-11	0	-0	-0	-1
SD (×1000)	**MLE**	782	236	151	106	83
	CP	518	233	151	106	82
	FLR	522	232	150	106	83
	LogF-LA	447	225	149	105	82
MSE (×1000)	**MLE**	1060	61	24	12	7
	CP	279	59	23	12	7
	FLR	283	59	23	12	7
	LogF-LA	207	54	23	11	7
Coverage* (×1000)	**MLE**	968	949	950	950	952
	CP	971	951	951	950	953
	FLR	959	948	950	950	952
	LogF-LA	970	951	951	950	952

Table 2.2: MAF binned averages of bias, SD, MSE and CI coverage probability of effect size estimates across 100 simulated data. * Coverage probability of two-sided nominal 95% confidence intervals for log-OR coefficient. Wald CIs were used for MLE and CP, whereas profile likelihood-based CIs were used for FLR, LogF-MCEM and LogF-LA. No results are available for LogF-MCEM when $n = 1500$ due to numerical underflow.

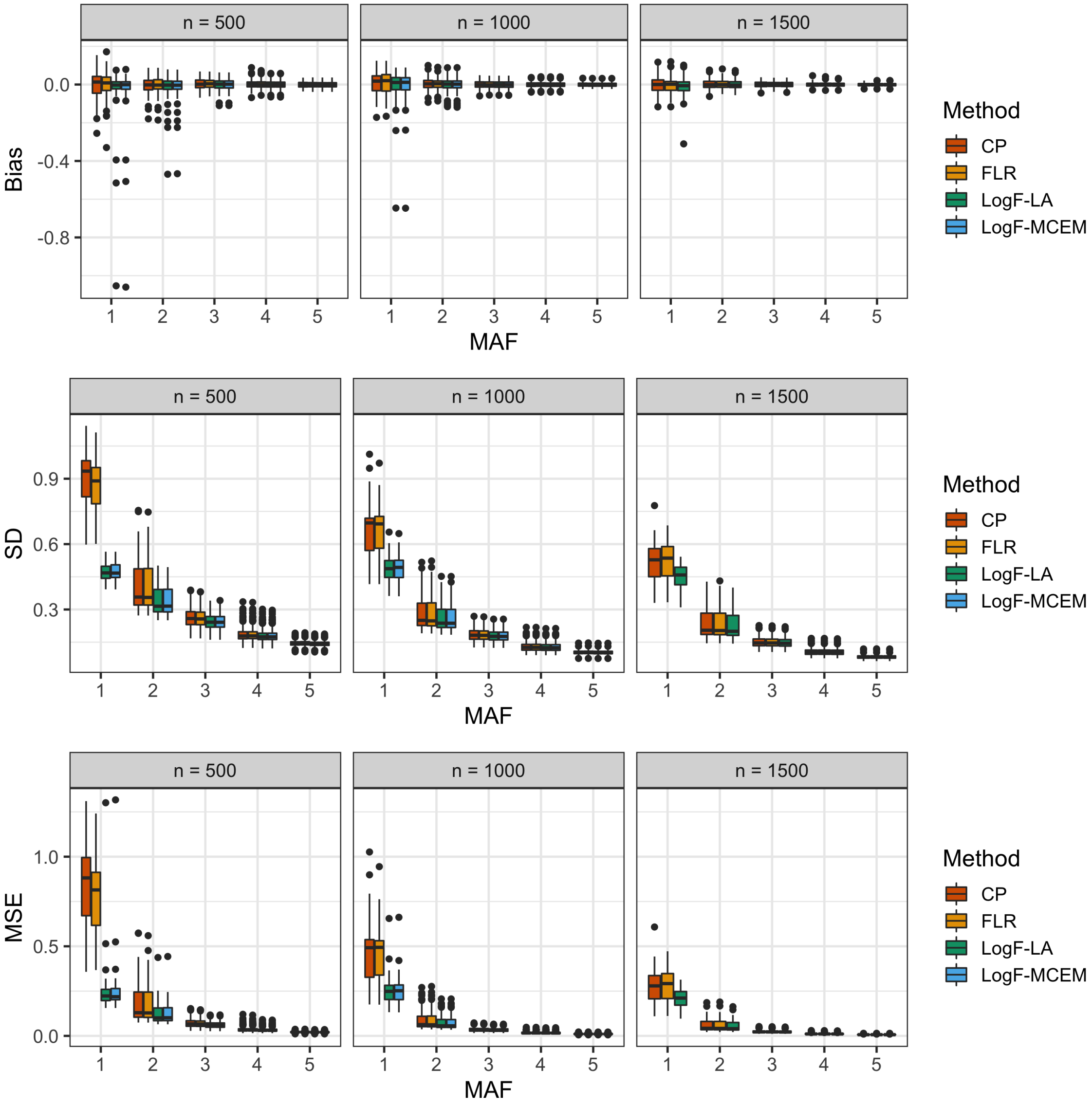


Figure 2.5: MAF binned boxplots of bias, SD and MSE of effect size estimates for LogF and other competing methods on simulated data. Each boxplot represents the distribution of the estimated quantity across 100 simulation replicates. MAF bins are: $1 = (0\%, 1\%)$, $2 = [1\%, 5\%)$, $3 = [5\%, 10\%)$, $4 = [10\%, 25\%)$ and $5 = [25\%, 50\%]$. No results are available for LogF-MCEM when $n = 1500$ due to numerical underflow.

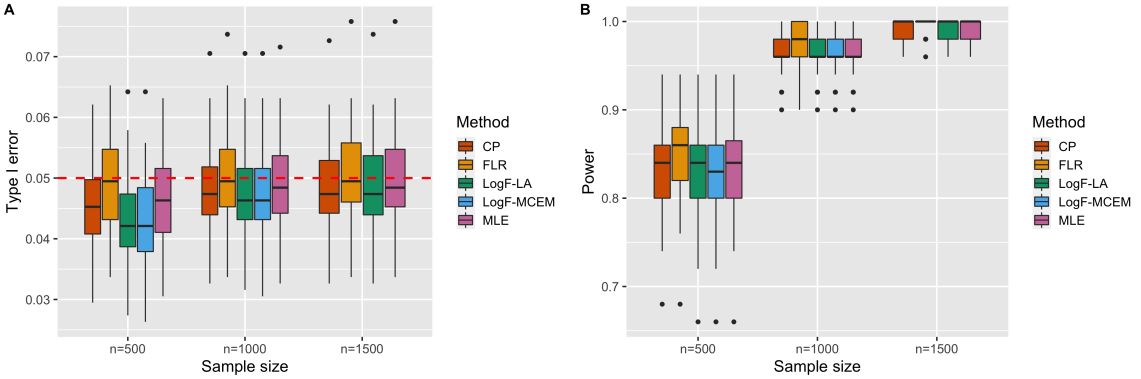


Figure 2.6: Type 1 error and power performance over simulated data sets. (A). Each boxplot represents the distribution of empirical type 1 error rates at nominal level 0.05 (red dashed horizontal line) across 100 simulation replicates computed at null SNVs. (B). Power computed at causal SNVs. No results are available for LogF-MCEM when $n = 1500$ due to numerical underflow.

2.4 Data Application

The Super Seniors data from the Brooks-Wilson laboratory was collected to investigate the association between genetic heritability and healthy aging of humans. The 'super seniors' are defined as those who are 85 or older and have no history of being diagnosed with the following 5 types of diseases: cardiovascular disease, cancer, diabetes, major pulmonary disease or dementia. In this study, 1162 samples of 4,559,465 markers were genotyped using a custom Infinium Omni5Exome-4 v1.3 BeadChip (Illumina, San Diego, California, USA) at the McGill University/Genome Quebec Innovation Centre (Montreal, Quebec, Canada) [31]. The data underwent extensive quality control after genotyping, including re-clustering, removal of replicate and tri-allelic single nucleotide polymorphisms (SNPs), and checking for sex discrepancies and relatedness. We also removed SNPs with MAF < 0.005, call rate $< 97\%$, or Hardy-Weinberg equilibrium p-value $< 1 \times 10^{-6}$ among controls. After a series of filtering steps, our final study includes 1044 self-reported Europeans, of which 427 are controls and 617 are cases (super seniors), and 2,678,703 autosomal SNPs.

A preliminary genome-wide scan identified 98 SNPs with p-values $< 5 \times 10^{-5}$. Of these, the 57 SNPs with no missing values were used to estimate the value of m. Our marginal likelihood approach for estimating m incorporates sex and the first 10 principal components as confounding variables. The m estimated by MCEM and LA are 7.01 and 6.89, respectively. To analyse 2,678,703 SNPs, the LogF approach (Step 2) takes 14 hours, which is 30× faster than FLR (437 hours). Manhattan plots (Figure 2.7) show very good agreement for the association detected between methods. Figure 2.8 shows the QQ-plot of p-values when applying MLE, LogF-LA (results of LogF-MCEM are close to LogF-LA, and are shown in

Supplementary Figure 5-7) and FLR to the Super Seniors data. All methods are close to the dashed line of slope one, though the FLR p-values veer up slightly above the line at $-\log_{10}$ p-values near 5. Figure 2.9 compares the parameter estimates of the MLE, FLR and LogF. Other than cases where the MLE appears grossly inflated (e.g., $|\hat{\beta}| > 5$), the estimates from the MLE and FLR are in surprisingly good agreement. The LogF estimates are shrunken more towards zero than those of FLR, and that the shrinkage is more pronounced for rare variants than for variants of frequency greater than 0.01. Figure 2.10 and Table 2.3 compare the p-values of the different approaches. The points below the dashed line of slope one in both panels of Figure 2.10 indicate that the FLR p-values are systematically lower than those of the MLE and LogF. This is also reflected in the confusion matrices of Table 2.3, which show that FLR flags more SNVs as significant at the 5×10^{-5} level than the other two methods. Taken together, these results suggest that the LogF approach may impose too much shrinkage on the SNV effect estimates.

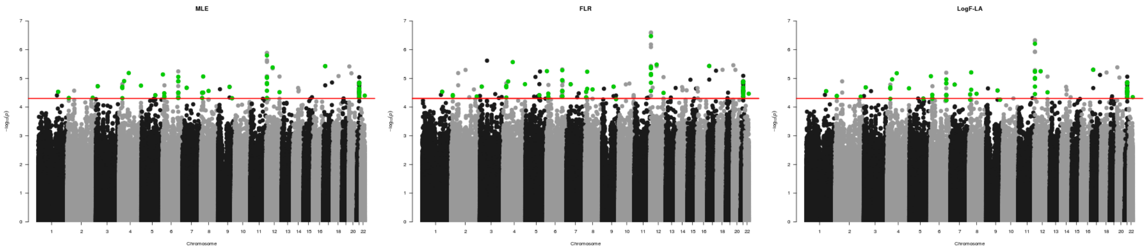


Figure 2.7: Manhattan plots comparing association results from different methods on Super Seniors data. The red horizontal line represents the liberal genome-wide significance threshold ($P = 5 \times 10^{-5}$) used to select SNPs in the preliminary scan. For LogF-LA, 57 SNPs (green points) below the threshold are used to estimate m in Step 1.

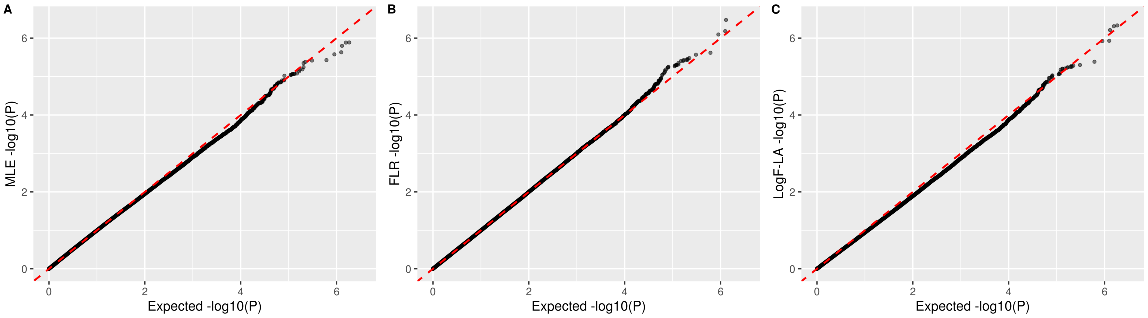


Figure 2.8: QQ-plot comparing p-values from different methods on Super Seniors data. The p-value for FLR and LogF-LA was obtained using the likelihood ratio test with a χ^2_1 test statistic.

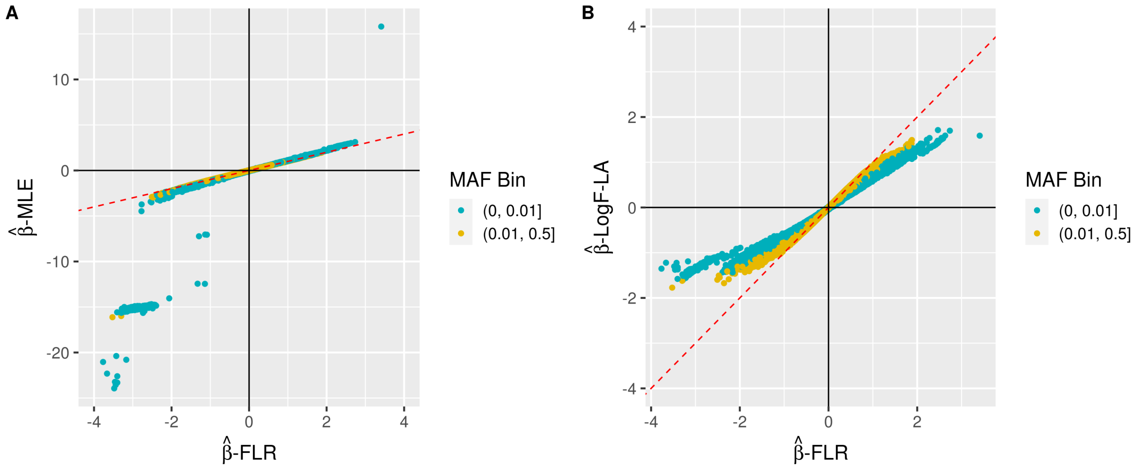


Figure 2.9: Scatterplots comparing effect size estimates from different methods for Super Seniors data. The plotting colors represents variant categories based on minor allele frequency (MAF) threshold of 1%.

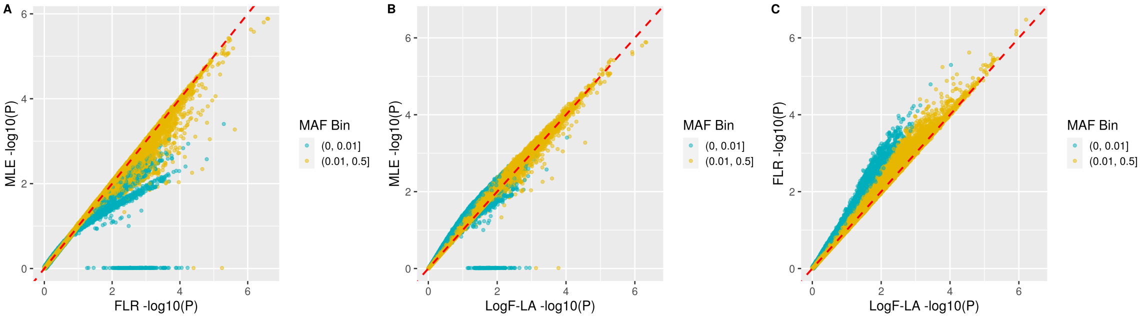


Figure 2.10: Scatterplots comparing p-values from different methods on Super Seniors data. The plotting colors represents variant categories based on minor allele frequency (MAF) threshold of 1%. For FLR and LogF-LA, the p-value for each variant was obtained by the likelihood ratio test with a χ_1^2 test statistic.

		MLE		**LogF-LA**		**LogF-MCEM**	
		0	1	0	1	0	1
FLR	0	2678557	1	2678558	0	2678558	0
	1	42	97	45	94	46	93

Table 2.3: Confusion matrices comparing association results from different methods on Super Seniors data, where ’1’ indicating the number of SNPs below the genome-wide significant threshold of 5×10^{-5} and ’0’ otherwise.

2.5 Discussion and Conclusion

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, we select K markers that show evidence of association with the phenotype in a preliminary scan and use these to estimate m. The value of m is the maximizer of a composite of K marginal likelihoods obtained by integrating the random effect out of the joint distribution of the observed data and the random effect. 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 two methods give similar results, with LA being faster for all sample sizes and more numerically stable for large sample sizes. Second, log-F penalties are conveniently implemented with standard logistic regression by translating the coefficient penalty into a pseudo-data record [19]. Our method requires extra computation time up front for the preliminary scan and selection of the shrinkage parameter m, but once selected, LogF approach (using LA in Step 1) is faster than Firth logistic regression (Table 2.1). Our simulation studies suggest that the proposed LogF approach has slightly lower bias and substantially lower MSE than the other methods considered for variants of frequency less than 5%, and similar bias and MSE for variants of frequency greater than 5%. However, the power results of our simulation study and the analysis of the Super Seniors data suggest that our current implementation of log-F penalization has a tendency to over-shrink estimates of truly-associated SNVs. We discuss generalizations of the penalization approach that might correct such over-shrinkage in what follows.

Penalization can be generalized by allowing the prior distribution to depend on characteristics of the SNV, such as MAF or annotation information. A straightforward extension is to stratify selection of the shrinkage parameter by, e.g., MAF. 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 and a larger variance for rare variants. In the context of heritability estimation [50] argue against stratified approaches and instead recommend modeling the variance of the SNV effects as proportional to $[f_i(1-f_i)]^{1+\alpha}$ for MAF f_i and a power α. Their analyses of real data suggested the value $\alpha = -0.25$. This corresponds to standardizing each SNV covariate by dividing by $[f_i(1-f_i)]^{(1+\alpha)/2}$ before analyses. In the context of modelling quantitative traits [58], proposing a double-exponential prior on SNV effects and a log-linear model for the scale parameter of the double exponential distribution allows the scale to depend on SNV characteristics such as annotation information. We plan to investigate the properties of both standardization and modelling of the shrinkage parameter on data from the UK Biobank. We also plan to use the UK Biobank data to investigate how the shrinkage parameter depends on phenotype characteristics such as prevalence and heritability. Application of the logF approach to data from the UK Biobank will also confirm that the methods scale to biobank-sized datasets. Please see Chapter 3 for the application of logF approach to the UK Biobank with the combination of REGENIE method.

It should be noted that in our simulations we used a simplified, binary confounding variable to represent population stratification. By contrast, the analysis of the Super Seniors data used an expanded set of confounding variables that included sex and 10 principal components. We have also mentioned adjustment for relatedness and population stratification by inclusion of an estimated polygenic effect as an "offset" in the model. Another extension of interest is to use log-F penalization for a SNV covariate of interest in a model that uses LMMs to correct for confounding due to population structure and genetic relatedness [32, 39]. LMMs can be viewed as regression with correlated errors, using a kinship matrix derived from anonymous SNVs to model correlations. It should be straightforward to extend this regression approach to include log-F penalization of the SNV 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 as well as provide insights into disease etiology. Recent studies have found that patients with low genetic risk scores (GRS) are more likely to carry rare pathogenic variants [42]. 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.

Our focus has been on single-SNV logistic regression, but log-F penalization generalizes to multiple-variant logistic regression. In general, we multiply the likelihood by as many log-F distributions as there are covariates whose coefficient we wish to penalize. This can also be implemented by a generalization of the data augmentation procedure described in Section 2.6 [19, 57]. Such an approach may be useful for performing the kinds of gene- or region-based tests that are commonly performed for rare variants, and investigation of its properties is ongoing.

In Section 2.2.4, we used the concept of composite likelihood to formulate the marginal likelihood of m by multiplying individual components of marginal likelihoods for each variant.

The composite likelihood serves as a "pseudo-likelihood" that inherits numerous desirable properties of inference derived from the full likelihood function. Notably, it does not require the assumption of independence between variants. Therefore, the methodology described in Section 2.2 allows for linkage disequilibrium (LD) between variants. For example, in the application study of the Super Seniors data, the 57 SNPs chosen do not appear to be independent. In the presence of LD, it is expected that adjacent variants will have similar strength of association to the outcome. When incorporating variants with LD into the shrinkage parameter estimation procedure, redundant information about the prior distribution of the log-OR is introduced. One may penalize such variants with high correlation by adjusting the corresponding weights in the composite likelihood. Currently, this form of adjustment is being explored and such investigation is a part of future work.

Chapter 3

Whole-genome regression analysis with log-F penalization: application to UK Biobank

3.1 Introduction

GWAS continue to grow in terms of sample size, the number of phenotypes, and the number of variants being analyzed. Standard methods, such as those employed in programs like PLINK, which use simple linear or logistic regression, have become outdated due to their inability to account for population structure and underlying genetic relatedness. To address this issue, LMMs and WGR models are now commonly used to account for population structure and relatedness. For example, the fastGWA LMM approach models genetic correlations within the sample using a sparse representation [29]; SAIGE utilizes a logistic mixed model approach with p-values obtained by a saddle-point approximation (SPA) to the distribution of the Wald test statistic [61]; BOLT-LMM and LEMMA are WGRs which assume non-Gaussian priors for SNP effect sizes [41, 34]; and BGENIE is able to process multiple quantitative phenotypes simultaneously and the entire genetic data are read only once [5]. These advances have improved the accuracy and efficiency of GWAS by accounting for population structure and relatedness in their analyses.

