Published in final edited form as:

J Stat Comput Simul. 2013 January 1; 83(7): 1191–1209. doi:10.1080/00949655.2011.654119.

Case-Control Genome-wide Joint Association Study Using Semiparametric Empirical Model and Approximate Bayes Factor

Jinfeng Xu¹, Gang Zheng², and Ao Yuan³

¹Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546

²Office of Biostatistics Research, DPPS, National Heart, Lung and Blood Institute, 6701 Rockledge Drive, Bethesda, MD 20892, USA

³National Human Genome Center, Howard University, 2216 Sixth Street N.W., Washington, DC 20059

Abstract

We propose a semiparametric approach for the analysis of case-control genome-wide association study. Parametric components are used to model both the conditional distribution of the case status given the covariates and the distribution of genotype counts, whereas the distribution of the covariates are modeled nonparametrically. This yields a direct and joint modeling of the case status, covariates and genotype counts, and gives better understanding of the disease mechanism and results in more reliable conclusions. Side information, such as the disease prevalence, can be conveniently incorporated into the model by empirical likelihood approach and leads to more efficient estimates and powerful test in the detection of disease-associated SNPs. Profiling is used to eliminate a nuisance nonparametric component, and the resulting profile empirical likelihood estimates are shown to be consistent and asymptotically normal. For the hypothesis test on disease association, we apply the approximate Bayes factor (ABF) which is computationally simple and most desirable in genome-wide association studies where hundreds of thousands to a million genetic markers are tested. We treat the approximate Bayes factor as a hybrid Bayes factor which replaces the full data by the maximum likelihood estimates of the parameters of interest in the full model and derive it under a general setting. The deviation from Hardy-Weinberg Equilibrium (HWE) is also taken into account and the ABF for HWE using cases is shown to provide evidence of association between a disease and a genetic marker. Simulation studies and an application are further provided to illustrate the utility of the proposed methodology.

Keywords

Approximate Bayes factor; association study; empirical likelihood; genetic model; Hardy-Weinberg Equilibrium; profile likelihood; robustness; side information

1 Introduction

In genome-wide case-control genetic association studies, the goal is to search across the genome to identify genetic loci which are associated with a disease phenotype. For the analysis of such data, various statistical methods have been proposed in the literature, adopting either a Bayesian or frequentist's approach. The former can easily incorporate prior knowledge into the inference while the latter is usually simpler in modeling and computation, which uses nonparametric models (as in Tzeng et al., 2003; Schaid et al., 2005; Yuan et al., 2007), or parametric models (as in Prentice and Pyke, 1979; Song and Elston, 2006). The nonparametric method is computationally simple and robust but has lower

efficiency and power. When correctly specified, the parametric method can achieve higher efficiency, facilitate the covariance analysis, and lead to easy interpretation. In a case-control genetic association study, association tests locate the genetic loci which are associated with the disease, while a regression analysis on the genotypes further identifies which genotype at this locus is potentially the cause of the disease. In addition to the genotypes, often other covariates are available which can adjust for the confounding factors and give a more accurate regression analysis. Hence we present a joint semiparametric modeling of the case status, covariates and genotypes to enhance our understanding of the disease mechanism and yield more reliable results. In our study the response variable is the case status which is binary, we adopt the standard logistic parametric regression model on it conditioning on the covariates. Also, as in many observational studies (as in Fan and Gijbels (1994) and Efromovich (1996)), we treat the covariates to be random. As the vector of the covariates is multi-dimensional, usually containing both discrete and continuous variables, it is more suitable to model its distribution nonparametrically since any given multivariate parametric distribution hardly fit them. Another advantage of such a nonparametric approach is that it can easily incorporate side information, using methods such as the empirical likelihood (EL). The nonparametric approach makes the modeling robust and the incorporation of side information improves the efficiency. The joint model we propose thus consists of the parametric logistic regression component which relates the case status to the covariates and the nonparametric component for the distribution of the covariates. To estimate the parametric component in the joint model, we profile the nonparametric part to get the maximum profile empirical likelihood estimates of the parameters. The association test is conducted via Bayesian hypothesis testing for each marker locus. The Bayes factor computes the posterior odds in favor of the null and is widely used in Bayesian testing. But apart from a few simple cases, the Bayes factor is computationally demanding, and numerically complicated methods such as the Laplace asymptotic approximation and Markov chain Monte Carlo simulation are often used. These methods can be computationally prohibitive in a genome-wide association study in which several million markers need to be investigated. On the other hand, the approximate Bayes factor (ABF) (Wakefield, 2007) has a very simple closed form. This makes it a valuable tool in computationally intensive applications such as genome-wide association studies. In this article, we combine these methods, a semiparametric model for the response and covariate, in which a logistic regression model for the response given the covariates, an empirical likelihood for the distribution of the covariates, the profile maximum likelihood estimates for the parameters, and then the test based on the approximate Bayes factor after the estimates are obtained. The Hardy-weinberg disequilibrium in the cases is also incorporated into the model to increase the power of the association test. Our method is applicable to both prospective and retrospective studies. Simulation studies are conducted to evaluate our methods, and as an illustration, application to a real data is performed for a genome-wide association analysis. The rest of article is organized as follows. In Section 2, we describe our model and estimation procedure. The ABF is outlined for the association study. Numerical results are presented in Section 3. Section 4 concludes with a general discussion.