In this chapter, we review a recently proposed machine-learning method called REGENIE (`https://rgcgithub.github.io/regenie/`) [44]. This method adjusts for population structure and hidden relatedness using a whole-genome regression (WGR) approach while offering significant computational efficiency compared to existing techniques. REGENIE operates in two main steps, applied one phenotype at a time (see Figure 3.1). In Step 1, we split the set of SNPs into consecutive blocks, and the set of SNPs from each block is used to fit a regression model. The model fits from each block are then aggregated into a single prediction, which is further decomposed into 23 predictions using a leave-one-chromosome-out (LOCO) scheme. These resulting predictions can be considered as estimated polygenic effects of individual phenotype values based on the genetic data. In Step 2, the LOCO predictions from Step 1 are utilized as an offset term in either linear or logistic regression, depending on the phenotype, to test each SNP in the genotyping array. This two-step process allows for efficient analysis while accounting for population structure and hidden genetic relatedness.

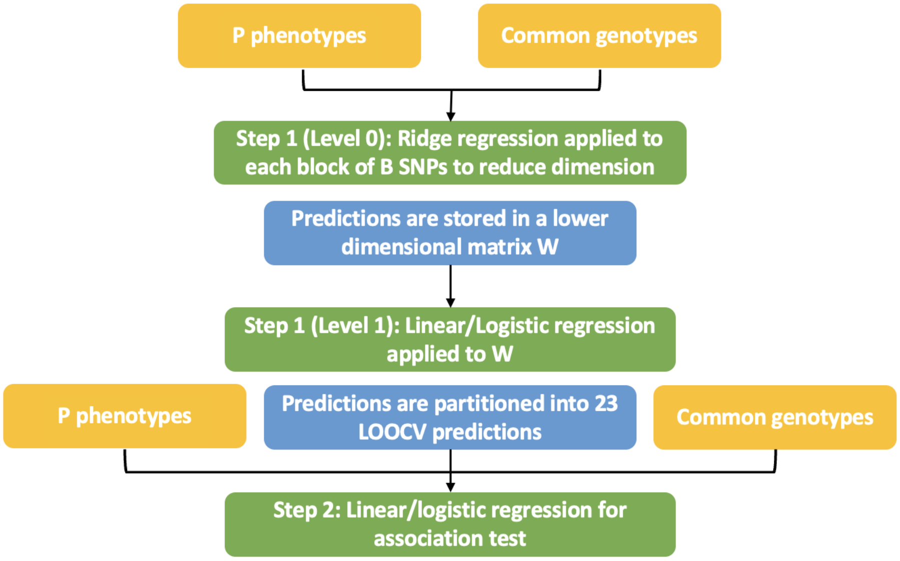

P phenotypes
Common genotypes
Step 1 (Level 0): Ridge regression applied to each block of B SNPs to reduce dimension
Predictions are stored in a lower dimensional matrix W
Step 1 (Level 1): Linear/Logistic regression applied to W
P phenotypes
Predictions are partitioned into 23 LOOCV predictions
Common genotypes
Step 2: Linear/logistic regression for association test

Figure 3.1: Overview of the REGENIE method, created based on Extended Data Figure 1 of [44].

This approach has several advantages over other competing approaches. First, the strategy of splitting SNPs into blocks reduces the memory usage. In Step 1, only a subset of SNPs is read at once for each block, which avoids the need to load the whole-genome into memory. This reduction in memory storage can lead to significant cost savings, particularly on cloud-based platforms. Second, this method is capable of analyzing multiple phenotypes in parallel. Files containing variants in both Step 1 and Step 2 can be reused instead of being repeatedly generated for each phenotype, so the genotype data only needs to be read once for all phenotypes, resulting in substantial gains in speed.

As discussed in Chapter 2, when dealing with phenotypes of extremely unbalanced case-control ratios, the asymptotic test used in MLE does not perform well for testing rare variant associations. REGENIE implements two solutions to address this issue in Step 2. The first solution is to use Firth logistic regression, which is a penalized likelihood method having the penalty based on the observed Fisher information matrix [10, 24]. The second solution is to use a SPA test rather than a normal approximation to model the distribution of score statistics. As previous approach we have proposed to control sparse data bias in single rare variant analysis in Chapter 2, we plan to explore extending REGENIE to incorporate the log-F-penalized logistic regression approach in its association testing step. Our simulation studies (see Section 2.3) demonstrated that the log-F-penalized approach exhibits several desirable properties. Although it initially requires extra computation time for the preliminary scan and the selection of the shrinkage parameter m, once the parameter is chosen, it is nearly three times faster than Firth logistic regression. Furthermore, the log-F approach is highly effective at controlling the bias and MSE of the log-OR estimates, particularly for variants with a frequency less than 5%.

Although we have applied the log-F-penalized method to a motivating example from the Super Seniors study in Section 2.4, it is worthwhile to explore its application to biobank-sized datasets, such as the UK Biobank (UKBB). As discussed in Section 2.5, the objective is to investigate the properties of modeling the shrinkage parameter in a more flexible way and to understand how the shrinkage parameter depends on phenotype characteristics, such as prevalence and heritability. More importantly, such an application would demonstrate the method's capability to scale to biobank-sized datasets. In Section 2.5, we also mentioned the possibility of including an estimated polygenic effect as an offset in the model, which leads to the idea of combining REGENIE with the log-F approach. Therefore, this chapter aims to conduct an application study using the UKBB, in which we incorporate the log-F penalization approach into REGENIE Step 2, and include REGENIE's adjustment for population stratification and hidden relatedness as an offset in the marginal likelihood for estimating m. We refer to this approach as REGENIE-LogF. The chapter is organized as follows: in Section 3.2, we introduce the workflow for combining the REGENIE and log-F approaches; in Section 3.3, we describe the steps taken for quality control and ethnicity checks of UKBB; we present our results in Section 3.4 and provide concluding remarks in Section 3.5.

3.2 Methods

REGENIE proceeds in two main steps that are applied one phenotype at a time. In this adaptation, we perform a whole-genome regression (WGR) identical to REGENIE Step 1. In Step 2, we use the log-F-penalized method for rare variant association testing. The estimated polygenic effect from REGENIE is included as an offset in the likelihood for both the estimation of the shrinkage parameter m and the association test. We introduce the REGENIE method in Section 3.2.1 and describe three modifications we have made to the log-F penalization approach to accommodate the UKBB data in Section 3.2.2.

3.2.1 Two-step Scheme of REGENIE

Step 1: Fit a WGR to a set of SNPs from across the whole genome

Given a sample of N subjects, let y denote the phenotype vector, G represent the genotype matrix, G^s be the standardized genotype matrix, comprised of columns of G centered by their mean and scaled by their SD, and X denote the covariate matrix. The standardized genotype matrix G^s are partitioned into blocks, each containing B consecutive and non-overlapping SNPs such that that SNPs in the same block are from the same chromosome. For each block i of B SNPs, we fit a ridge regression

$$\tilde{y} = \tilde{G}_i \gamma + \epsilon \text{ with } L_2 \text{ penalty} \tag{3.1}$$

where $\tilde{y} = P_X y$ and $\tilde{G}_i = P_X G^s_i$ are genotype and phenotype residuals by removing the covariate effects using a projection matrix

$$P_X = I_N - X(X^T X)^{-1} X^T. \tag{3.2}$$

The estimate $\hat{\gamma}$ can be viewed from a Bayesian framework as the maximum *a posteriori* (MAP) estimator where a Gaussian prior is used for the marker effect sizes. The ridge regression penalty parameter can be written as a function of the SNP heritability $\lambda = M(1-h^2_g)/h^2_g$. For each block, J different ridge parameters $\lambda_1, \ldots, \lambda_J$ are generated based on evenly spaced h^2_g values within the range of [0.01, 0.99]. Consequently, a set of J ridge regression predictions are obtained from this Level 0 ridge regression.

The resulting Level 0 ridge regression predictors are re-scaled to have variance of one and are stored in matrix W. The dimension of W should be much smaller than the number of markers M, so that memory usage would be much lower than reading the whole genotype matrix at once. For example, if $M = 1,000,000, B = 5,000$ and $J = 10$ are used, then the reduced dataset will have $J \times M/B = 2,000$ predictors. For quantitative phenotypes, a ridge regression is preformed on the lower dimensional matrix W

$$\tilde{y} = W\eta + \epsilon \text{ with } L_2 \text{ penalty} \tag{3.3}$$

where the ridge regression parameter is chosen using K-fold CV scheme; this is referred to as the Level 1 ridge regression. Let $\hat{y}$ refer to the final predictions obtained from the Level 1 ridge regression. We further decompose the genome-wide polygenic effects, $\hat{y}$, into 22 LOCO predictions, denoted as $\hat{y}_{LOCO}$. Using this LOCO scheme ensures that the resulting predictions capture polygenic effects only on the chromosomes other than the one that contains the SNP being tested.

For binary phenotypes, we use exactly the same Level 0 ridge regression approach, but a logistic ridge regression model is used at Level 1. We first fit a null model that only has covariate effects

$$\text{logit}(p_i) = X_i^T \alpha. \tag{3.4}$$

Then the effects estimated from equation (3.4), denoted as $\hat{\alpha}$, is included as an offset in the model

$$\text{logit}(p_i) = X_i^T \hat{\alpha} + W_i^T \eta \text{ with } L_2 \text{ penalty}, \tag{3.5}$$

where W_i is the Level 0 ridge regression predictions for the ith individual, and $\eta = \eta_1, ..., \eta_{BR}$ with $\eta_j \sim N(0, 1/\lambda)$. In order to avoid an extremely unbalanced case-control ratio situation in which one particular fold does not have cases, the LOOCV instead of K-fold CV is used to choose λ. Similar to with quantitative phenotypes, a LOCO scheme is applied to split resulting predictions into 22 predictions (denoted as $\hat{y}_{LOCO}$), which are then used for the association test in Step 2.

Step 2: Linear/logistic regression for association test

REGENIE tests for association of phenotype with each single variant in Step 2. The association test is carried out using the LOCO scheme, where each SNP on a chromosome is tested conditional on the Step 1 predictions ignoring the chromosome containing the SNP being tested. For quantitative phenotypes, consider a simple linear regression model for a variant g

$$\hat{y}_{resid,LOCO} = \tilde{g}\beta + \tilde{\epsilon}, \tag{3.6}$$

where $\hat{y}_{resid,LOCO} = \tilde{y} - \hat{y}_{LOCO}$ refers to the phenotype residuals, $\tilde{g} = P_X g$ and $\tilde{\epsilon} = P_X \epsilon$. The key idea is that REGENIE fits a WGR once and includes estimated polygenic effect $\hat{y}_{LOCO}$ as an offset in the model.

To test a variant $g = (g_1, ..., g_N)$ for association with a binary phenotype, we consider the following model

$$\text{logit}(p_i) = X_i^T \alpha + g_i \beta + \hat{y}_{i,LOCO}, i = 1, ..., N \tag{3.7}$$

where the polygenic effects predictions $\hat{y}_{LOCO}$ stored from step 1 are included as a fixed offset in the logistic model.

When dealing with quantitative phenotypes, any individuals with missing values for any of the phenotypes are eliminated during both the null-fitting and association testing steps. For binary phenotypes, we replace any missing phenotypes by mean-imputed values in Level 0 ridge regression (equation (3.3)); however, all observations with missing phenotypes are dropped when fitting Level 1 logistic ridge regression (equation (3.7)). The same approach is followed during the testing step, where missing observations are removed when fitting the logistic regression model and using Firth/SPA corrections.

The standard firth logistic regression is computationally demanding as it requires an iterative algorithm. Given that difficulty, [44] have developed an approximate Firth regression method that is much faster. The approach involves fitting a null model that only has covariate effects and then include both the estimated covariate effects and the LOCO predictions from Step 1 as offset terms in the Firth logistic regression test. When working with binary phenotypes, REGENIE implements either Firth or SPA corrections when the p-value is below some threshold. In our analysis, we set the threshold to be 0.05. To be more specific, we will only implement LogF, Firth and SPA corrections to SNPs with p-value less than 0.05 in order to maintain consistency for comparison.

3.2.2 Modifications on log-F Penalization

The log-F-penalized regression method we have proposed in Chapter 2 consists of two major steps: (1) The shrinkage parameter m is selected by maximizing a marginal likelihood that constructed by K markers that show evidence of association with the phenotype. Our maximization algorithm uses either a Monte carlo EM approach or a Laplace approximation; (2) Once m is selected, we implement the log-F penalty with standard logistic regression by using a data augmentation approach. To optimize the application of the methodology for practical biobank data usage and enhance its adaptability on a genome-wide scale data such as UKBB, we have implemented three significant modifications.

First, as we discussed in Chapter 2 Section 2.2.4, it is possible to extend the marginal likelihood approach for estimating m to include any non-genetic confounders and estimated polygenic effects. Here we incorporate confounding variables and the LOCO predictions $\hat{y}_{LOCO}$ obtained from REGENIE Step 1 as an offset in the regression model. The extended logistic regression likelihood is

$$L(\alpha^*, \gamma, \beta) = \prod_{i=1}^{n} \frac{\exp\left(Y(\alpha^* + Z_i\gamma + \hat{y}_{LOCO} + X_i\beta)\right)}{1 + \exp\left(\alpha^* + Z_i\gamma + \hat{y}_{LOCO} + X_i\beta\right)}, \tag{3.8}$$

where α^* is the logistic intercept, Z_i is vector of confounders, and γ is a corresponding confounding effects.

Second, we allow the prior distribution to be indexed by a variant-frequency-specific parameter instead of a common parameter for all variants. We multiply the global shrinkage parameter m by a frequency-specific parameter λ_k, so the kth variant effect has a log-$F(\lambda_k m, \lambda_k m)$ prior distribution. We choose $\lambda_k = [MAF_k(1-MAF_k)]^{\alpha}$ so that rare variants receive more shrinkage than common variants. To determine suitable α values, we conduct a simulation study analyzing four values (-1/4, -1/3, -1/2, -3/4) and find that -1/4 provides a better estimate than the standard log-F method with a common m, without causing over-shrinkage in extreme cases (see Supplementary Figure B.1).

Lastly, analyzing the UKBB suggests that the shrinkage parameter selection also depends on phenotype characteristics such as prevalence. The log-F-penalized method aims to mitigate the sparseness of genetic covariates in the data, but the sparseness of outcomes also plays an important role in the selection of m. For common phenotypes, such as glaucoma (case: control = 1:23) and diabetes (case: control = 1:18), the likelihood of m is monotonically increasing, suggesting an infinite value of m. This is potentially because most of the SNP effects associated with common phenotypes are densely centered at zero (Figure 1.1 shows the log-$F(m, m)$ distribution tends towards a point mass at zero as m approaches infinity). In situations where we are not able to obtain a finite estimate of m, we recommend setting $m = 1$, which provides the least shrinkage to log-ORs. Though somewhat counter-intuitive to set the value of m to minimal shrinkage when the data suggest maximal shrinkage, we treat the monotone likelihood as a case where the estimator of m does not exist and choose a default value that gives similar shrinkage to Firth's method in our analyses (see Results below).

3.3 Data Preprocessing

3.3.1 Select a High Quality Dataset by Setting Thresholds

The UKBB (`https://www.ukbiobank.ac.uk/`) is a large-scale observational database that contains genetic and health records of about 500,000 UK individuals. The UKBB has already undergone comprehensive quality control (QC) as described in [5]. However, some of the results of this QC are provided as additional information along with the UKBB data. We aim to obtain a high-quality subset by setting stricter thresholds:

1. Remove makers with missingness rate ≥ 0.02.

2. Remove individuals with missingness rate ≥ 0.02.

3. Remove imputed markers with imputed score ≤ 0.8.

4. Remove markers with Hardy-Weinberg equilibrium exact test p-value $\leq 1 \times 10^{-7}$.

5. Remove markers with MAF < 0.01.

3.3.2 Ethnicity Check

Differences in ancestral background can be a source of confounding that leads to inflated p-values in GWAS. Therefore, we choose to analyze only a set of individuals with relatively homogeneous ancestry, such as white British or European ancestry. Ancestry classification of individuals can be established by performing principal component analysis (PCA) on the combined genotype panel of the study data and a reference dataset with known ethnicities. In this case, we use the 1000 Genomes Project [8] as the reference dataset and select European ancestral samples from the UKBB. The major steps involved are:

1. Genetic variants in the study data (i.e. UKBB) are pruned in linkage disequilibrium (LD) with an $r^2 > 0.2$ in a 50kb window. We then filter the reference data (i.e. 1000 Genome Project) for the same list of pruned variants in the study data.

2. Study genotypes and reference data are merged. After removing problematic SNPs (e.g., multi-allelic, mismatched, duplicate), a total of 214,599 SNPs are identified with the same ID and extracted from both data sets.

3. We perform PCA on the merged data using PLINK 2.0. Identifying individuals of divergent ancestry is implemented in `R` using `check_ancestry()` in package `plinkQC`.

4. After removing non-European samples, the final dataset contains 437,979 markers and 458,676 individuals, in which 209,686 males, 248,803 females, and 187 of unspecified sex.

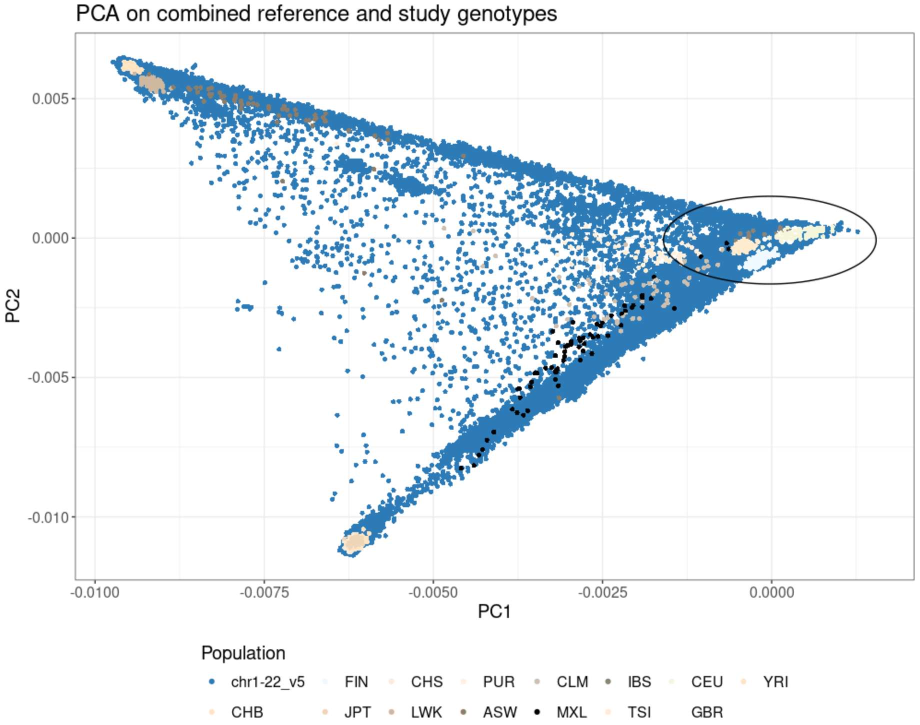


Figure 3.2

3.4 Results

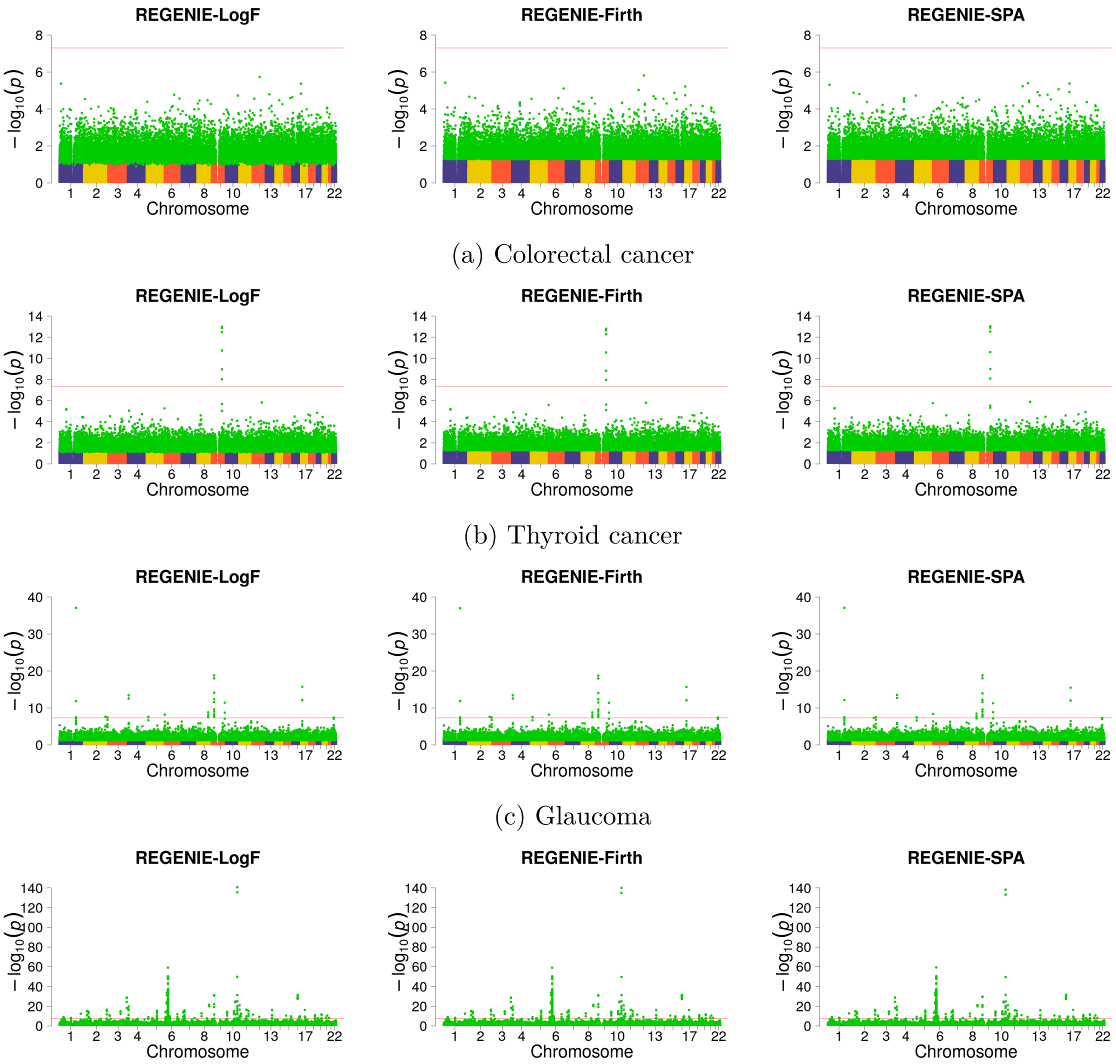
All of our analyses were conducted on the UK Biobank Research Analysis Platform (RAP), which can be found at https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform. It is supported by DNAnexus technology (`https://www.dnanexus.com/`) and driven by the power of Amazon Web Services (AWS). A variety of software environments are pre-installed on this platform for research purposes. In the steps of data pre-processing, we utilized `PLINK` (v1.90) and `PLINK2` (v2.00). The steps involved are thoroughly detailed in Appendix B.1. Our main analysis was performed using `REGENIE` (v3.1.1). Some of the useful commands that we used during this step can be found in Appendix B.2. In REGENIE Step 2, we wrote an `R` function that can be used as a `PLINK` plug-in to implement the log-F penalization (see `R` plug-in commands in Appendix B.3).

In our analysis, we focused on four binary phenotypes, notably those with unbalanced case-control ratios. The four binary traits are colorectal cancer (case:control = 1:60), thyroid cancer (case:control = 1:83), glaucoma (case:control = 1:23) and diabetes (case:control = 1:18). REGENIE includes implementation of LogF, Firth and SPA corrections were applied if p-value is less than 0.05. In the implementation of log-F penalization method, we selected the value of m following the methodologies outlined in Section 4.2.2, which include updates specific to the UKBB data. We incorporate the estimated polygenic effect from REGENIE

Step 1 into the logistic likelihood as an offset, adjusting covariates such as age, sex, and the first 10 UKBB PCs. For the colorectal and thyroid cancers, we estimated m to be 1.83 and 1.58 respectively, with each genetic variant receiving a degree of shrinkage equals to $\lambda_k m$ that varies based on the frequency of the variant - rarer variants receive more shrinkage than common ones (see Supplementary Figure B.2). On the other hand, for glaucoma and diabetes, a standard m value of one was assigned to all variants, as we were unable to obtain a finite estimate of m due to the high prevalence of these phenotypes.

The association results of the four binary traits, each with various case-control ratios, are plotted in the Manhattan plots shown in Figure 3.3. All three approaches demonstrated very good agreement for the associations tested, but the LogF correction tends to inflate some p-values beyond the typical threshold of 0.05 following penalization. Figure 3.4 shows the QQ-plot of p-values from the three approaches. In the case of colorectal cancer, the p-values from all methods closely follow the expected red line with a slope of one. However, the p-values derived from the LogF correction exhibit a slight deviation below the line at $\log_{10}$ p-values around 4.5. For other binary traits, the p-values have inflated tails, indicating minor SNP effects.

Figure 3.5 and Figure 3.6 compare the p-values and effect size estimates of the three approaches, stratified by MAF. When comparing LogF with Firth penalization, rare variants are more likely to receive a larger p-value, suggesting a decreased type I error. Overall, the p-values for LogF and SPA are quite similar, especially for common variants. The estimated effect sizes align well across the methods, although the LogF estimates show a greater degree of shrinkage towards zero compared to the Firth estimates. This shrinkage is particularly significant for rare variants compared to those with a frequency exceeding 0.05. Figure3.6 suggests that the LogF method may shrink negative estimates more than positive estimates, but this may be due to the fact that there are more extreme negative estimates from rare variants, which are shrunken more. Further investigation of any preferential shrinkage for negative estimates is ongoing. In the cases of glaucoma and diabetes, the LogF results closely match those obtained using the Firth method. This similarity suggests that the Firth correction is somewhat analogous to the log-$F(1,1)$ penalization at a genome-wide scale.



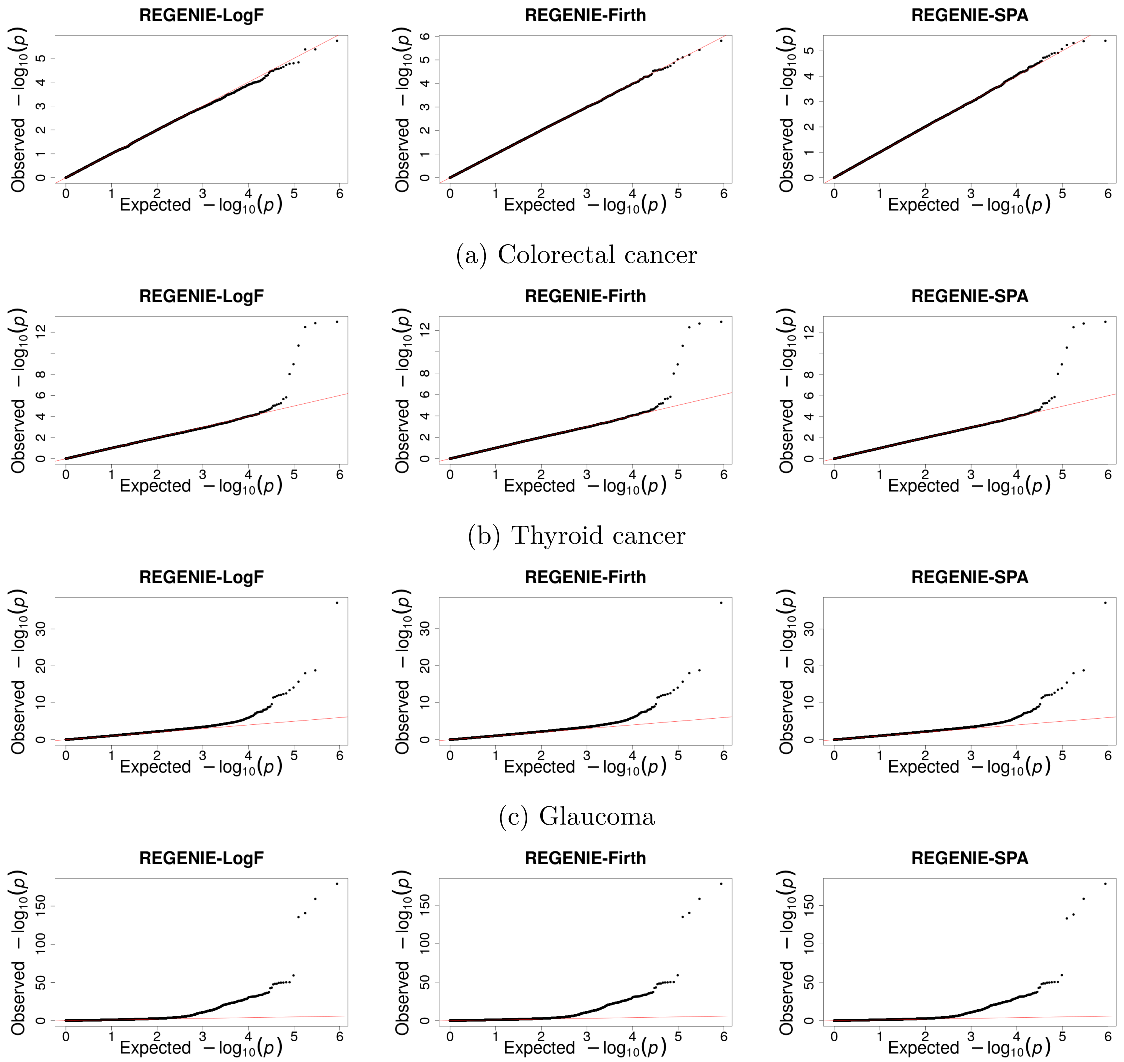
(a) Colorectal cancer

(b) Thyroid cancer

(c) Glaucoma

(d) Diabetes

Figure 3.3: Manhattan plots comparing GWAS results from REGENIE using LogF, Firth and SPA corrections for four binary phenotypes using UK Biobank European samples for (a) colorectal cancer (case:control = 1:60), (b) thyroid cancer (case:control = 1:83), (c) glaucoma (case:control = 1:23) and (d) diabetes (case:control = 1:18). SNPs with p-values less than 0.05 are highlighted green, where 0.05 is the p-value threshold below which LogF/Firth/SPA correction is applied. The orange horizontal line marks the genome-wide significance $P = 5 \times 10^{-8}$.



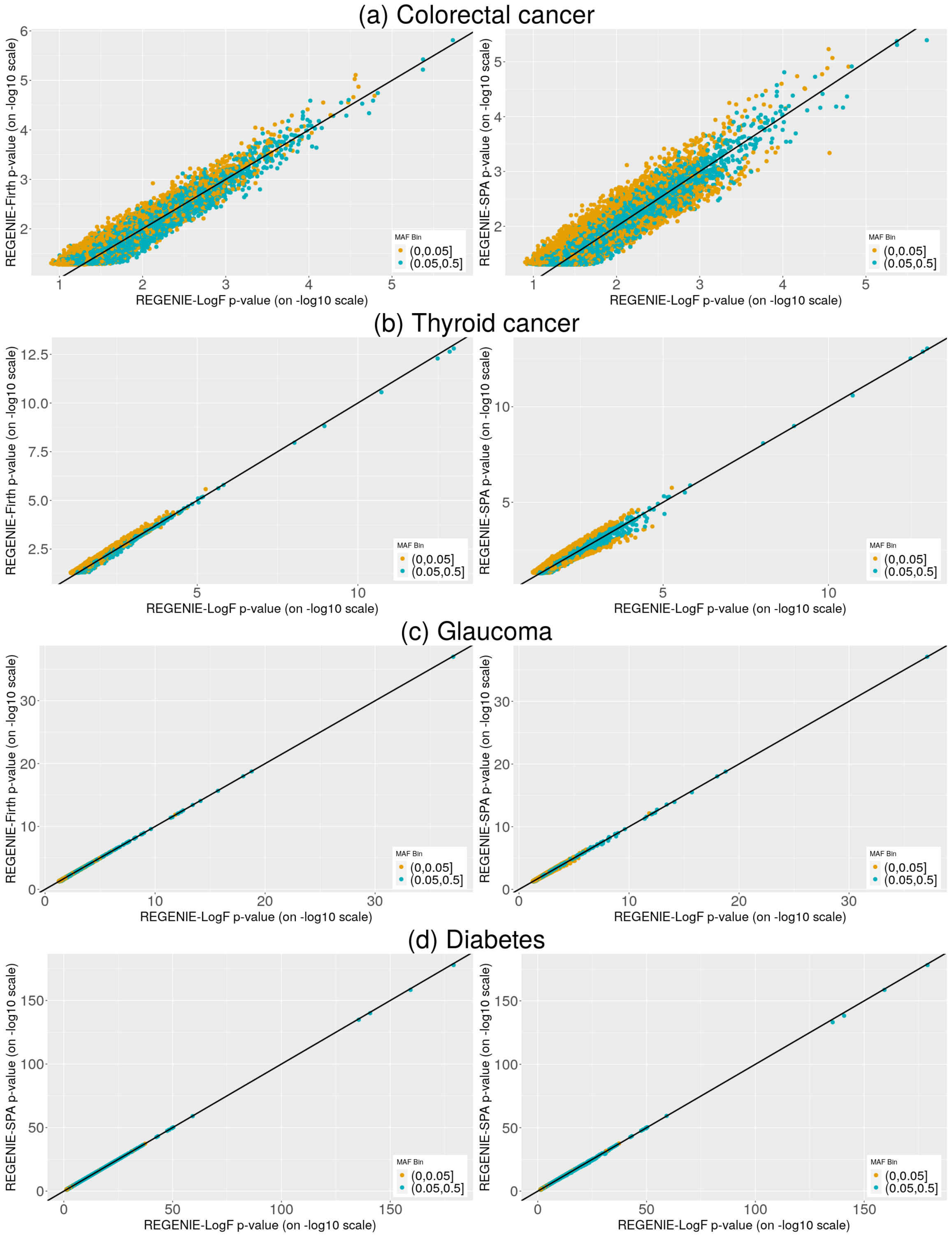
(a) Colorectal cancer

(b) Thyroid cancer

(c) Glaucoma

(d) Diabetes

Figure 3.4: Quantile–quantile plots comparing GWAS results from REGENIE using LogF, Firth and SPA corrections for four binary phenotypes using UK Biobank European samples for (a) colorectal cancer (case:control = 1:60), (b) thyroid cancer (case:control = 1:83), (c) glaucoma (case:control = 1:23) and (d) diabetes (case:control = 1:18).

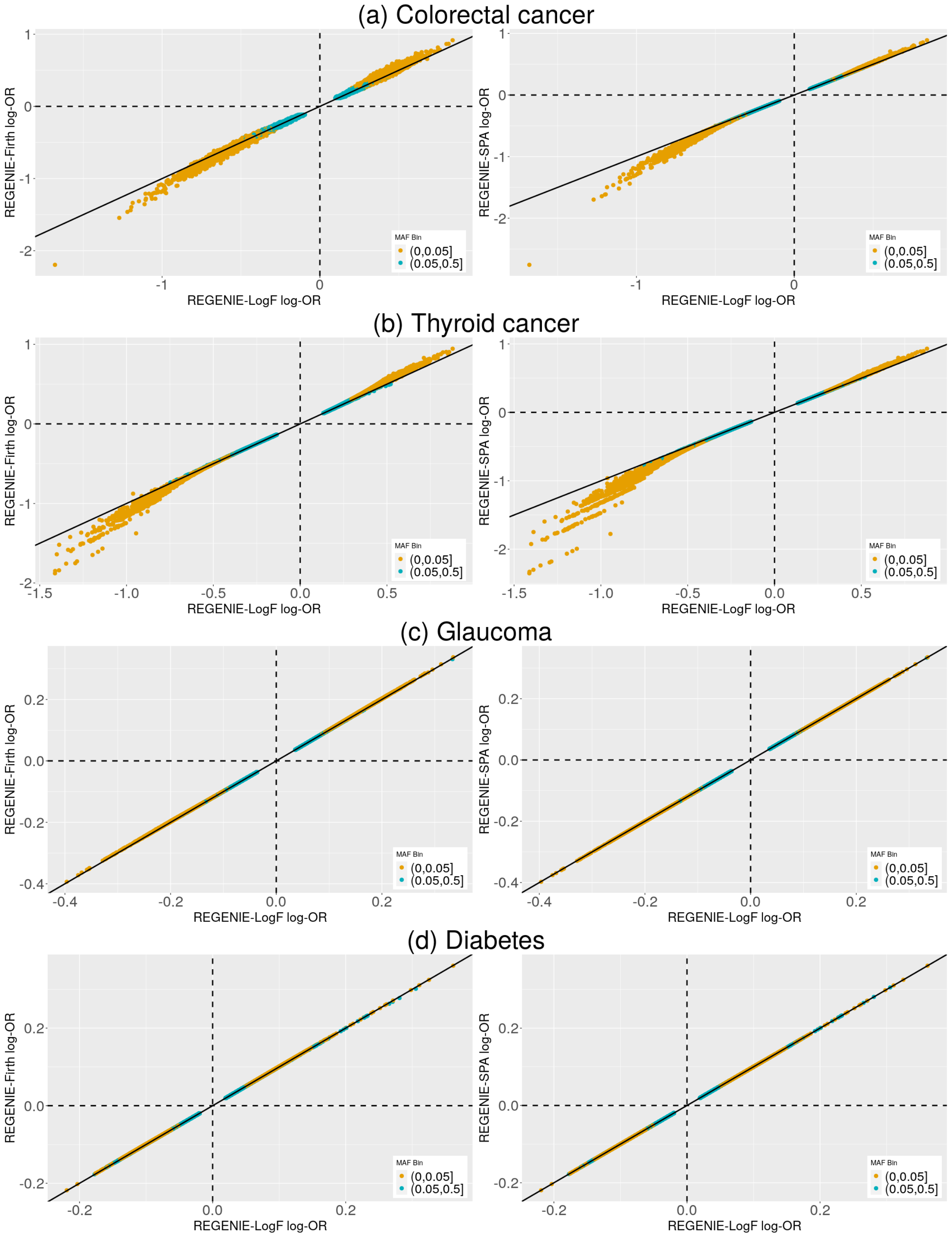

(a) Colorectal cancer

(b) Thyroid cancer

(c) Glaucoma

(d) Diabetes

Figure 3.5: Scatterplots comparing p-values from REGENIE using LogF, Firth and SPA corrections for four binary phenotypes using UK Biobank European samples. The plotting colors represents variant categories based on minor allele frequency (MAF) threshold of 5%. Only SNPs with p-values less than 0.05 are plotted, where 0.05 is the p-value threshold below which LogF/Firth/SPA correction is applied.



(a) Colorectal cancer

(b) Thyroid cancer

(c) Glaucoma

(d) Diabetes

Figure 3.6: Scatterplots comparing effect size estimates $\hat{\beta}$ from REGENIE using LogF and Firth corrections or no correction (SPA) for four binary phenotypes using UK Biobank European samples. The plotting colors represents variant categories based on minor allele frequency (MAF) threshold of 5%. Only SNPs with p-values less than 0.05 are plotted, where 0.05 is the p-value threshold below which LogF/Firth/SPA correction is applied.

3.5 Discussion

In this chapter, our focus lies in the application of the log-F penalization method proposed in Chapter 2 to the efficient whole-genome regression method REGENIE. We have introduced the log-F penalty within its association testing step. To better adapt the method to real-world biobank data and improve its flexibility, we have made several updates to our approach described in Chapter 2. Firstly, similar to many existing mixed-model-based approaches, REGENIE enables adjustment of genetic relatedness that exists within the sample. We now include predictions obtained from its null model fitting step as an offset in the likelihood calculation in order to account for population structure. Secondly, we have introduced a more flexible approach by allowing for varying degrees of shrinkage for each variant, rather than applying a common value of m for all variants. Lastly, we investigate the properties of shrinkage parameter selection based on the phenotype's prevalence, which provides valuable insights into the adaptability of our method.

Analysis of large scale bio-bank data for which binary phenotypes often results in substantial case-control unbalance, REGENIE provides Firth and SPA corrections under this scenario. We propose to use log-F penalization as an alternative. Our study demonstrates comparable results between the Firth, SPA, and log-F approaches. As expected, the log-F method shrinks the coefficients more towards zero and leads to less significant p-values, especially for rare variants. In our analysis of glaucoma and diabetes, we observed a monotonically increasing likelihood of m. This finding indicates that there are a numerous casual variants with relatively small effects to the phenotype, so the distribution of the regression coefficients are concentrated around zero, which suggests for a large or even infinite value of m. When the likelihood for m is monotone the MLE does not exist and it is not possible to use a data-informed prior. We instead choose to use a diffuse prior by setting $m = 1$, which we observed to give similar shrinkage to Firth's method. One item for future work is to select SNVs for estimating m based on their estimated effect sizes from the preliminary GWAS, rather than based on their p-values. Population-genetic principles suggest that rare variants would tend to have larger effect sizes, but power to detect these effects for rare variants is notoriously low. Thus, screening on p-values may miss large effect sizes and consequently inflate estimates of m, leading to over-shrinkage. Conversely, screening on effect sizes is likely to use estimates that are biased away from zero (the so-called "winner's curse" [35]), which could lead to under-shrinkage. Evaluation of different approaches to selecting SNVs for estimation of m is an area of future work.

Interestingly, we show that the log-$F(1,1)$ is somewhat identical to the Firth correction, illustrating the interchangeability between the two methods. A significant benefit of using log-F is that it is almost three times faster than Firth, even though it requires extra time for the preliminary scan and parameter selection. It should be noticed that REGENIE-Firth uses an approximate Firth regression approach, which is almost equivalent to the exact Firth regression but much faster [44]. In REGENIE-LogF, since we perform the association testing step using `PLINK` with `R` plug-in functions, we cannot compare the compute time of it to REGENIE-Firth Step 2, which is implemented as a C++ program [44]. Therefore, no proper conclusion can be made about the comparison of computational costs between REGENIE-LogF and REGENIE-Firth in the current analyses. The equivalency between the log-$F(1,1)$ and Firth methods implies that selecting $m = 1$ can be an alternative solution in situations where we want to impose minimal shrinkage to log-ORs or when getting a finite estimate of m is not feasible due to specific properties of the phenotype. This choice ensures that the amount of shrinkage applied to the estimates is approximately equivalent to that obtained using the Firth method, but may efficiently reduce the time of computation.

As discussed in Section 2.5, an extension worth considering is to include log-F penalization within a LMM to adjust for population structure and genetic relatedness. Further investigation suggested that such an extension is quite challenging due to the difficulty in deriving the genetic relationship matrix within the random effect component in the presence of pseudo-observations. Consequently, a more practical approach to account for population stratification and relatedness is to include the estimated polygenic effect as an offset in the model, which is exactly what we have done in this application. We also investigate the properties of centering and scaling the genotype data for the purpose of estimating m using data from the UKBB, but found no advantage to doing so. After standardizing, we observe that the majority of log-ORs tend to concentrate closely around zero, leading to a monotonically increasing likelihood of m.

In conclusion, our study demonstrates the feasibility and adaptability of applying log-F penalization to the analysis of genome-wide, biobank-scale data. We have shown that this approach, while offering some advantages in terms of computational cost, remains competitive compared to the well-established Firth method. However, the performance of REGENIE-LogF and REGENIE-Firth is not completely understood without exact computing time comparison between the two methods. Future work on implementing the log-F methodology directly into the source code of REGENIE would be beneficial. This integration would allow for a more optimized implementation of the log-F method within the REGENIE framework. In addition, given the limitations of time and computing resources, our investigation only focuses on four binary phenotypes from the UKBB dataset. We should acknowledge that these phenotypes may not be representative of all phenotypes. Therefore, repeating our analyses with a broader range of phenotypes could be helpful to gain a more comprehensive understanding of the properties of our approach. Furthermore, an area for future work would be to explore the relationship between the shrinkage parameter and the heritability of each phenotype.