2 The method

In genome-wide case-control association studies, each genetic locus, in a large set of selected marker loci across the whole genome, is examined for association with a phenotype, in the presence of extra covariates. The association test can be conducted by the approximate Bayes factor at each locus. The ABF can be calculated easily after the parameters are estimated by the profile empirical likelihood method. Furthermore, to increase the power of the association test, the information inferred from the Hardy-weinberg disequilibrium in the cases can be conveniently incorporated into the ABF. For ease of exposition, we describe our method for a given locus. We first introduce some notations.

The Data—The observations consist of $(y_i, \mathbf{x}_i, g_i)$, $(i = 1, \dots, n)$ from n independent individuals, where y_i is the case (disease) status indicator, $y_i = 1(0)$ for case (control), $\mathbf{x}_i = (x_{i1}, \dots, x_{id})'$ is the vector of covariates, and g_i is the genotype, for the i th individual. We assume a diallelic marker with two alleles denoted by A and B , and three genotypes by $(G_0, G_1, G_2) = (AA, AB, BB)$. Let the genotype counts in cases be $\mathbf{r} = (r_0, r_1, r_2)$, those in controls be $\mathbf{s} = (s_0, s_1, s_2)$, $r = r_0 + r_1 + r_2$, $s = s_0 + s_1 + s_2 = n - r$, and the total genotype counts G_i is $n_i = r_i + s_i$, $(i = 0, 1, 2)$. Denote $\mathbf{y} = (y_1, \dots, y_n)$, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, and $\mathbf{g} = (g_1, \dots, g_n)'$. For ease of exposition we also classify the covariates according to genotypes, as $\mathbf{X} = \{(x_{ij1}, \dots, x_{ijd})' : i = 0, 1, 2; j = 1, \dots, n_i\}$, where $(x_{ij1}, \dots, x_{ijd})'$ is a covariate vector for the genotype G_i . In our study (n, r, s) are treated as fixed constants.

The Model—We consider a joint model for the data $(\mathbf{y}, \mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s})$, which as shown below, can be decomposed into a regression part and a genetic part. The regression part describes $\mathbf{y} | (\mathbf{X}, \mathbf{g}, \mathbf{r})$ through the parameters $(\boldsymbol{\alpha}, \beta)$, and the genetic part models the counts $(\mathbf{r}, \mathbf{s})$ which involves the genetic parameters $(\mathbf{p}, \mathbf{q}, f)$ to be specified. Let $h(\cdot, \cdot)$ be the probability density/mass function of $(\mathbf{X}, \mathbf{g})$. Given the parameters, the joint likelihood takes the form

$$P(\mathbf{y}, \mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s} | \boldsymbol{\alpha}, \beta, \mathbf{p}, \mathbf{q}, h, f) = P(\mathbf{y} | \mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta, \mathbf{p}, \mathbf{q}, f) h(\mathbf{X}, \mathbf{g} | \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta, \mathbf{p}, \mathbf{q}, f) \times P(\mathbf{r}, \mathbf{s} | \boldsymbol{\alpha}, \beta, \mathbf{p}, \mathbf{q}, f) = [P(\mathbf{y} | \mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta) h(\mathbf{X}, \mathbf{g})] \times P(\mathbf{r}, \mathbf{s} | \mathbf{p}, \mathbf{q}, f).$$