Chapter 4

Log-F-penalized conditional logistic regression for sparse data

4.1 Introduction

Conditional logistic regression (CLR) is used for the analysis of binary outcomes when subjects are obtained by stratified or matched sampling with many small strata or matched sets [23]. Conditioning on sufficient statistics for stratum-specific nuisance parameters eliminates them from the likelihood and allows consistent maximum likelihood inference of the regression parameters of interest [3]. Matching is especially useful when a measurable matching variable, such as school or family, acts as a surrogate for confounders that are difficult to measure, such as environmental exposures, socioeconomic factors or genetic ancestry [43]. In such cases, an unconditional logistic regression analysis that incorporates confounders may not be possible and conditional logistic regression analysis may be the only option [4].

Conditional logistic regression also arises in other contexts, such as studies involving genetic data on children affected with a disease and their parents. The analysis of data from such case-parent trio studies is typically conditional on the parental genotypes. For a given genetic marker, conditioning on parental genotypes creates a matched set comprised of the alleles transmitted to the affected child and "matched" pseudo-controls that represent other possible combinations of alleles that the parents could have transmitted.

For small or sparse datasets, the conditional maximum likelihood estimator (CMLE) of regression coefficients is known to be biased away from zero [15, 17, 23] and is infinite when there is separation [1]. Here sparse means small numbers in categories of a categorical exposure or confounding variable [17], and separation means that there is a linear combination of covariates that can perfectly differentiate between cases and controls. Separation is illustrated by the data of Herbst et al. [25]. These authors sought to identify exposures that might explain a cluster of vaginal cancers in young women in Boston in the late 1960s. This

Exposure	**Cases**	**Controls**	**OR**
yes	7	0	NA
no	1	32	

Table 4.1: Data on DES exposure in cases and controls from Herbst *et al.* [25, Table 2]. The odds ratio is estimated without regard to matching.

was the first study to identify maternal treatment with diethylstilbesterol (DES) as a risk factor for vaginal cancer in exposed daughters. The study follows a matched case-control design. Using hospital records, patients were matched to four controls based on birth date and type of room (ward *versus* private). The matching was intended to provide control for unmeasured socioeconomic factors. The DES exposure data are summarized in Table 4.1. Coding DES exposure as one for exposed and zero for unexposed, all cases have DES ≥ 0 and all controls have DES ≤ 0, and hence the CMLE of the DES effect is infinite. Standard conditional logistic regression software reports a lack of convergence when applied to these data. We say that *near* separation occurs when a linear combination of the covariates nearly, but not perfectly, distinguishes cases from controls (e.g., this linear combination of covariate vectors *tends* to be greater in cases than in controls, though there is some overlap between the two groups). Separation and near separation can also occur with continuous exposures. The bias of the CMLE from near separation is illustrated in the simulation study of Section 4.3.

For matched pairs data, there is a simple data-augmentation approach to reducing bias and avoiding possibly infinite CMLEs. To illustrate the approach, Table 4.1 shows DES exposure status for a subset of the data of Herbst *et al.* that includes all eight matched sets, but only one control per matched set. The CMLE of the OR for DES exposure from these data is the ratio of the case-exposed/control-unexposed to case-unexposed/control-exposed cells [3], which is $7/0 = \infty$. The data-augmentation approach is to add an arbitrary number k to each cell of Table 4.1, and optionally re-scale the numbers in each cell so that the augmentation does not change the table total [2]. Adding $k = 1/2$ is known as Haldane's method and adding $k = 1$ is Laplace's method [17]. In the example, and without re-scaling, Haldane's method yields an estimated OR of 7.5/0.5=15 and Laplace's method yields an estimate of 8/1=8. We show in section 4.2 that Haldane's and Laplace's methods are penalized conditional logistic regression estimates that use the log-$F(1,1)$ and log-$F(2,2)$ distributions, respectively, for the penalty term. Greenland and Mansournia [19] have developed log-F-penalized likelihood methods for *unconditional* logistic regression. This approach has the advantage of being easily implemented by applying standard logistic regression to a dataset augmented with m pseudo-observations per penalized regression coefficient. Greenland and Mansournia suggest choosing the degrees of freedom of the log-F distribution such that the

		Control	
		exposed	unexposed
Case	exposed	0	7
	unexposed	0	1

Table 4.2: DES exposure status for the eight cases and a single matched control

spread of the prior corresponds to the research's prior belief about the plausible range of coefficient values.

Heinze and Puhr [23] developed point- and interval-estimators of odds-ratios from matched case-control data following Firth [10]. Firth's method is penalized-likelihood inference, where the penalty term derives from the Jeffreys prior [27], a data-dependent distribution and hence not a true subjective prior [13, 19]. Heinze and Puhr found that the conditional Firth-penalized likelihood (CFL) point estimator had lower bias and the associated interval estimator better coverage than other methods in their study.

In this chapter, we investigate the use of log-F-penalized likelihood methods for conditional logistic regression and compare them to Firth-penalized and unpenalized conditional logistic regression. Such an approach adapts the unconditional logistic-regression method of [19] to conditional logistic regression, or alternatively the Haldane and Laplace estimators for matched-pairs data to more general strata, and to continuous exposures and confounding variables. In Section 4.2 we develop point and interval estimators based on a log-F-penalized conditional logistic likelihood. In Section 4.3 we conduct a simulation study of the statistical properties of the log-F-penalized method and compare to conditional logistic and Firth-penalized conditional logistic regression. In Section 4.4 we apply the log-F-penalized approach to the DES data of Herbst *et al.* [25] and case-parent trio data on children with type 2 diabetes from Frayling *el al.* [11]. Section 4.5 gives concluding remarks and items for future work.

4.2 Methods

4.2.1 Conditional Logistic Regression

Suppose we have data sampled from I strata, indexed $i \in (1, ..., I)$. For notational simplicity we assume that each stratum contains one case and M_i controls. Let $j \in (0, ..., M_i)$ index subjects within strata, with the case having index $j = 0$ and controls $j = 1 \ldots, M_i$. Then let $\boldsymbol{X}_j^i$ denote the random covariate vector for the jth individual in the ith stratum, and $\boldsymbol{x}_j^i$ the vector of observed values. The parameters of interest are the log-ORs $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)^T$ corresponding to risk factors $X_1, ..., X_K$. Following [46], we can get the profile likelihood derived from the stratified retrospective likelihood for the matched case-control data

$$\begin{aligned} L(\boldsymbol{\alpha}^*, \boldsymbol{\beta}) &= \prod_{i=1}^{I} \frac{\exp\left(\alpha_i^* + \mathbf{x}_0^{iT}\boldsymbol{\beta}\right)}{1 + \exp\left(\alpha_i^* + \mathbf{x}_0^{iT}\boldsymbol{\beta}\right)} \prod_{j=1}^{M_i} \frac{1}{1 + \exp\left(\alpha_i^* + \mathbf{x}_j^{iT}\boldsymbol{\beta}\right)} \\ &= \prod_{i=1}^{I} \prod_{j=0}^{M_i} \frac{\exp\left(y_{ij}\left(\alpha_i^* + \mathbf{x}_j^{iT}\boldsymbol{\beta}\right)\right)}{1 + \exp\left(\alpha_i^* + \mathbf{x}_j^{iT}\boldsymbol{\beta}\right)}, \end{aligned} \tag{4.1}$$

where α_i^* is the ith stratum effect in the stratified logistic regression [14]. There are $K + I$ parameters in the unconditional logistic regression likelihood, with I increasing as we collect more matched sets. Consistency of the MLE fails because the number of parameters increases with the sample size [3].

Stratum effects can be eliminated from the likelihood by conditioning on their sufficient statistics. We describe the conditioning approach for a single matched set, and then use the product of conditional likelihoods over the I matched sets for inference about $\boldsymbol{\beta}$. For matched set i, the sufficient statistic for $\boldsymbol{\beta}$ is $T_i = \sum_{j=0}^{M_i} y_{ij}\boldsymbol{X}_j^i$ and the sufficient statistic for the nuisance intercept is the values of the covariate vectors in the matched set [3]. The distribution of T_i given the covariate vector values is of the form [3].

$$L_i(\boldsymbol{\beta}) = f(T_i = t_i|\boldsymbol{\beta}) = \frac{C(t)\exp(t_i^T\boldsymbol{\beta})}{\sum C(t_i^*)\exp(t_i^{*T}\boldsymbol{\beta})} \tag{4.2}$$

with $t_i = \sum_{j=0}^{M_i} y_{ij}\boldsymbol{x}_j^i$ is the observed value of T_i, $t_i^* = \sum_{j=0}^{M_i} y_{ij}^*\boldsymbol{x}_j^i$, $\boldsymbol{y}^{i*} = (y_{i1}^*, ..., y_{iM_i}^*)$ a permutation of $\boldsymbol{y}^i$, and $C(t^*)$ denotes the number of permutations of $\boldsymbol{y}^{i*}$ that lead to the particular value t_i^*. The sum in the denominator is over all t_i^* that can be obtained in this way. This distribution varies over permutations of the disease-status vectors $\boldsymbol{y}^i = (y_{i1}, ..., y_{iM_i})$ within each matched set.

Notice that $\boldsymbol{y}^i$ consists of a single one at y_{i0} and M_i zeros elsewhere, so that $t = \boldsymbol{x}_0^i$ and $\exp(t^T\boldsymbol{\beta}) = \exp(\boldsymbol{x}_0^{iT}\boldsymbol{\beta})$. Therefore, the conditional logistic likelihood for matched set i is

$$L_i(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_0^{iT}\boldsymbol{\beta})}{\sum_{j=0}^{M_i}\exp(\boldsymbol{x}_j^{iT}\boldsymbol{\beta})}. \tag{4.3}$$

Therefore, the likelihood and log-likelihood over all strata can be written as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{I} \frac{\exp(\boldsymbol{x}_0^{iT}\boldsymbol{\beta})}{\sum_{j=0}^{M_i}\exp(\boldsymbol{x}_j^{iT}\boldsymbol{\beta})} \text{ and} \tag{4.4}$$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{I} \left[\boldsymbol{x}_0^{iT} \boldsymbol{\beta} - \log \left(\sum_{j=0}^{M_i} \exp(\boldsymbol{x}_j^{iT} \boldsymbol{\beta}) \right) \right], \tag{4.5}$$

respectively. The conditional MLE, $\hat{\boldsymbol{\beta}}$, is the argument that maximizes (4.4). The conditional logistic likelihood has the same form as the Cox proportional hazards model [3], and Cox regression software is often used to maximize the conditional likelihood.

4.2.2 log-F Penalized Likelihood Method

The log-F penalized likelihood method was proposed by Greenland and Mansournia [19] for unconditional logistic regression. In this paper, we adapt it to conditional logistic regression. Penalization is by a class of log-F priors indexed by shrinkage parameter m. In this method, the log-OR parameters, $\beta_1, \ldots, \beta_K$, are assumed to be independent and have a log-$F(m,m)$ prior distribution with density

$$f(\beta_k|m) = \frac{1}{B(\frac{m}{2}, \frac{m}{2})} \frac{\exp(\frac{m}{2}\beta_k)}{(1+\exp(\beta_k))^m}, \quad k = 1, \ldots, K,$$

where $B(\cdot,\cdot)$ is the beta function. When $m = 2$ this is the standard logistic distribution [30]. The penalized conditional logistic likelihood is obtained by multiplying the likelihood by the product of K independent log-$F(m,m)$ densities, leading to the following penalized conditional likelihood and log-likelihood:

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) \times \prod_{k=1}^{K} f(\beta_k|m) = \prod_{i=1}^{I} \frac{\exp(\boldsymbol{x}_0^{iT}\boldsymbol{\beta})}{\sum_{j=0}^{M_i} \exp(\boldsymbol{x}_j^{iT}\boldsymbol{\beta})} \times \prod_{k=1}^{K} \frac{\exp(\frac{m}{2}\beta_k)}{(1+\exp(\beta_k))^m} \tag{4.6}$$

$$l^*(\boldsymbol{\beta}) = \sum_{i=1}^{I} \left[\boldsymbol{x}_0^{iT} \boldsymbol{\beta} - \log \left(\sum_{j=0}^{M_i} \exp(\boldsymbol{x}_j^{iT} \boldsymbol{\beta}) \right) \right] + \sum_{k=1}^{K} \left[\frac{m}{2}\beta_k - m\log(1+\exp(\beta_k)) \right] \tag{4.7}$$

With 1:1 pair matching, the data can be summarized into a 2 × 2 contingency table (see Table 4.1). Haldane's method adds 1/2 to each cell of the table. Apart from the constant, the augmented (penalized) likelihood then equals

$$\left[\frac{1}{2}\right]^a \left[\frac{\exp(\beta)}{1+\exp(\beta)}\right]^b \left[\frac{1}{1+\exp(\beta)}\right]^c \left[\frac{1}{2}\right]^d \times \left[\frac{\exp(\beta)}{1+\exp(\beta)}\right]^{\frac{1}{2}} \left[\frac{1}{1+\exp(\beta)}\right]^{\frac{1}{2}}. \tag{4.8}$$

Note that the augmenting factor $\exp(\frac{1}{2}\beta)/1+\exp(\beta)$ is proportional to a log-$F(1,1)$ density for β. Similarly, the augmenting factor for Laplace's method is $\exp(\beta)/(1+\exp(\beta))^2$, which is proportional to a log-$F(2,2)$ density for β. Thus, Haldane's and Laplace's methods are special cases of penalized conditional logistic regression using log-$F(1,1)$ and log-$F(2,2)$

distributions, respectively, for the penalty term.

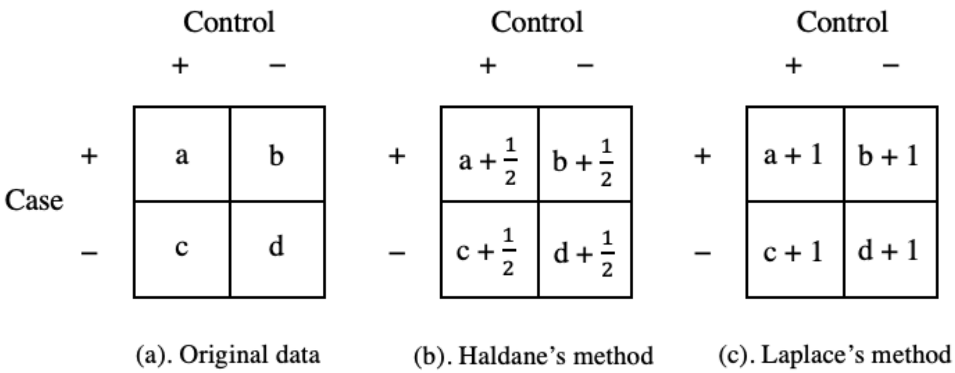
The log-F-penalized estimator, $\hat{\boldsymbol{\beta}}_F$, is obtained by solving the K modified score equations

$$\frac{\partial l^*(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^{I} \left[x_{0k}^i - \frac{\sum_{j=0}^{M_i} x_{jk}^i \exp(\boldsymbol{x}_j^{iT} \boldsymbol{\beta})}{\sum_{j=0}^{M_i} \exp(\boldsymbol{x}_j^{iT} \boldsymbol{\beta})} \right] + \frac{m}{2} - \frac{m \exp(\beta_k)}{1 + \exp(\beta_k)} = 0 \tag{4.9}$$

for $k = 1, ..., K$. Standard errors of $\hat{\boldsymbol{\beta}}_F$ can be obtained from the the inverse of the observed Fisher information

$$[I(\hat{\beta}_F)]^{-1} = \left[- \frac{\partial^2 l^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \bigg|_{\hat{\beta}_F} \right]^{-1}. \tag{4.10}$$

Specifically, the standard error for $\hat{\beta}_{Fk}$ is the square-root of the kth diagonal element of $[I(\hat{\beta}_F)]^{-1}$. Level α confidence intervals are based on the profile penalized conditional likelihood (PPCL) by solving $2[l^*(\hat{\boldsymbol{\beta}}_F) - l^*(\tilde{\boldsymbol{\beta}}_F | \beta_k)] = \chi^2_{1,1-\alpha}$ for β_k, where $l^*(\tilde{\boldsymbol{\beta}}_F | \beta_k)$ is the penalized conditional log-likelihood maximized over $\boldsymbol{\beta}_{-k}$, holding β_k fixed, and $\chi^2_{1,1-\alpha}$ is the $1 - \alpha$ quantile of the chi-square distribution with one degree of freedom.



		Control +	Control −
Case	+	a	b
	−	c	d

(a). Original data

	Control +	Control −
+	$a + \frac{1}{2}$	$b + \frac{1}{2}$
−	$c + \frac{1}{2}$	$d + \frac{1}{2}$

(b). Haldane's method

	Control +	Control −
+	a + 1	b + 1
−	c + 1	d + 1

(c). Laplace's method

Figure 4.1: Summary of a matched paired case-control study with two simple data-augmentation approaches.

4.2.3 Implementation of log-F Penalization by Data Augmentation

We can maximize the penalized likelihood by solving the modified score equations, but for even m there is a convenient data augmentation approach that can use well-tested CLR software. In unconditional logistic regression, the log-F penalization can be implemented by translating each coefficient penalty into a pseudo-data record [19]. The same idea can be adapted to conditional logistic regression, but restricting to an even degree of freedom m. We start with a simple case where we only penalize the kth covariate. The resulting penalized likelihood is

$$\begin{aligned} L^*(\boldsymbol{\beta}) &= L(\boldsymbol{\beta}) \times \frac{\exp(\frac{m}{2}\beta_k)}{(1+\exp(\beta_k))^m} \\ &= L(\boldsymbol{\beta}) \times \left[\frac{\exp(\beta_k)}{1+\exp(\beta_k)}\right]^{m/2} \times \left[\frac{1}{1+\exp(\beta_k)}\right]^{m/2} \\ &= L(\boldsymbol{\beta}) \times \prod^{m/2} \frac{\exp(\beta_k)}{1+\exp(\beta_k)} \times \prod^{m/2} \frac{1}{1+\exp(\beta_k)} \qquad (4.11) \end{aligned}$$

Notice that the log-F penalty can be transformed into a likelihood over m pseudo-observations. In particular, penalizing the conditional logistic likelihood by a log-$F(m,m)$ prior is equivalent to adding m matched-pairs (1 case and 1 control) for each covariate of interest, such that:

1. in $m/2$ strata, the case has a 1 at the covariate of interest and 0 elsewhere, and the control has 0 at all covariates, and
2. in $m/2$ strata, the case has 0 at all covariates, and the control has a 1 at the covariate of interest and 0 elsewhere.

Analyzing the augmented dataset with standard conditional logistic regression will give us the penalized CMLE and the corresponding standard errors. For odd values of m, we obtain the log-F-penalized estimator by solving equations (4.9) and the corresponding standard errors through the inverse of the observed Fisher information presented in (4.10). The optimization can be achieved using `optim()` function in `R`.

4.3 A Simulation Study

We compared the empirical performance of the log-F estimators for conditional logistic regression presented in Section 4.2 in a simulation study. We simulated data under different numbers of matched sets, exposure types (continuous or binary), exposure effect sizes and number of covariates included in the conditional logistic regression. For each simulated dataset we fit regular CMLE (Section 4.2.1), CFL [23], and log-F-penalized conditional logistic regression with (1,1), (2,2) or (3,3) degrees of freedom. Each estimator has an associated 95% confidence interval (which were derived by PPCL described in Section 4.2.2) and 5% significance test. We summarize the simulations with the number of simulations in which the estimator failed to converge, the bias, SD, and MSE of the estimator, the coverage of the confidence interval and the power of the test.

We first simulate a population of size 100,000 with a disease prevalence between 5-10% and then draw 10, 50 or 100 matched case-control sets from the population. The population data consists of a hidden variable H, exposure E, covariates Z and binary disease status

D. Strata are defined by the hidden variable H that confounds the association between the exposure and disease status. We think of the hidden variable as an unmeasured confounder that is relatively homogeneous within matched sets defined by a matching variable M that we can observe. For example, H could be an unmeasured environmental exposure that is relatively homogeneous within families, which we can observe and match on. Without information on H, we can use an analysis matched on M to control confounding.

We simulate H from a standard normal distribution. The exposure E depends on H and is either continuous or binary. For continuous E, we simulate $E|H$ from a linear regression with slope 1 and intercept 0; i.e., $E|H = h \sim N(\mu_h = h, 1)$, where μ_h is not set to impose any constraint on the population mean of the exposure. For binary E, we simulate $E|H$ from a logistic regression with slope 1 and intercept $\beta_0 = -3.37$, -2.56 or -1.65; i.e., $E|H = h \sim Bernoulli(p_h)$ where $logit(p_h) = \beta_0 + h$ and the β_0's were chosen to achieve target exposure prevalences (unconditionally) of 1/20, 1/10 and 1/5, by simulation. Disease status D depends on both E and H. D is a Bernoulli random variable with a logistic regression for the success probability

$$logit[P(D = 1|E = e, H = h)] = \beta_0 + \beta_E e + \beta_H h.$$

The exposure effect β_E is chosen to be 0.5, 1 or 1.5. We set β_H to be 2 to mimic a fairly strong confounder. We set β_0 to be -5. Empirically, this gives a disease prevalence between 5.4% and 8.4% in our simulations with continuous E and a prevalence between 3.4% and 5.8% in our simulations with binary E. The covariates Z are independent of H, E and D, and are also independent of each other and have standard normal distributions. We simulate 0, 1 or 5 covariates with the exposure. Figure 4.2 shows exposure and hidden variable values for cases and controls from 10 simulated matched sets simulated under an exposure effect size of 1.5 and a single covariate. As an aside we note that eight of 10 cases have exposure value > 1.5 and nine of 10 controls have exposure < 1.5; thus an exposure value of 1.5 nearly separates cases from controls. Increasing the exposure effect to 3 tends to produce complete separation (results not shown).

Table 4.3 shows the simulation results for continuous exposure of 10 matched case-control sets. In most settings, bias is proportional to the true parameter value and decreases in magnitude as the sample size increases (see Supplementary Material Tables). As expected, as the "events per variable" decrease with more covariates, the bias will increase [20]. When β_E becomes large, there is a substantial proportion of simulated datasets with separation, leading to infinite CMLEs. Thus, the simulation results for CMLEs were calculated based on datasets where CMLEs are finite out of 1000 replications. Overall, log-F and CFL estimates outperform CMLEs in all settings, and they could be obtained in all samples even in

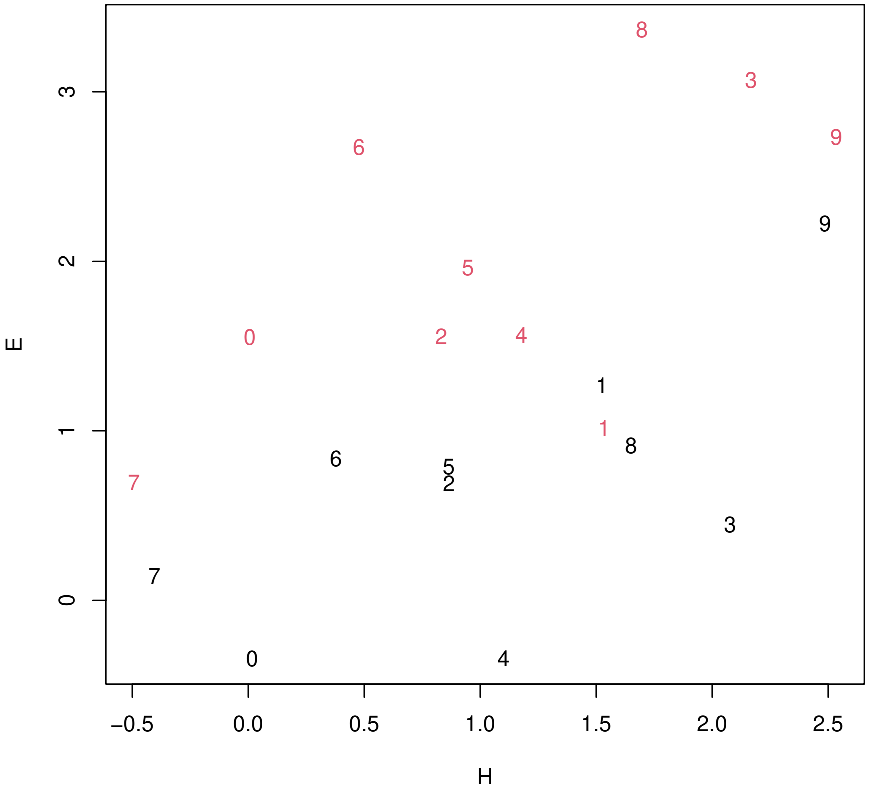


Figure 4.2: Plot of exposure E *versus* hidden variable H for one simulated dataset of 10 matched sets, with exposure effect 1.5. Matched sets are indicated by the integers 0,...9, with the case plotted in red and the matched control plotted in black. H is used only to spread the matched sets out in the horizontal direction on the plot.

situations of infinite CMLEs, although CFL may fail in case of non-convergence in extreme datasets. There is small bias in the log-F and CFL estimates when β_E is small, whereas the bias increases for large β_E. As anticipated from theory, log-F estimates exhibit greater shrinkage of the coefficient parameters as m increases, resulting in larger bias but smaller variance. Compared to CFL, it is remarkable that log-$F(2,2)$ and log-$F(3,3)$ estimates achieve smaller variance in exchange for a slight increase in bias, leading to a substantial reduction in MSE, especially for small sample sizes. In some settings, the lowest MSE obtained by log-F estimates could be twice as small as the MSE obtained by CFL.

For binary exposure, the same pattern was observed. We show the simulation results for binary exposure of 10 matched case-control sets in Table C.10. When we change the exposure prevalence from 0.05 to 0.20, the MSE slightly decreases for CFL and log-F estimates. We see that CFL and log-$F(1,1)$ generate the same results in all the combinations of simulation parameters. Mathematically, under balanced case-control ratios, CFL for a single-binary-covariate model is equivalent to imposing a log-$F(1,1)$ prior. The algebraic identity of these two models is given in the Appendix C.1.

The coverage probabilities of two-sided 95% confidence intervals based on the PPCL are reasonable in all settings. The power for log-F estimates is slightly higher than the Firth estimates for continuous exposure, and the power increases with β_E, sample size, and binary exposure prevalence. In both scenarios, log-$F(3,3)$ estimates are less powerful compared to other competing methods. This is potentially because over-shrinkage of estimates towards zero leads to a lower probability of excluding zero within the confidence interval.

ncov*	**Method**	$\boldsymbol{\beta_E}$ **= 0.5**					$\boldsymbol{\beta_E}$ **= 1.0**					$\boldsymbol{\beta_E}$ **= 1.5**				
		Bias	**SD**	**MSE**	**CP**†	**Power**‡	**Bias**	**SD**	**MSE**	**CP**†	**Power**‡	**Bias**	**SD**	**MSE**	**CP**†	**Power**‡
0	**CMLE**$^{\phi}$	30	156	252	941	187	65	207	469	958	525	73	236	611	979	776
	CFL§	4	67	46	958	127	2	76	58	965	413	-15	76	61	964	685
	log-F(1,1)	12	68	47	948	175	18	70	53	974	516	7	69	48	979	776
	log-F(2,2)	4	54	29	963	160	0	52	27	971	476	-20	47	26	977	753
	log-F(3,3)	-1	47	22	972	137	-11	43	19	968	444	-36	37	27	964	721
1	**CMLE**$^{\phi}$	28	120	152	943	190	64	196	423	963	485	69	214	507	979	700
	CFL§	1	67	44	972	120	-4	70	49	964	372	-27	72	59	949	600
	log-F(1,1)	15	75	58	956	186	26	75	64	975	518	15	70	52	977	743
	log-F(2,2)	5	58	34	976	159	3	54	29	971	475	-18	47	25	969	702
	log-F(3,3)	0	50	25	978	144	-9	44	21	970	440	-35	37	26	956	681
5	**CMLE**$^{\phi}$	6	119	141	971	46	24	131	178	957	118	32	145	220	973	216
	CFL§	-8	62	39	982	25	-37	53	41	977	71	-79	42	80	978	128
	log-F(1,1)	35	89	92	971	152	39	77	75	989	328	18	63	42	992	499
	log-F(2,2)	14	63	42	984	109	6	54	29	992	289	-21	43	23	987	458
	log-F(3,3)	5	52	27	985	82	-10	44	20	988	248	-40	35	28	984	422

Table 4.3: Simulation results for continuous exposure of 10 matched case-control sets. Bias, SD and MSE ×100, CP and Power ×1000.

* ncov = number of covariates.

$^{\phi}$ Estimate was calculated based on the datasets where CMLEs are finite out of 1000 replications. When ncov = 0, the number of infinite datasets is 17, 47 and 138 for β_E = 0.5, 1.0 and 1.5, respectively; when ncov = 1, the number of infinite datasets is 58, 156 and 320 for β_E = 0.5, 1.0 and 1.5, respectively; when ncov = 5, the number of infinite datasets is 827, 907 and 963 for β_E = 0.5, 1.0 and 1.5, respectively.

§ CFL failed in case of non-convergence. When ncov = 1, the number of fails is 2 for β_E = 1.0; When ncov = 5, the number of fails is 14, 11 and 6 for β_E = 1.5, 1.0 and 1.5, respectively.

† Coverage probability of two-sided nominal 95% confidence intervals based on PPCL.

‡ Relative frequency of confidence intervals for log-OR coefficient excluding 0.

ExpPrev*	**Method**	**β_E = 0.5**					**β_E = 1.0**					**β_E = 1.5**				
		Bias	**SD**	**MSE**	**CP†**	**Power‡**	**Bias**	**SD**	**MSE**	**CP†**	**Power‡**	**Bias**	**SD**	**MSE**	**CP†**	**Power‡**
0.05	**CMLE$^\phi$**	-29	67	54	1000	0	-54	66	73	980	6	-80	64	105	971	8
	CFL	-13	102	106	978	7	-17	95	94	970	28	-29	87	84	981	91
	log-F(1,1)	-13	102	106	978	7	-17	95	94	970	28	-29	87	84	981	91
	log-F(2,2)	-23	72	58	997	7	-39	68	61	970	28	-61	63	76	958	91
	log-F(3,3)	-28	58	41	997	2	-51	54	55	970	3	-77	51	86	958	31
0.10	**CMLE$^\phi$**	-18	76	61	997	7	-37	73	67	960	40	-53	64	69	980	101
	CFL	2	97	94	981	31	-11	91	83	966	78	-12	85	74	988	216
	log-F(1,1)	2	97	94	981	31	-11	91	83	966	78	-12	85	74	988	216
	log-F(2,2)	-13	72	54	992	31	-31	68	56	966	78	-43	63	58	969	216
	log-F(3,3)	-19	59	39	992	12	-43	56	49	966	31	-61	51	63	969	114
0.20	**CMLE$^\phi$**	-19	78	65	995	15	-29	73	62	957	47	-48	64	64	965	137
	CFL	-3	93	86	982	39	-5	90	81	965	118	-15	86	76	977	249
	log-F(1,1)	-3	93	86	982	39	-5	90	81	965	118	-15	86	76	977	249
	log-F(2,2)	-14	71	52	991	39	-25	68	52	965	118	-44	64	60	952	249
	log-F(3,3)	-20	59	38	992	16	-37	56	45	965	59	-61	52	64	952	130

Table 4.4: Simulation results for binary exposure of 10 matched case-control sets with no covariate.

Bias, SD and MSE ×100, CP and Power ×1000.

* ExpPrev = exposure prevalence.

$^\phi$ Estimate was calculated based on the datasets where CMLEs are finite out of 1000 replications. When ExpPrev = 0.05, the number of infinite datasets is 402, 461 and 517 for β_E = 0.5, 1.0 and 1.5, respectively; when ExpPrev = 0.10, the number of infinite datasets is 238, 281 and 413 for β_E = 0.5, 1.0 and 1.5, respectively; when ExpPrev = 0.20, the number of infinite datasets is 173, 241 and 341 for β_E = 0.5, 1.0 and 1.5, respectively.

† Coverage probability of two-sided nominal 95% confidence intervals based on PPCL.

‡ Relative frequency of confidence intervals for log-OR coefficient excluding 0.

4.4 Data Application

4.4.1 DES

Our method is examined using real data and compared with the CFL regression proposed by [23]. The data aims to examine the effect of diethylstilbestrol (DES) exposure during pregnancy on the subsequent development of vaginal cancer in daughters [25]. The study consists of eight young women with vaginal cancer, each matched with four controls. The matching is intended to reflect socioeconomic factors. The use of DES by their mothers during pregnancy is compared to see if DES treatment is more prevalent among mothers of the cases. Among the mothers of the eight cases, seven were exposed to DES during pregnancy, while none of the mothers of the controls had received DES treatment. The data on DES exposure and maternal smoking are summarized in Table 4.5.

We aim to examine the effect of DES exposure using conditional logistic regression, adjusting for the effect of maternal smoking. As mentioned earlier, separation occurs in the data because the DES covariate effectively separates cases from controls: DES ≥ 0 for all cases and DES ≤ 0 for all controls. A standard conditional logistic regression analysis of these data using the `clogit()` function from the `survival` package in `R` warns that the DES coefficient has not converged, resulting in infinite Conditional Maximum Likelihood Estimation (CMLE) (see Figure 4.3 (a)).

Exposure	Level	Cases	Controls	OR
DES	yes	7	0	`NA`
	no	1	32	
Maternal smoking	yes	7	21	3.57
	no	1	11	

Table 4.5: Data on cases and controls in the Herbst et al. study reconstructed from their Table 2. Odds ratios are estimated by conditional logistic regression.

The results are presented in Table 4.7. These findings align with our simulation results, which demonstrated that the estimator of log-F penalized logistic regression is pulled towards zero even more than the CFL estimators for large values of m. In terms of shrinkage of the point estimate towards zero, the log-$F(1,1)$ penalty shrinks the least, and the log-$F(3,3)$ penalty shrinks the most. As expected, the standard errors and confidence interval lengths are ordered inversely to the amount of shrinkage, with log-$F(1,1)$ having the largest spreads, and log-$F(3,3)$ having the smallest spreads. Figure 4.7 displays the Firth- and log-F-penalized profile log-likelihoods for the DES effect. The log-F penalized likelihood estimate is shrunken more towards the origin as m increases, while the standard error, which measures the curvature of the log-likelihood function at its maximum, decreases with m.

As suggested in [19], we can choose a prior based on what we believe are plausible values for the effect. For example, an $F(1,1)$ prior gives a 95% probability of the OR falling between 1/648 and 648, which extends to an order of magnitude beyond effects typically seen in health studies. In contrast, the exact 95% prior intervals for the ORs from the corresponding $F(2,2)$ and $F(3,3)$ distributions are (1/39, 39) and (1/15, 15) respectively. These ranges would still be considered plausible prior ranges for the OR, providing a more reasonable basis for analysis and interpretation.

4.4.2 Case-parent Trio Data

Another application we would like to showcase is a study that investigates genetic risk factors for type 2 diabetes. We applied the log-F penalized approach to case-parent trio data on children with type 2 diabetes, aiming to estimate genotype relative risks of a polymorphism known as the GCK1 Z+2 allele [11]. The trio data were reconstructed from Table 4 in [11]. The data, presented in Table 4.6, display the number of affected children with each genotype, stratified by the parental mating type. This analysis allows us to examine the potential association between the GCK1 Z+2 allele and the risk of developing type 2 diabetes in children, taking into account the genetic background of their parents. We denote the informative mating types by

$$G_p = \begin{cases} 0 \times 1 & \text{if one parent is heterozygous, and one parent is homozygous for the alternative allele,} \\ 1 \times 1 & \text{if if both parents are heterozygous,} \\ 1 \times 2 & \text{if one parent is heterozygous, and one parent is homozygous for the reference allele.} \end{cases} \tag{4.12}$$

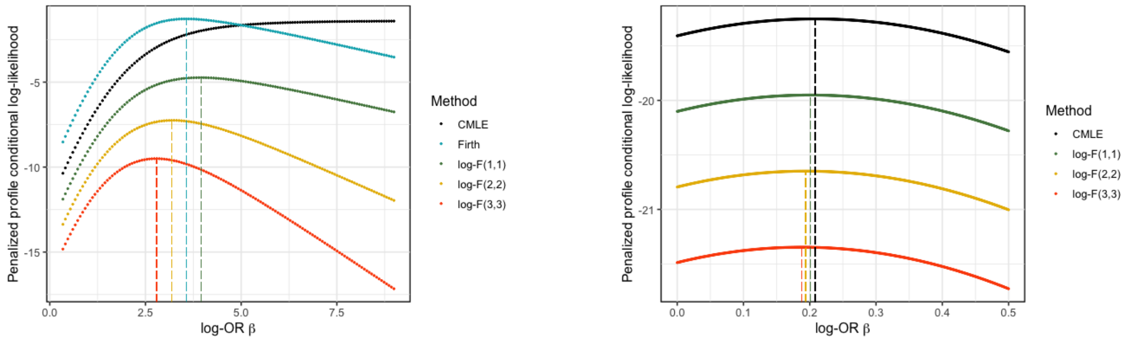
We can reconstruct Table 4.6 into a matched case-control dataset, where the cases are children affected by the disease, and the matched pseudo-controls are all possible combinations of alleles not transmitted from parents to their children, conditional on the parents' genotypes. The likelihood for the genotype effect on disease risk comes from the conditional probability of the child's genotype given the parents' genotypes. To utilize conditional logistic regression for estimating GRR, an offset term should be incorporated into the model. The explicit use of an offset is provided in Table 1 of [49]. The results are summarized in Table 4.7 and Figure 4.3 (b). We cannot obtain results for CFL because `coxphf()` issues a non-convergence warning. We are currently investigating the reason for this non-convergence. In Figure 4.3, we observed that the use of the penalized method did not yield substantial improvement over CMLE. However, the amount of shrinkage increases as the degrees of freedom increase, demonstrating the influence of penalization on the final estimates.

Parental mating type (G_p)	**Genotype (g)**		
	0	**1**	**2**
0×1	10	15	NA
1×1	1	1	0
1×2	NA	NA	NA

Table 4.6: Summary of case-parent trio data

	DES		**Case-parent trio data**	
Method	**OR estimate (95% CI)**	**Std. err**	**OR estimate (95% CI)**	**Std. err**
CFL	35.49 (5.58, 4150.47)	1.291	NA	NA
log-$F(1,1)$	50.86 (6.04, 6634.92)	1.467	1.22 (0.60, 2.55)	0.367
log-$F(2,2)$	24.80 (4.37, 465.99)	1.074	1.21 (0.60, 2.48)	0.361
log-$F(3,3)$	16.17 (3.47, 155.07)	0.900	1.20 (0.60, 2.46)	0.355

Table 4.7: Estimates, standard errors, and 95% confidence intervals for the OR in both DES and case-parent trio data, using different penalized conditional logistic regression approaches.



(a) Penalized profile conditional log-likelihoods for the effect of exposure of DES, adjusting for the effect of maternal smoking.

(b) Penalized conditional log-likelihoods for the genotype relative risk of GCK1 Z+2 allele on type 2 diabetes.

Figure 4.3: Conditional log-likelihoods curves with Firth and log-F penalties. The corresponding maximum penalized-profile-likelihood estimators are indicated by vertical long-dashed lines.

4.5 Discussion

In this paper, we present a novel approach for small sample conditional logistic regression analysis by using a class of log-F penalties. Our proposed approach improves on existing methods in several ways. First, it provides finite OR estimates in cases where the CMLE is not defined. Second, the log-F-penalized estimator has smaller MSE than the CFL estimator proposed by [23] in most scenarios. Third, the log-F-penalized method is simple to implement using data augmentation and standard software.

The choice of shrinkage parameter, m, is crucial for the log-F-penalized method. Although large values of m will result in smaller MSE, our simulation results show no advantages in preferring large values of m for inference of large ORs. When we include binary covariates in the model, the log-F estimator shows larger bias than than the CFL estimator as m increases, but is superior in terms of SD and MSE. Overall, our simulation results suggest that the log-$F(2,2)$ penalty is a good choice, providing a substantial decrease in MSE without over-shrinking estimates towards zero. In addition, the log-$F(2,2)$-penalized estimation can be implemented by a simple data augmentation approach that is more computationally efficient than CFL estimation.

Our recommendation of a log-$F(2,2)$-penalized estimator is based on limited simulation results and the convenience of its implementation. An alternative is to choose the prior based on what the analyst believes is a plausible range of values for the odd-ratio coefficient [19]. For example, a log-$F(1,1)$ distribution puts 95% of coefficients between 1/648 and 648 and a log-$F(2,2)$ distribution puts 95% of coefficients between 1/39 and 39. Another alternative is to adopt the Jeffreys prior as in CFL. However, authors such as [19] argue against such a data-dependent prior. Among other objections, the Jeffreys prior has the disadvantage of changing the marginal prior for a given log-odds-ratio as covariates are added to the model [13]. Regardless of the prior, we echo the sentiment of [12] that penalization is under-utilized in medical statistics and penalized likelihood methods such as the one we have proposed should be considered.

In Chapter 2, we have developed an EB method to estimate the shrinkage parameter m in the log-F prior. The EB approach operates under the assumption that we can effectively estimate the parameter of the random effects distribution if we have multiple independent and identically distributed realizations (i.i.d) of random effects. In the context of genetic epidemiology, we have data from a single case-control study involving multiple genetic variants. In this scenario, we can assume the genetic variants are independent, and the regression coefficient of each genetic variant are i.i.d realizations from the common log-F prior distribution. However, for conditional logistic regression, we only have data from one matched case-control study with a single exposure variable. To implement the EB approach, we need multiple independent studies sharing identical outcome and exposure variables, and then we could combine data from these studies to estimate m. This would be one area of future work.

Chapter 5

Conclusion

In this thesis, we investigate penalized likelihood methods using a class of log-$F(m,m)$ priors to address the issue of sparse-data bias in logistic and conditional logistic regression analysis, with applications to genetic epidemiology. To start with, we developed log-F penalized logistic regression in Chapter 2. This includes the construction of a marginal likelihood for the shrinkage parameter m, and its maximization using approximate maximization algorithms. Once m is determined, a standard logistic regression can be used on an augmented dataset to implement the log-F penalization approach. Finally, we provide simulation studies and a real data application to understand the properties of the proposed method and illustrate its performance relative to existing approaches.

In Chapter 3, we present an application of the log-F penalization method combined with the REGENIE approach to the UK Biobank data. We integrate the log-F penalty into REGENIE Step 2, and include the polygenic effect estimated from REGENIE Step 1 as an offset in the logistic likelihood. We allow for various degrees of shrinkage for different variants by introducing a frequency-specific parameter into the log-F density function. Our findings reveal that using a conservative approach by assuming $m = 1$ for minimal shrinkage yields similar results as that obtained by the Firth correction. This suggests that the log-$F(1,1)$ penalty can be a viable option that ensures approximately the same amount of shrinkage achieved through the Firth method while taking the advantage of computational efficiency.

In Chapter 4, we extend the log-F penalization method to conditional logistic regression for the analysis of stratified data, such as data from matched case-control studies. Penalization can be implemented by adding pseudo-observations into the original dataset only if m is even. We demonstrate the equivalence between conditional Firth logistic regression and log-$F(1,1)$ penalization in a matched-pair case-control study with a single binary exposure covariates. Based on our findings, we recommend the log-$F(2,2)$ penalty because it is easily implemented using a data augmentation approach and has substantially lower MSE than other estimators without over-shrinkage.

In conclusion, in this thesis we have expanded the scope of log-F-penalized regression, by adapting the penalization method to conditional logistic regression, investigating methods for selecting the shrinkage parameter for logistic regression, and exploring the feasibility and properties of log-F-penalized logistic regression on large-scale biobank data. We have primarily concentrated on single-covariate logistic regression, but the concept of log-F penalization can be extended to multi-covariate logistic regression, which might be useful for performing the gene- or region-based tests typically employed for rare variants. The essence of such an extension is to incorporate one log-F density into the likelihood function for each covariate that we intend to penalize. This can be implemented by a generalization of the data augmentation procedure outlined in Section 2.6. Research on the properties of this extension is ongoing.

Bibliography

[1] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):pp. 1–10, 1984.

[2] Y. M. Bishop, S. E. Feinberg, and P. W. Holland. *Discrete Multivariate Analysis*. Springer-Verlag New York, 1975.

[3] N. E. Breslow and N. E. Day. *Statistical Methods for Cancer Research: Volume 1 – the analysis of case-control data*. IARC Scientific Publications, 1980.

[4] B. A. Brumback, A. B. Dailey, and H. W. Zheng. Adjusting for confounding by neighborhood using a proportional odds model and complex survey data. *Am J Epidemiol*, 175(11):1133–1141, Jun 2012.

[5] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[6] H. Chen, C. Wang, M. P. Conomos, A. M. Stilp, Z. Li, T. Sofer, A. A. Szpiro, W. Chen, J. M. Brehm, J. C. Celedón, S. Redline, G. J. Papanicolaou, T. A. Thornton, C. C. Laurie, K. Rice, and X. Lin. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am J Hum Genet*, 98(4):653–666, Apr 2016.

[7] 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.

[8] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

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

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

[11] Timothy M Frayling, Mark Walker, Mark I McCarthy, Julie C Evans, Lisa I Allen, Steve Lynn, Susan Ayres, Barbara Millauer, Catherine Turner, Robert C Turner, et al. Parent-offspring trios: a resource to facilitate the identification of type 2 diabetes genes. *Diabetes*, 48(12):2475–2479, 1999.

[12] Sarah Friedrich, Andreas Groll, Katja Ickstadt, Thomas Kneib, Markus Pauly, Jörg Rahnenführer, and Tim Friede. Regularization approaches in clinical biostatistics: A review of methods and their applications. *Statistical Methods in Medical Research*, 32(2):425–440, 2023.

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

[14] 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.

[15] S. Greenland. Small-sample bias and corrections for conditional maximum-likelihood odds-ratio estimators. *Biostatistics*, 1(1):113–122, Mar 2000.

[16] S. Greenland, M. A. Mansournia, and D. G. Altman. Sparse data bias: a problem hiding in plain sight. *BMJ*, 352:i1981, Apr 2016.

[17] S. Greenland, J. A. Schwartzbaum, and W. D. Finkle. Problems due to small samples and sparse data in conditional logistic regression analysis. *Am J Epidemiol*, 151(5):531–539, Mar 2000.

[18] Sander Greenland. Prior data for non-normal priors. *Stat Med*, 26(19):3578–3590, 2007.

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

[20] Sander Greenland, Mohammad Ali Mansournia, and Douglas G Altman. Sparse data bias: a problem hiding in plain sight. *BMJ*, 352, 2016.

[21] J. Halaschek-Wiener, L. C. Tindale, J. A. Collins, S. Leach, B. McManus, K. Madden, G. Meneilly, N. D. Le, J. M. Connors, and A. R. Brooks-Wilson. The Super-Seniors Study: Phenotypic characterization of a healthy 85+ population. *PLoS One*, 13(5):e0197578, 2018.

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

[23] Georg Heinze and Rainer Puhr. Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Statistics in medicine*, 29(7-8):770–777, 2010.

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

[25] Arthur L Herbst, Howard Ulfelder, and David C Poskanzer. Adenocarcinoma of the vagina: association of maternal stilbestrol therapy with tumor appearance in young women. *New England journal of medicine*, 284(16):878–881, 1971.

[26] R. R. Hudson. Two-locus sampling distributions and their application. *Genetics*, 159(4):1805–1817, Dec 2001.

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

[28] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proc R Soc Lon Ser-A*, 186(1007):453–461, 1946.

[29] Longda Jiang, Zhili Zheng, Ting Qi, Kathryn E Kemper, Naomi R Wray, Peter M Visscher, and Jian Yang. A resource-efficient tool for mixed model association analysis of large-scale data. *Nature genetics*, 51(12):1749–1755, 2019.

[30] Norman Lloyd. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous univariate distributions.* Wiley, 1994.

[31] Samantha Jean Jones. *Characterization of environmental and genetic factors in multiple-case lymphoid cancer families.* PhD thesis, University of British Columbia, 2020.

[32] 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. *Nat Genet*, 42(4):348–354, 2010.

[33] Hyun Ming Kang, Mickaël Canouil, and Phuong Nguyen. *EPACTS (Efficient and Parallelizable Association Container Toolbox)*, 2022.

[34] Matthew Kerin and Jonathan Marchini. Inferring gene-by-environment interactions with a bayesian whole-genome regression model. *The American Journal of Human Genetics*, 107(4):698–713, 2020.

[35] P. Kraft. Curses–winner's and otherwise–in genetic epidemiology. *Epidemiology*, 19(5):649–651, Sep 2008.

[36] Peter Kraft, Eleftheria Zeggini, and John P. A. Ioannidis. Replication in Genome-Wide Association Studies. *Stat Sci*, 24(4):561 – 573, 2009.

[37] Fabrice Larribe and Paul Fearnhead. On composite likelihood in statistical genetics. *Stat Sinica*, 21:43–69, 01 2011.

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

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

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

[41] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 47(3):284–290, 2015.

[42] 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.

[43] M. A. Mansournia, N. P. Jewell, and S. Greenland. Case-control matching: effects, misconceptions, and recommendations. *Eur J Epidemiol*, 33(1):5–14, 01 2018.

[44] 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*, pages 1097–1103, May 2021.

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

[46] 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.

[47] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020.

[48] Alastair J Scott and CJ Wild. Maximum likelihood for generalised case-control studies. *J Stat Plan Infer*, 96(1):3–27, 2001.

[49] Ji-Hyung Shin, Claire Infante-Rivard, Brad McNeney, and Jinko Graham. A data-smoothing approach to explore and test gene-environment interaction in case-parent trios. *Statistical Applications in Genetics and Molecular Biology*, 13(2):159–171, 2014.

[50] D. Speed, N. Cai, M. R. Johnson, S. Nejentsev, and D. J. Balding. Reevaluation of SNP heritability in complex human traits. *Nat Genet*, 49(7):986–992, Jul 2017.

[51] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*, 12(3):e1001779, Mar 2015.

[52] Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *J Am Stat Assoc*, 81(393):82–86, 1986.

[53] Christiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Stat Sinica*, 21(1):5–42, 2011.

[54] Greg CG Wei and Martin A Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J Am Stat Assoc*, 85(411):699–704, 1990.

[55] 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. *Am J Hum Genet*, 89(1):82–93, 2011.

[56] Y. Yu, S. Chen, S. J. Jones, R. Hoque, O. Vishnyakova, A. Brooks-Wilson, and B. McNeney. Penalized Logistic Regression Analysis for Genetic Association Studies of Binary Phenotypes. *Hum Hered*, Jun 2022.

[57] Ying Yu. Shrinkage parameter estimation for penalized logistic regression analysis of case-control data. Master's thesis, Simon Fraser University, 2019.

[58] C. Zeng, D. C. Thomas, and J. P. Lewinger. Incorporating prior knowledge into regularized regression. *Bioinformatics*, 37(4):514–521, 05 2021.

[59] Biao Zhang. Bias-corrected maximum semiparametric likelihood estimation under logistic regression models based on case–control data. *J Stat Plan Infer*, 136(1):108–124, 2006.

[60] W. Zhou, J. B. Nielsen, L. G. Fritsche, R. Dey, M. E. Gabrielsen, B. N. Wolford, J. LeFaive, P. VandeHaar, S. A. Gagliano, A. Gifford, L. A. Bastarache, W. Q. Wei, J. C. Denny, M. Lin, K. Hveem, H. M. Kang, G. R. Abecasis, C. J. Willer, and S. Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet*, 50(9):1335–1341, 09 2018.

[61] Wei Zhou, Jonas B Nielsen, Lars G Fritsche, Rounak Dey, Maiken E Gabrielsen, Brooke N Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A Gagliano, Aliya Gifford, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics*, 50(9):1335–1341, 2018.

[62] X. Zhou, P. Carbonetto, and M. Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet*, 9(2):e1003264, 2013.

[63] X. Zhou and M. Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods*, 11(4):407–409, Apr 2014.

Appendix A

Supplementary Material for Chapter 2

A.1 Computational Considerations for MCEM

We use the weighted logistic regression approach to maximize equation (2.13) 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 (2.13) 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 Supplementary Figure A.1 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 (2.13). The offsets account for known quantities in the logistic model. In particular, the linear prediction in the logistic model for observation i in the 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 (2.13). Thus the term $x_{ik}\beta_{kj}$ is a known offset. By constructing the augmented dataset in Supplementary Figure A.1, maximizing equation (2.13) 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.

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

Figure A.1: $\boldsymbol{Y}_{(Nn\times1)}$ is a vector containing N replicates of $\boldsymbol{y}$ and $\boldsymbol{X}_{(Nn\times1)}$ 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 $\boldsymbol{O} = \{\boldsymbol{x}\beta_{kj}\}_{j=1}^{N}$.

A.2 Useful Links

The source code and scripts for the methods presented in Chapter 2 can be found at `https://github.com/SFUStatgen/logistlogF`.

A.3 Supplementary Figures

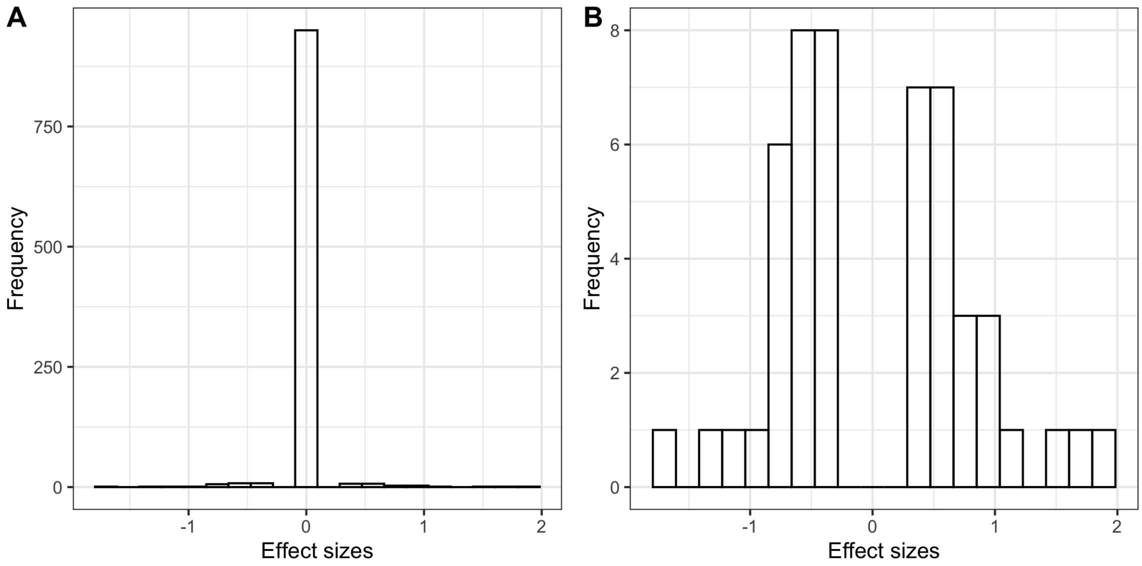


Figure A.2: A. Histogram of 1000 SNV-effect-sizes used for data simulation, in which are 5 casual SNVs and 950 null SNVs. B. Histogram of effect sizes of causal SNPs, where $\beta_j = \frac{\log 5}{2} |\log_{10} \text{MAF}_j|$.

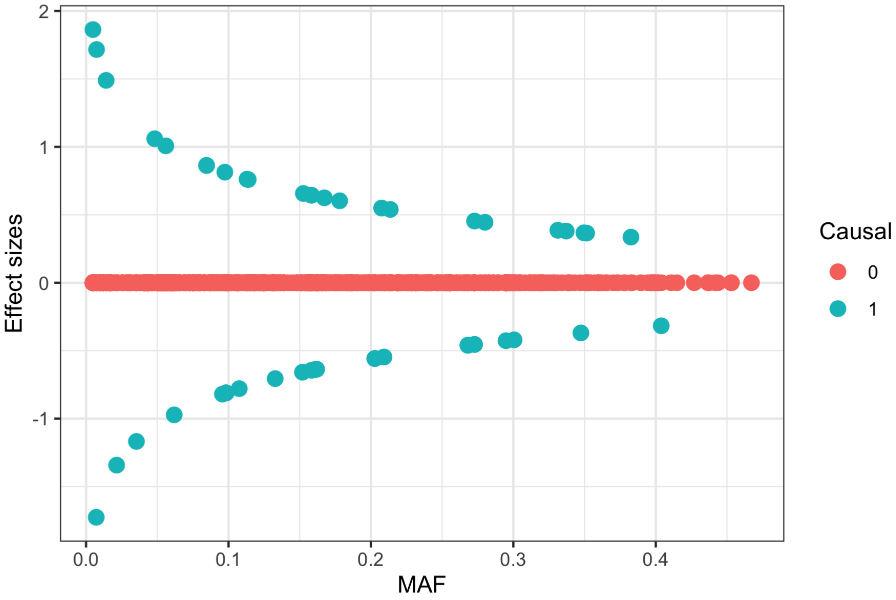


Figure A.3: Effect sizes of 1000 SNVs generated used for data simulation by minor allele frequency. Red dots indicate casual SNVs and blue dots indicate non-casual SNVs.

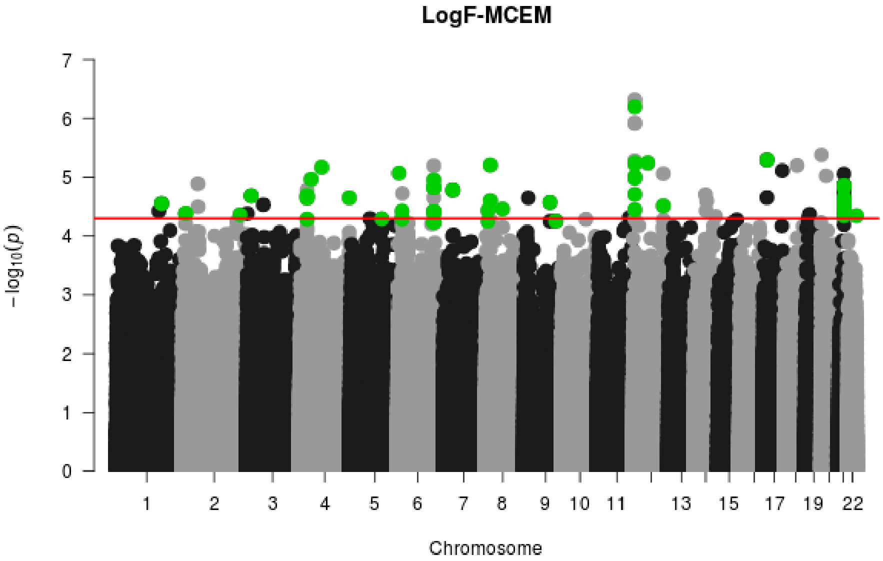


Figure A.4: Manhattan plots showing association results from LogF-MCEM on Super Seniors data. The red horizontal line represents the liberal genome-wide significance threshold ($P = 5 \times 10^{-5}$) used to select SNPs in the preliminary scan. 57 SNPs (green points) below the threshold are used to estimate m in Step 1.

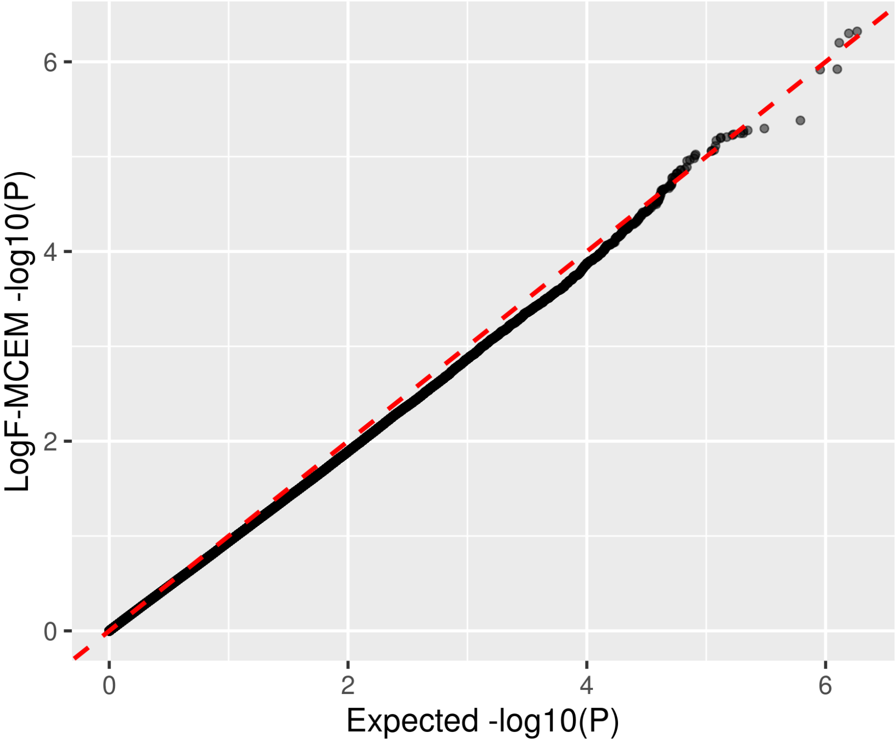


Figure A.5: QQ-plot showing p-values from LogF-MCEM on Super Seniors data. The p-value was obtained using the likelihood ratio test with a χ^2_1 test statistic.

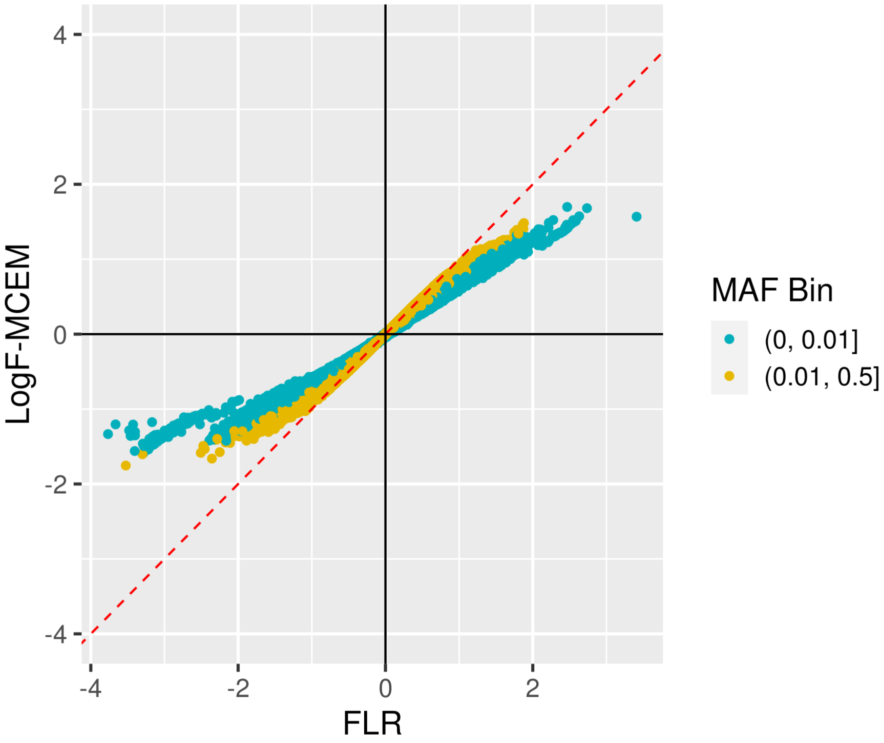


Figure A.6: Scatterplots showing effect size estimates from LogF-MCEM for Super Seniors data. The plotting colors represents variant categories based on minor allele frequency (MAF) threshold of 1%.

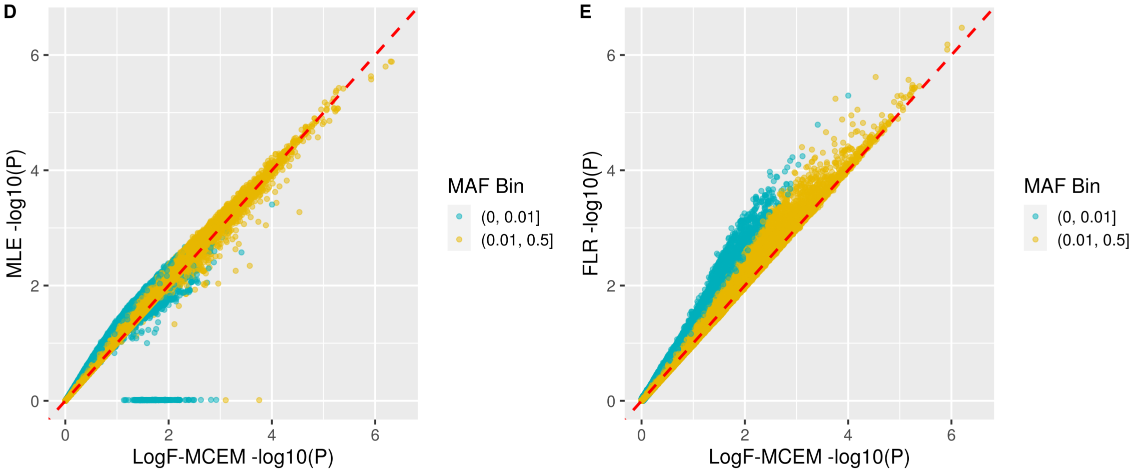


Figure A.7: Scatterplots comparing p-values from different methods on Super Seniors data. The plotting colors represents variant categories based on minor allele frequency (MAF) threshold of 1%. For FLR and LogF-MCEM, the p-value for each variant was obtained by the likelihood ratio test with a χ_1^2 test statistic.

Appendix B

Supplementary Material for Chapter 3

B.1 PLINK Useful Commands

Step 1: Combine all autosomal SNPs

```
plink --merge-list data_by_chr/all_my_files.txt \
      --make-bed \
      --out qc/chr1-22
```

Step 2: Select a high quality dataset by setting thresholds

1. Remove markers with missingness rate ≥ 0.02.

   ```
   plink --bfile qc/chr1-22 \
         --geno 0.02 \
         --make-bed \
         --out qc/chr1-22_v1
   ```

2. Remove individuals with missingness rate ≥ 0.02.

   ```
   plink --bfile qc/chr1-22_v1 \
         --mind 0.02 \
         --make-bed \
         --out qc/chr1-22_v2
   ```

3. Remove imputed markers with INFO ≤ 0.8. Resource 1967 (Imputation MAF and information scores) on `https://biobank.ctsu.ox.ac.uk/crystal/search.cgi` contains MAF and info score for each of the markers in the imputed data. The file `imp_exclude_snp.txt` lists all imputed markers with INFO ≤ 0.8.

```
plink --bfile qc/chr1-22_v2 \
      --exclude imp_info_score/imp_exclude_snp.txt \
      --make-bed \
      --out qc/chr1-22_v3
```

4. Remove markers with Hardy-Weinberg equilibrium exact test p-value $\leq 1 \times 10^{-7}$.

```
plink --bfile qc/chr1-22_v3 \
      --hwe 1e-7 \
      --make-bed \
      -out qc/chr1-22_v4 \
```

5. Remove markers with MAF < 0.01.

```
plink --bfile qc/chr1-22_v4 \
      --maf 0.01 \
      --make-bed \
      --out qc/chr1-22_v5
```

Step 3: Ethnicity check
Ethnicity check was done using commands from `https://cran.r-project.org/web/packages/plinkQC/vignettes/AncestryCheck.pdf`. Below are the summarized steps:

1. Define bash variables and the directory of study and reference data.

```
qcdir='qc'
refdir='1000gp'
name='chr1-22_v5'
refname='1kg_phase1_all'
```

2. Match study genotypes and reference data.

   (a) Filter non A-T or G-C SNPs in reference and study data, and remove them from both the datasets.

```
qcdir='qc'
refdir='1000gp'
name='chr1-22_v5'
refname='1kg_phase1_all'

awk 'BEGIN {OFS="\t"} ($5$6 == "GC" || $5$6 == "CG" \
|| $5$6 == "AT" || $5$6 == "TA") {print $2}' \
$qcdir/$name.bim > \
$qcdir/ethnicity_check/$name.ac_gt_snps

awk 'BEGIN {OFS="\t"} ($5$6 == "GC" || $5$6 == "CG" \
|| $5$6 == "AT" || $5$6 == "TA") {print $2}' \
$refdir/$refname.bim > \
$qcdir/ethnicity_check/$refname.ac_gt_snps
```

```
plink --bfile $refdir/$refname \
      --exclude $qcdir/ethnicity_check/$refname.ac_gt_snps \
      --make-bed \
      --out $qcdir/ethnicity_check/$refname.no_ac_gt_snps

plink --bfile $qcdir/$name \
      --exclude $qcdir/ethnicity_check/$name.ac_gt_snps \
      --make-bed \
      --out $qcdir/ethnicity_check/$name.no_ac_gt_snps
```

(b) Prune variants in linkage disequilibrium (LD) with an $r^2 > 0.2$ in a 50kb window.

```
plink --bfile $qcdir/ethnicity_check/$name.no_ac_gt_snps \
      --indep-pairwise 50 5 0.2 \
      --out $qcdir/ethnicity_check/$name.no_ac_gt_snps

plink --bfile $qcdir/ethnicity_check/$name.no_ac_gt_snps \
      --extract $qcdir/ethnicity_check/$name.no_ac_gt_snps.prune.in \
      --make-bed \
      --out $qcdir/ethnicity_check/$name.pruned
```

(c) Filter the reference data by using the list of pruned variants obtained from the study data.

```
plink --bfile $qcdir/ethnicity_check/$refname.no_ac_gt_snps \
      --extract $qcdir/ethnicity_check/$name.no_ac_gt_snps.prune.in \
      --make-bed \
      --out $qcdir/ethnicity_check/$refname.pruned
```

(d) Check that variants in the reference data have the same chromosome number as in the study data.

```
awk 'BEGIN {OFS="\t"} FNR==NR {a[$2]=$1; next} \
($2 in a && a[$2] != $1) {print a[$2],$2}' \
$qcdir/ethnicity_check/$name.pruned.bim \
$qcdir/ethnicity_check/$refname.pruned.bim | \
sed -n '/^[XY]/!p' > $qcdir/ethnicity_check/$refname.toUpdateChr

plink --bfile $qcdir/ethnicity_check/$refname.pruned \
      --update-chr $qcdir/ethnicity_check/$refname.toUpdateChr 1 2 \
      --make-bed \
      --out $qcdir/ethnicity_check/$refname.updateChr
```

(e) Update variants positions and possible flip alleles.

```
awk 'BEGIN {OFS="\t"} FNR==NR {a[$2]=$4; next} \
($2 in a && a[$2] != $4) {print a[$2],$2}' \
$qcdir/ethnicity_check/$name.pruned.bim \
$qcdir/ethnicity_check/$refname.pruned.bim > \
$qcdir/ethnicity_check/${refname}.toUpdatePos

awk 'BEGIN {OFS="\t"} FNR==NR {a[$1$2$4]=$5$6; next} \
($1$2$4 in a && a[$1$2$4] != $5$6 && a[$1$2$4] != $6$5) {print $2}' \
$qcdir/ethnicity_check/$name.pruned.bim \
$qcdir/ethnicity_check/$refname.pruned.bim > \
$qcdir/ethnicity_check/$refname.toFlip

plink --bfile $qcdir/ethnicity_check/$refname.updateChr \
      --update-map $qcdir/ethnicity_check/$refname.toUpdatePos 1 2 \
      --flip $qcdir/ethnicity_check/$refname.toFlip \
      --make-bed \
      --out $qcdir/ethnicity_check/$refname.flipped
```

(f) Remove alleles that do not match after allele-flips.

```
awk 'BEGIN {OFS="\t"} FNR==NR {a[$1$2$4]=$5$6; next} \
($1$2$4 in a && a[$1$2$4] != $5$6 && a[$1$2$4] != $6$5) {print $2}' \
$qcdir/ethnicity_check/$name.pruned.bim \
$qcdir/ethnicity_check/$refname.flipped.bim > \
$qcdir/ethnicity_check/$refname.mismatch

plink --bfile $qcdir/ethnicity_check/$refname.flipped \
      --exclude $qcdir/ethnicity_check/$refname.mismatch \
      --make-bed \
      --out $qcdir/ethnicity_check/$refname.clean
```

3. Merge genotypes of study data and reference data.

```
plink --bfile $qcdir/ethnicity_check/$name.pruned \
      --bmerge $qcdir/ethnicity_check/$refname.clean.bed \
               $qcdir/ethnicity_check/$refname.clean.bim \
               $qcdir/ethnicity_check/$refname.clean.fam \
      --make-bed \
      --out $qcdir/ethnicity_check/$name.merge.$refname
```

4. Perform PCA on the merged data using `PLINK 2.0`.

```
plink2 --bfile $qcdir/ethnicity_check/$name.merge.$refname \
       --pca approx \
       --out $qcdir/ethnicity_check/$name.$refname
```

5. Check ancestry. We use the `.eigenvec` file generated from last step to estimate the ancestry of the study data, which is implemented in `R` using `check_ancestry()` in package `plinkQC`. Currently, this function only supports the identification of European ancestry.

```
library(plinkQC)

indir <- "/mnt/project/qc"
qcdir <- "/mnt/project/qc/ethnicity_check"
name <- 'chr1-22_v5'
refname <- '1kg_phase1_all'
prefixMergedDataset <- paste(name, ".", refname, sep="")
```

```
exclude_ancestry <-
evaluate_check_ancestry(indir = indir, name = name, qcdir = qcdir,
                  prefixMergedDataset = prefixMergedDataset,
                  defaultRefSamples = "1000Genomes",
                  legend_text_size = 12,
                  legend_title_size = 15,
                  axis_text_size = 12,
                  axis_title_size = 15,
                  title_size = 15,
                  studyColor = "#2c7bb6",
                  interactive = TRUE)

fail_ancestry <- exclude_ancestry$fail_ancestry
write.table(fail_ancestry,file="fail_ancestry.txt",quote=F,row.names=F)
```

6. Remove non-European samples.

```
plink --bfile qc/chr1-22_v5 \
      --remove qc/ethnicity_check/fail_ancestry.txt \
      --make-bed --out qc/chr1-22_v6
```

B.2 REGENIE Useful Commands

The developers of REGENIE provide comprehensive documentation online at their GitHub pages `https://rgcgithub.github.io/regenie/`, which includes information on how to install and use the program along with various practical examples. Here are some example commands that we used:

1. Fitting the whole-genome regrssion:

```
regenie-3.1.3 \
  --step 1 \
  --bed plink_data \
  --phenoFile example.pheno \
  --covarFile example.cov  \
  --bsize 1000 \
  --loocv \
  --bsize 1000 \
  --lowmem \
  --out regenie_step1
```

In this command, `plink_data` is your input file in PLINK binary format, `example.pheno` is your phenotype file, `example.cov` is your covariate file, `--bsize1000` is the block size for fitting Level 0 ridge regression and `--loocv` flags to use leave-one out cross validation scheme. We recommend to use option `--lowmem` to reduce memory usage when analyzing more than 10 traits.

2. Running the association test:

```
regenie-3.1.3 \
  --step 2 \
  --bgen plink_data.bgen \
  --sample plink_data.sample \
  --phenoFile example.pheno \
  --covarFile example.cov \
  --bsize 1000 \
  --firth --approx --pThresh 0.05 \
  --pred regenie_step1_pred.list \
  --out regenie_firth
```

In this command, `plink_data.bgen` is your input file in BGEN format, `plink_data.sample` is your sample file, `--firth --approx --pThresh 0.05` flags to use approximate Firth regression for computational speedup with p-values less than the threshold of 0.05, and `regenie_step1_pred.list` is the list of predicted phenotype files generated from Step 1. To use SPA correction for p-values less than the threshold, we can use option `--spa`.

B.3 PLINK R Plug-in Commands

PLINK enables the extension of its basic functionality by using the R-based 'plug-in' functions. PLINK uses the `Rserve` package to communicate with R. Here is the commands for basic usage of R plug-ins:

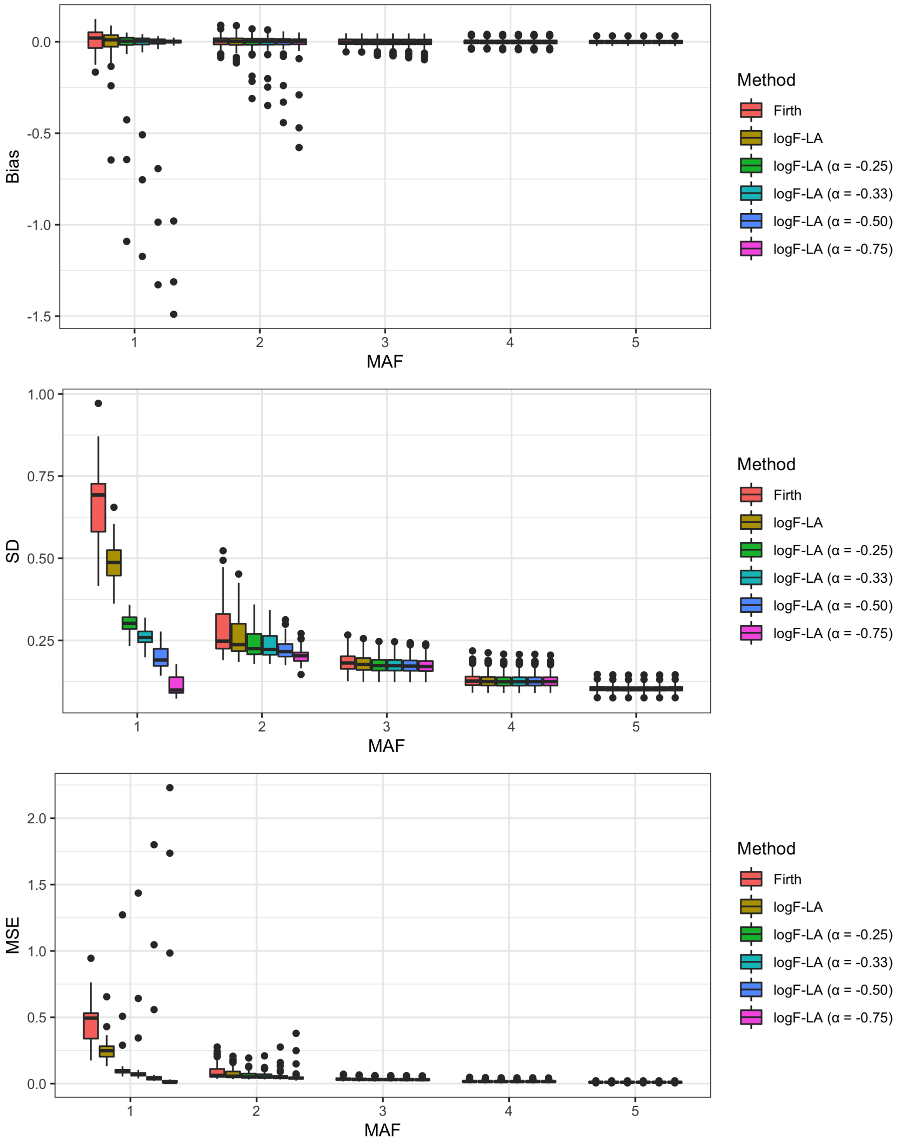
```
plink  --bfile plink_data \
       --pheno example.pheno \
       --covar example.cov \
       --R myscript.R \
       --out output_file
```

In this command, `myscript.R` is the file containing the R code with a standard format specifically for the purpose of PLINK plug-in. For details of how to define a `R` plug-in function, please refer to `https://zzz.bwh.harvard.edu/plink/rfunc.shtml`.

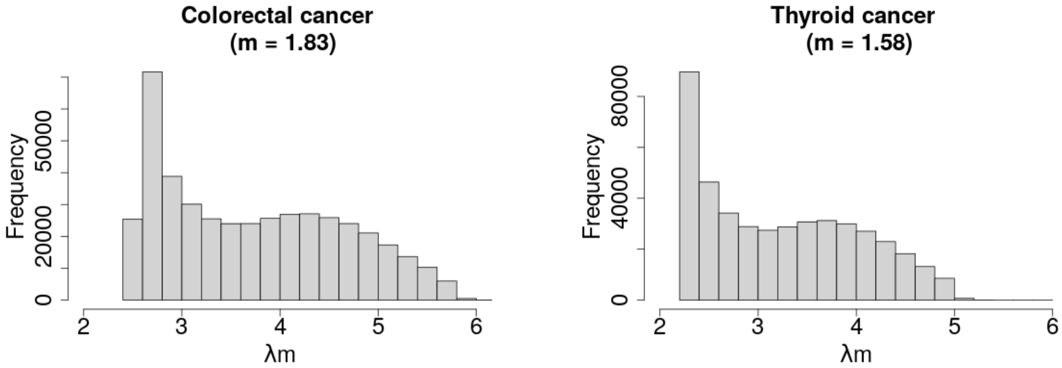
B.4 UKBB Useful Links

For phenotype information in the UKBB data, please refer to `https://biobank.ctsu.ox.ac.uk/crystal/search.cgi`.

B.5 Supplementary Figures



Supplementary Figure B.1: MAF binned boxplots of bias, SD and MSE of effect size estimates on simulated data. Each boxplot represents the distribution of the estimated quantity across 100 simulation replicates. MAF bins are: 1 = $(0\%, 1\%)$, 2 = $[1\%, 5\%)$, 3 = $[5\%, 10\%)$, 4 = $[10\%, 25\%)$ and 5 = $[25\%, 50\%]$.



Supplementary Figure B.2: Distribution of the estimated shrinkage parameter λm in the log-F prior density for colorectal and thyroid cancers, where $\lambda = [f(1-f)]^{-1/4}$ varies based on the minor allele frequency f of the variant.

Appendix C

Supplementary Material for Chapter 4

C.1 The Algebraic Identity of the CFL and log-$F(1,1)$-penalized Logistic Regression

Let us assume a matched-pair ($M_i = 1$) case-control study with a single binary exposure covariate. The penalized log-likelihood with a log-$F(1,1)$ prior is

$$l_{LogF}{}^{*}(\beta) = l(\beta) + \frac{1}{2}\beta - \log(1 + \exp(\beta)) + \text{Constant}. \tag{C.1}$$

According to [23], applying Firth's penalization method to (4.4) leads to Firth-type penalized conditional log-likelihood

$$l_{Firth}{}^{*}(\beta) = l(\beta) + \frac{1}{2}\log(|I(\beta)|) \tag{C.2}$$

where the observed information matrix $I(\beta)$, which is the minus of the second derivative of the conditional log-likelihood, is given by

$$I(\beta) = \sum_{i=1}^{I} \frac{(x_0^i - x_1^i)^2 \exp(x_0^i\beta + x_1^i\beta)}{(\exp(x_0^i\beta) + \exp(x_1^i\beta))^2}. \tag{C.3}$$

Note that if the ith pair of matched set is discordant on the exposure indicator, we have $\frac{(x_0^i - x_1^i)^2 \exp(x_0^i\beta + x_1^i\beta)}{(\exp(x_0^i\beta) + \exp(x_1^i\beta))^2} = \frac{\exp(\beta)}{(1+\exp(\beta))^2}$; otherwise, $\frac{(x_0^i - x_1^i)^2 \exp(x_0^i\beta + x_1^i\beta)}{(\exp(x_0^i\beta) + \exp(x_1^i\beta))^2} = 0$. Thus, $I(\beta) = I_0 \frac{\exp(\beta)}{(1+\exp(\beta))^2}$, where I_0 is the number of discordant pairs within data. This suggests that solving the score equations of $l_{LogF}{}^{*}(\beta)$ and $l_{Firth}{}^{*}(\beta)$ results in the same estimate of β.

C.2 Supplementary Tables

ncov*	**Method**	**β_E = 0.5**					**β_E = 1.0**					**β_E = 1.5**				
		Bias	**SD**	**MSE**	**CP**†	**Power**‡	**Bias**	**SD**	**MSE**	**CP**†	**Power**‡	**Bias**	**SD**	**MSE**	**CP**†	**Power**‡
0	**CMLE**$^\phi$	5	25	7	943	713	8	35	13	958	995	17	63	43	941	1000
	CFL	2	24	6	952	687	1	32	10	965	994	3	48	23	957	1000
	log-F(1,1)	4	24	6	946	707	5	33	11	959	995	8	47	22	953	1000
	log-F(2,2)	4	24	6	950	703	3	30	9	967	995	1	40	16	965	1000
	log-F(3,3)	3	23	5	951	700	0	29	8	966	994	-5	36	13	962	1000
1	**CMLE**$^\phi$	5	26	7	938	722	11	38	16	941	996	26	88	84	936	1000
	CFL	1	24	6	949	693	1	33	11	954	996	3	53	29	961	1000
	log-F(1,1)	4	25	6	942	719	7	35	12	948	996	13	51	28	959	1000
	log-F(2,2)	3	24	6	944	716	4	32	10	952	996	5	43	19	968	1000
	log-F(3,3)	3	24	6	949	712	2	30	9	956	996	-2	38	14	975	1000
5	**CMLE**$^\phi$	12	32	12	928	693	28	53	36	892	986	66	118	182	862	1000
	CFL	2	26	7	955	630	2	35	12	954	981	5	57	33	967	1000
	log-F(1,1)	10	30	10	938	686	21	43	23	914	985	39	68	62	913	1000
	log-F(2,2)	8	29	9	942	679	16	38	17	938	985	23	51	32	950	1000
	log-F(3,3)	7	27	8	946	677	11	34	13	948	984	13	43	20	979	1000

Table C.1: Simulation results for continuous exposure of 50 matched case-control sets. Bias, SD and MSE ×100, CP and Power ×1000.

* ncov = number of covariates.

$^\phi$ Estimate was calculated based on the datasets where CMLEs are finite out of 1000 replications. When ncov = 5, the number of infinite datasets is 15 for β_E = 1.5.

† Coverage probability of two-sided nominal 95% confidence intervals based on PPCL.

‡ Relative frequency of confidence intervals for log-OR coefficient excluding 0.

ncov*	**Method**	**β_E = 0.5**					**β_E = 1.0**					**β_E = 1.5**				
		Bias	**SD**	**MSE**	**CP**†	**Power**‡	**Bias**	**SD**	**MSE**	**CP**†	**Power**‡	**Bias**	**SD**	**MSE**	**CP**†	**Power**‡
0	**CMLE**	2	17	3	941	930	4	23	6	943	1000	7	32	11	955	1000
	CFL	0	17	3	946	922	1	22	5	952	1000	1	30	9	963	1000
	log-F(1,1)	1	17	3	942	927	3	23	5	947	1000	4	30	9	959	1000
	log-F(2,2)	1	17	3	944	926	2	22	5	948	1000	1	28	8	964	1000
	log-F(3,3)	1	17	3	944	926	1	22	5	954	1000	-2	27	7	966	1000
1	**CMLE**	3	17	3	947	940	6	24	6	944	1000	10	34	13	943	1000
	CFL	1	16	3	952	932	1	22	5	939	1000	1	31	10	953	1000
	log-F(1,1)	2	17	3	952	938	5	23	6	943	1000	6	32	10	953	1000
	log-F(2,2)	2	17	3	949	938	4	22	5	943	1000	3	30	9	960	1000
	log-F(3,3)	2	17	3	952	936	2	22	5	943	1000	0	28	8	961	1000
5	**CMLE**	5	18	4	927	938	11	26	8	930	1000	22	44	25	912	1000
	CFL	1	17	3	947	925	0	23	5	960	1000	1	34	12	935	1000
	log-F(1,1)	4	18	4	926	937	9	25	7	936	1000	17	39	18	927	1000
	log-F(2,2)	4	18	3	929	937	7	24	6	947	1000	12	36	14	941	1000
	log-F(3,3)	4	18	3	933	936	6	23	6	955	1000	8	33	11	946	1000

Table C.2: Simulation results for continuous exposure of 100 matched case-control sets. Bias, SD and MSE ×100, CP and Power ×1000.

* ncov = number of covariates.

† Coverage probability of two-sided nominal 95% confidence intervals based on PPCL.

‡ Relative frequency of confidence intervals for log-OR coefficient excluding 0.

ExpPrev*	**Method**	β_E = 0.5					β_E = 1.0					β_E = 1.5				
		Bias	**SD**	**MSE**	**CP**†	**Power**‡	**Bias**	**SD**	**MSE**	**CP**†	**Power**‡	**Bias**	**SD**	**MSE**	**CP**†	**Power**‡
0.05	**CMLE**$^\phi$	-18	116	138	984	10	-37	122	163	977	32	-70	114	179	963	20
	CFL	-11	116	135	984	13	-14	114	131	974	46	-33	107	125	961	94
	log-F(1,1)	-10	99	98	985	8	-15	96	95	965	31	-34	93	98	973	77
	log-F(2,2)	-22	72	57	993	4	-38	72	66	980	19	-65	69	90	953	49
	log-F(3,3)	-26	56	38	999	2	-49	55	55	965	6	-81	54	94	946	25
0.10	**CMLE**$^\phi$	-4	118	140	974	27	-27	106	119	968	42	-34	131	182	970	101
	CFL	-3	107	114	971	40	-12	108	118	973	86	-10	104	109	972	216
	log-F(1,1)	-5	94	89	973	28	-14	93	88	960	71	-13	87	78	982	220
	log-F(2,2)	-15	74	56	980	13	-33	72	62	970	47	-43	68	65	966	155
	log-F(3,3)	-21	58	38	987	3	-45	56	52	960	27	-62	53	66	960	99
0.20	**CMLE**$^\phi$	-8	137	189	968	32	-14	126	160	958	73	-23	110	126	965	142
	CFL	-4	109	118	966	45	-3	106	112	966	131	-9	98	97	969	230
	log-F(1,1)	-6	93	86	977	35	-5	94	89	952	111	-12	86	75	975	238
	log-F(2,2)	-15	75	58	986	28	-25	75	62	965	95	-41	65	60	961	164
	log-F(3,3)	-21	59	39	989	12	-38	58	48	952	59	-59	52	62	958	136

Table C.3: Simulation results for binary exposure of 10 matched case-control sets with one covariate.

Bias, SD and MSE ×100, CP and Power ×1000.

* ExpPrev = exposure prevalence.

$^\phi$ Estimate was calculated based on the datasets where CMLEs are finite out of 1000 replications. When ExpPrev = 0.05, the number of infinite datasets is 389, 467 and 544 for β_E = 0.5, 1.0 and 1.5, respectively; when ExpPrev = 0.10, the number of infinite datasets is 221, 289 and 428 for β_E = 0.5, 1.0 and 1.5, respectively; when ExpPrev = 0.20, the number of infinite datasets is 168, 270 and 375 for β_E = 0.5, 1.0 and 1.5, respectively.

† Coverage probability of two-sided nominal 95% confidence intervals based on PPCL.

‡ Relative frequency of confidence intervals for log-OR coefficient excluding 0.

ExpPrev*	**Method**	β_E = 0.5					β_E = 1.0					β_E = 1.5				
		Bias	**SD**	**MSE**	**CP**†	**Power**‡	**Bias**	**SD**	**MSE**	**CP**†	**Power**‡	**Bias**	**SD**	**MSE**	**CP**†	**Power**‡
0.05	**CMLE**$^\phi$	-16	198	396	986	0	-24	249	624	970	15	-84	197	459	960	10
	CFL	-20	165	276	989	10	-24	143	209	989	18	-53	130	198	980	24
	log-F(1,1)	-10	98	97	986	6	-14	95	92	966	38	-25	86	79	981	74
	log-F(2,2)	-27	72	60	994	1	-45	71	71	991	10	-74	67	99	981	11
	log-F(3,3)	-26	55	38	994	0	-49	55	53	966	8	-75	50	82	960	24
0.10	**CMLE**$^\phi$	3	212	448	972	23	-20	199	399	969	25	-36	164	281	964	18
	CFL	-12	132	176	985	10	-25	126	165	985	26	-45	109	138	977	51
	log-F(1,1)	-4	99	98	970	29	-12	94	91	952	80	-14	88	80	965	219
	log-F(2,2)	-19	78	65	989	10	-38	75	71	987	18	-54	71	79	979	56
	log-F(3,3)	-21	61	41	989	11	-44	58	53	952	30	-63	53	68	944	101
0.20	**CMLE**$^\phi$	-15	188	355	988	6	-1	207	429	979	21	-29	177	321	988	12
	CFL	-9	126	159	984	5	-28	116	143	976	34	-40	98	112	992	60
	log-F(1,1)	1	96	91	976	41	-5	92	84	960	120	-8	85	73	981	261
	log-F(2,2)	-14	78	63	992	8	-33	76	68	981	34	-47	70	72	990	74
	log-F(3,3)	-17	60	39	989	12	-37	57	47	960	56	-57	51	59	962	143

Table C.4: Simulation results for binary exposure of 10 matched case-control sets with five covariates.

Bias, SD and MSE ×100, CP and Power ×1000.

* ExpPrev = exposure prevalence.

$^\phi$ Estimate was calculated based on the datasets where CMLEs are finite out of 1000 replications. When ExpPrev = 0.05, the number of infinite datasets is 861, 865 and 901 for β_E = 0.5, 1.0 and 1.5, respectively; when ExpPrev = 0.10, the number of infinite datasets is 824, 837 and 888 for β_E = 0.5, 1.0 and 1.5, respectively; when ExpPrev = 0.20, the number of infinite datasets is 829, 856 and 914 for β_E = 0.5, 1.0 and 1.5, respectively.

† Coverage probability of two-sided nominal 95% confidence intervals based on PPCL.

‡ Relative frequency of confidence intervals for log-OR coefficient excluding 0.

ExpPrev[⋆]	**Method**	**β_E = 0.5**					**β_E = 1.0**					**β_E = 1.5**				
		Bias	**SD**	**MSE**	**CP**[†]	**Power**[‡]	**Bias**	**SD**	**MSE**	**CP**[†]	**Power**[‡]	**Bias**	**SD**	**MSE**	**CP**[†]	**Power**[‡]
0.05	**CMLE**[ϕ]	5	67	45	921	179	11	63	41	950	502	10	62	40	977	803
	CFL	1	61	38	932	144	5	62	39	946	469	2	65	42	964	799
	log-F(1,1)	1	62	38	932	145	5	62	39	946	469	2	65	42	964	799
	log-F(2,2)	-3	55	31	935	125	-4	54	29	956	440	-11	54	30	961	779
	log-F(3,3)	-7	51	26	940	109	-11	48	25	957	415	-21	47	27	954	759
0.10	**CMLE**[ϕ]	2	51	26	937	202	5	52	27	949	616	9	56	32	964	931
	CFL	-1	48	23	944	178	-1	48	23	957	574	0	53	28	960	920
	log-F(1,1)	-1	48	23	944	178	-1	48	23	956	574	0	53	28	960	920
	log-F(2,2)	-3	45	20	948	170	-6	44	20	965	564	-9	47	23	950	918
	log-F(3,3)	-6	42	18	952	151	-11	41	18	962	518	-16	43	21	945	907
0.20	**CMLE**[ϕ]	7	45	21	944	245	7	49	24	949	665	7	54	29	956	943
	CFL	4	44	19	958	227	1	46	21	958	632	-1	50	25	958	937
	log-F(1,1)	4	44	19	958	227	1	46	21	958	632	-1	50	25	958	937
	log-F(2,2)	1	41	17	962	225	-4	42	18	963	629	-9	45	21	953	935
	log-F(3,3)	-1	39	15	968	200	-8	40	16	965	612	-16	41	20	942	928

Table C.5: Simulation results for binary exposure of 50 matched case-control sets with no covariate.

Bias, SD and MSE ×100, CP and Power ×1000.

[⋆] ExpPrev = exposure prevalence.

[ϕ] Estimate was calculated based on the datasets where CMLEs are finite out of 1000 replications. When ExpPrev = 0.05, the number of infinite datasets is 3, 16 and 36 for β_E = 0.5, 1.0 and 1.5, respectively; when ExpPrev = 0.10, the number of infinite datasets is 1, 1 and 8 for β_E = 0.5, 1.0 and 1.5, respectively; when ExpPrev = 0.20, the number of infinite datasets is 1, 1 and 4 for β_E = 0.5, 1.0 and 1.5, respectively.

[†] Coverage probability of two-sided nominal 95% confidence intervals based on PPCL.

[‡] Relative frequency of confidence intervals for log-OR coefficient excluding 0.

ExpPrev[⋆]	**Method**	**β_E = 0.5**					**β_E = 1.0**					**β_E = 1.5**				
		Bias	**SD**	**MSE**	**CP**[†]	**Power**[‡]	**Bias**	**SD**	**MSE**	**CP**[†]	**Power**[‡]	**Bias**	**SD**	**MSE**	**CP**[†]	**Power**[‡]
0.05	**CMLE**[ϕ]	8	63	40	951	173	11	68	47	937	474	12	67	46	960	817
	CFL	3	57	33	958	157	0	60	36	948	450	3	69	47	936	802
	log-F(1,1)	2	57	32	957	154	1	60	36	947	464	2	68	46	933	813
	log-F(2,2)	-1	52	27	967	150	-7	54	30	951	431	-9	58	34	938	783
	log-F(3,3)	-5	47	22	969	129	-14	48	25	946	417	-22	50	29	930	776
0.10	**CMLE**[ϕ]	7	51	26	953	218	12	58	35	936	644	12	58	36	948	931
	CFL	3	47	22	963	196	4	52	27	951	613	1	56	32	952	919
	log-F(1,1)	3	47	22	958	197	3	51	26	948	616	1	55	30	955	925
	log-F(2,2)	1	45	20	968	186	-1	48	23	957	596	-6	50	25	953	917
	log-F(3,3)	-2	41	17	966	161	-8	44	20	955	577	-16	43	22	952	907
0.20	**CMLE**[ϕ]	5	45	21	952	230	8	51	26	949	661	15	59	37	944	949
	CFL	2	42	18	962	212	1	47	22	955	637	1	52	27	960	940
	log-F(1,1)	1	42	17	961	215	1	47	22	958	637	1	51	26	956	949
	log-F(2,2)	0	40	16	967	207	-3	44	19	961	628	-5	47	22	962	935
	log-F(3,3)	-3	38	14	970	186	-9	40	17	962	601	-14	42	20	955	934

Table C.6: Simulation results for binary exposure of 50 matched case-control sets with one covariate.

Bias, SD and MSE ×100, CP and Power ×1000.

[⋆] ExpPrev = exposure prevalence.

[ϕ] Estimate was calculated based on the datasets where CMLEs are finite out of 1000 replications. When ExpPrev = 0.05, the number of infinite datasets is 4, 5 and 42 for β_E = 0.5, 1.0 and 1.5, respectively; when ExpPrev = 0.10, the number of infinite datasets is 1, 1 and 12 for β_E = 0.5, 1.0 and 1.5, respectively; when ExpPrev = 0.20, the number of infinite datasets is 2 and 2 for β_E = 1.0 and 1.5, respectively.

[†] Coverage probability of two-sided nominal 95% confidence intervals based on PPCL.

[‡] Relative frequency of confidence intervals for log-OR coefficient excluding 0.

ExpPrev*	Method	β_E = 0.5					β_E = 1.0					β_E = 1.5				
		Bias	SD	MSE	CP†	Power‡	Bias	SD	MSE	CP†	Power‡	Bias	SD	MSE	CP†	Power‡
0.05	**CMLE$^\phi$**	12	74	56	931	183	22	72	57	949	485	30	79	71	951	803
	CFL	1	61	37	959	147	4	63	39	963	419	4	70	49	967	764
	log-F(1,1)	-1	56	31	958	142	4	60	36	958	478	2	65	42	964	813
	log-F(2,2)	0	58	33	960	135	1	57	33	966	421	0	62	38	973	768
	log-F(3,3)	-8	47	23	968	108	-12	47	24	960	408	-22	48	27	959	777
0.10	**CMLE$^\phi$**	10	59	36	936	213	20	64	46	924	604	34	72	63	924	926
	CFL	1	50	25	958	177	1	53	28	955	547	5	59	35	948	902
	log-F(1,1)	0	47	22	954	192	0	49	24	954	585	2	55	30	955	918
	log-F(2,2)	2	50	25	958	182	3	52	27	956	567	7	55	31	951	909
	log-F(3,3)	-5	41	17	961	165	-11	42	18	957	546	-16 7 43	21	951	901	
0.20	**CMLE$^\phi$**	12	56	32	936	240	18	64	44	918	650	30	72	60	919	930
	CFL	3	47	22	955	196	1	56	31	947	598	0	55	30	961	913
	log-F(1,1)	1	43	19	955	196	-1	49	24	951	629	-2	51	26	964	93
	log-F(2,2)	4	47	23	953	205	3	53	28	950	617	4	53	28	967	918
	log-F(3,3)	-4	39	15	965	178	-10	42	18	957	606	-17	42	20	948	933

Table C.7: Simulation results for binary exposure of 50 matched case-control sets with five covariates.

Bias, SD and MSE ×100, CP and Power ×1000.

* ExpPrev = exposure prevalence.

$^\phi$ Estimate was calculated based on the datasets where CMLEs are finite out of 1000 replications. When ExpPrev = 0.05, the number of infinite datasets is 14 and 36 for β_E = 1.0 and 1.5, respectively; when ExpPrev = 0.10, the number of infinite datasets is 1 and 11 for β_E = 1.0 and 1.5, respectively; when ExpPrev = 0.20, the number of infinite datasets is 4 and 4 for β_E = 1.0 and 1.5, respectively.

† Coverage probability of two-sided nominal 95% confidence intervals based on PPCL.

‡ Relative frequency of confidence intervals for log-OR coefficient excluding 0.

ExpPrev*	Method	β_E = 0.5					β_E = 1.0					β_E = 1.5				
		Bias	SD	MSE	CP†	Power‡	Bias	SD	MSE	CP†	Power‡	Bias	SD	MSE	CP†	Power‡
0.05	**CMLE$^\phi$**	3	41	17	951	269	5	45	20	947	765	6	51	27	942	980
	CFL	1	39	16	958	261	1	42	17	951	759	-1	47	22	938	979
	log-F(1,1)	1	39	16	958	261	1	42	17	951	759	-1	47	22	937	979
	log-F(2,2)	-1	38	14	962	250	-3	39	16	956	749	-8	43	19	933	978
	log-F(3,3)	-3	36	13	962	238	-7	37	14	951	741	-13	40	18	928	977
0.10	**CMLE$^\phi$**	3	34	12	940	355	4	38	14	936	887	6	38	15	951	1000
	CFL	2	33	11	945	345	1	36	13	948	884	1	36	13	958	1000
	log-F(1,1)	2	33	11	945	345	1	36	13	948	884	1	36	13	958	1000
	log-F(2,2)	0	32	10	945	336	-2	34	12	948	877	-4	34	12	962	999
	log-F(3,3)	-1	31	10	953	331	-5	33	11	950	874	-8	32	11	960	999
0.20	**CMLE$^\phi$**	3	33	11	938	426	5	34	12	944	923	4	38	15	944	1000
	CFL	2	32	10	940	424	3	33	11	955	923	0	36	13	949	1000
	log-F(1,1)	2	32	10	940	424	3	33	11	954	922	0	36	13	949	1000
	log-F(2,2)	1	31	9	950	413	0	32	10	956	917	-4	34	12	950	1000
	log-F(3,3)	0	30	9	951	413	-2	31	10	959	914	-8	33	11	948	1000

Table C.8: Simulation results for binary exposure of 100 matched case-control sets with no covariate.

Bias, SD and MSE ×100, CP and Power ×1000.

* ExpPrev = exposure prevalence.

$^\phi$ Estimate was calculated based on the datasets where CMLEs are finite out of 1000 replications. When ExpPrev = 0.05, the number of infinite datasets is 1 for β_E = 1.5.

† Coverage probability of two-sided nominal 95% confidence intervals based on PPCL.

‡ Relative frequency of confidence intervals for log-OR coefficient excluding 0.

ExpPrev*	Method	β_E = 0.5					β_E = 1.0					β_E = 1.5				
		Bias	**SD**	**MSE**	**CP†**	**Power‡**	**Bias**	**SD**	**MSE**	**CP†**	**Power‡**	**Bias**	**SD**	**MSE**	**CP†**	**Power‡**
0.05	**CMLE$^\phi$**	6	43	19	937	285	6	43	19	951	782	11	51	27	943	992
	CFL	3	41	17	946	269	0	40	16	962	763	2	46	21	959	990
	log-F(1,1)	3	41	17	946	267	0	40	16	960	768	2	46	21	956	993
	log-F(2,2)	2	39	15	950	253	-3	38	14	963	755	-3	43	18	962	989
	log-F(3,3)	-1	37	14	957	236	-7	36	13	960	741	-10	39	16	952	991
0.10	**CMLE$^\phi$**	4	34	12	950	378	4	36	13	944	894	8	43	19	937	997
	CFL	2	32	11	956	361	0	34	12	955	888	1	40	16	932	997
	log-F(1,1)	2	32	10	957	364	0	34	12	956	891	1	39	16	940	997
	log-F(2,2)	1	32	10	962	353	-2	33	11	957	887	-2	38	14	941	997
	log-F(3,3)	0	30	9	965	352	-5	32	10	956	886	-7	36	13	943	997
0.20	**CMLE$^\phi$**	4	32	10	943	425	4	34	12	952	918	8	38	15	949	999
	CFL	2	31	10	952	410	1	33	11	954	916	2	36	13	953	999
	log-F(1,1)	2	31	9	951	415	1	33	11	959	919	3	36	13	955	999
	log-F(2,2)	1	30	9	952	408	-1	32	10	959	911	0	34	12	954	999
	log-F(3,3)	-1	29	8	954	401	-4	31	10	962	908	-5	32	11	959	999

Table C.9: Simulation results for binary exposure of 100 matched case-control sets with one covariate.

Bias, SD and MSE ×100, CP and Power ×1000.

* ExpPrev = exposure prevalence.

$^\phi$ Estimate was calculated based on the datasets where CMLEs are finite out of 1000 replications.

† Coverage probability of two-sided nominal 95% confidence intervals based on PPCL.

‡ Relative frequency of confidence intervals for log-OR coefficient excluding 0.

ExpPrev*	Method	β_E = 0.5					β_E = 1.0					β_E = 1.5				
		Bias	**SD**	**MSE**	**CP†**	**Power‡**	**Bias**	**SD**	**MSE**	**CP†**	**Power‡**	**Bias**	**SD**	**MSE**	**CP†**	**Power‡**
0.05	**CMLE$^\phi$**	7	44	20	949	278	12	48	25	925	780	14	52	30	927	971
	CFL	2	40	16	959	255	3	43	19	943	750	0	46	22	941	961
	log-F(1,1)	2	39	15	963	262	2	42	18	948	775	-1	45	20	947	974
	log-F(2,2)	2	40	16	960	254	3	42	18	944	750	-1	45	20	945	963
	log-F(3,3)	-2	36	13	961	245	-6	38	15	953	756	-12	39	17	947	973
0.10	**CMLE$^\phi$**	4	36	13	950	350	10	41	18	925	885	17	47	25	920	997
	CFL	0	33	11	957	325	1	37	14	938	872	3	41	17	942	996
	log-F(1,1)	-1	32	10	952	328	1	36	13	942	891	2	39	16	949	999
	log-F(2,2)	1	34	11	955	331	3	37	14	938	875	5	41	17	942	996
	log-F(3,3)	-3	31	9	957	310	-4	33	11	949	886	-7	35	13	942	999
0.20	**CMLE$^\phi$**	6	35	13	934	410	9	39	16	930	902	16	43	21	928	997
	CFL	2	32	11	944	380	1	35	12	938	893	2	38	15	943	997
	log-F(1,1)	2	32	10	940	405	0	34	11	941	914	2	37	14	940	998
	log-F(2,2)	4	33	11	943	387	3	36	13	942	899	5	38	15	943	997
	log-F(3,3)	0	30	9	948	383	-5	31	10	942	908	-6	34	12	943	998

Table C.10: Simulation results for binary exposure of 100 matched case-control sets with five covariates.

Bias, SD and MSE ×100, CP and Power ×1000.

* ExpPrev = exposure prevalence.

$^\phi$ Estimate was calculated based on the datasets where CMLEs are finite out of 1000 replications. When ExpPrev = 0.05, the number of infinite datasets is 1 for β_E = 1.5.

† Coverage probability of two-sided nominal 95% confidence intervals based on PPCL.

‡ Relative frequency of confidence intervals for log-OR coefficient excluding 0.