Note that $P(\mathbf{y} | \mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta, \mathbf{p}, \mathbf{q}, f)$ does not depend on $(\mathbf{p}, \mathbf{q}, h, f)$, $P(\mathbf{r}, \mathbf{s} | \boldsymbol{\alpha}, \beta, \mathbf{p}, \mathbf{q}, f)$ only depends on $(\mathbf{p}, \mathbf{q}, f)$. We specify $h(\mathbf{X}, \mathbf{g} | \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta, \mathbf{p}, \mathbf{q}, f)$ nonparametrically. Different from many existing methods, we allow side information to be incorporated into the joint likelihood by postulating certain constraints on $h(\mathbf{X}, \mathbf{g})$. More detail will be given in Section 2.2. We see that the likelihood is decomposed into two parts, $P(\mathbf{y} | \mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta) h(\mathbf{X}, \mathbf{g})$ for the regression part, and $P(\mathbf{r}, \mathbf{s} | \mathbf{p}, \mathbf{q}, f)$ for the genetic part. In the following, we will describe the semiparametric joint model and inference in detail. We will first address the parametric regression part, the estimation of the parameters. Then we will move on to the nonparametric part. Finally we will describe the hypothesis testing using ABF and the gene frequency model.

2.1 The Semiparametric Model and Parameter Estimation

A genetic regression model depends on the disease genetic mode, which can be either recessive (REC), dominance (DOM) or additive (ADD). We first consider the case where the genetic mode is known. As in a typical case-control study, given the covariates for the conditional density of $\mathbf{y}$, we fit a prospective logistic model for $P(\mathbf{y} | \mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta)$ as in Prentice and Pyke (1979),

$$P(\mathbf{y} | \mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta) = \frac{\exp[\sum_{i=0}^2 \sum_{j=1}^{n_i} \boldsymbol{\alpha}' \mathbf{x}_{ij} + \beta \sum_{i=0}^2 r_i c(G_i)]}{\prod_{i=0}^2 \prod_{j=1}^{n_i} [1 + \exp\{\boldsymbol{\alpha}' \mathbf{x}_{ij} + \beta c(G_i)\}]}, \quad (1)$$

where $c(G_i)$ is a genotype code depending on the underlying genetic mode. We choose $c(G_0) = 0$, $c(G_1) = t_1$, and $c(G_2) = t_2$, where $t_2 - t_1 \geq 0$ and $t_2 > 0$. Under REC, DOM and ADD modes, $t_1 = 0$ and $t_2 = 1$, $t_1 = t_2 = 1$, and $t_1 = 1$ and $t_2 = 2$ are chosen respectively. For

robustness, we specify $h(\mathbf{X}, \mathbf{g}) = \prod_{i=1}^n h(\mathbf{x}_i, g_i) = \prod_{i=1}^n dH(\mathbf{x}_i, g_i)$ nonparametrically. Compared to other nonparametric methods, the empirical likelihood (for example, Owen, 1988; Qin and Lawless, 1994) is easy to use, which assigns a non-negative weight at each observation as

$$dH(\mathbf{x}_i, g_i) = w_i, \quad \text{subject to} \quad w_i \geq 0, \quad \sum_{i=1}^n w_i = 1. \quad (2)$$

The advantages of using a nonparametric $h(\mathbf{X}, \mathbf{g})$ in the analysis is for robustness and incorporating side information into the estimation to improve the efficiency. Such information can be implemented using the constraints through $h(\mathbf{X}, \mathbf{g})$ (Qin and Lawless (1994)). For our case, we consider one such application below, other forms of information, when available, can also be implemented in a similar way. Let $\rho_0 = P(y_i = 1)$ be the disease prevalence rate among the selected population, which is sometimes available to us, and is the side information we intend to incorporate into the estimation. By (1) and (2), we have

$$\rho_0 = \int P(y_i = 1 | \mathbf{x}_i, g_i; \boldsymbol{\alpha}, \beta) dH(\mathbf{x}_i, g_i) = \int \frac{\exp[\boldsymbol{\alpha}' \mathbf{x}_i + \beta c(g_i)]}{1 + \exp[\boldsymbol{\alpha}' \mathbf{x}_i + \beta c(g_i)]} h(\mathbf{x}_i, g_i) d\mathbf{x}_i dg_i.$$

Here and throughout, $d\mathbf{x}$ means differential for continuous components of $\mathbf{x}$ and difference for discrete components. We see that the nonparametric part $h(\mathbf{x}_i, g_i)$ and $(\boldsymbol{\alpha}, \beta)$ are implicitly connected through the above relationship, and $h(\cdot, \cdot)$ can not be simply left out in the estimation of the parameter $(\boldsymbol{\alpha}, \beta)$, as it provides auxiliary information about the parameters. Similar works of inferring parameters using empirical likelihood in the presence of nonparametric nuisance distribution with side information can be found in Leung and Qin (2006) and Tan (2007). Here, $h(\cdot, \cdot)$ or equivalently $\{w_i; i = 1, \dots, n\}$ is a set of nuisance parameters in our joint likelihood. Write $h(\mathbf{X}, \mathbf{g}) = \prod_{i=1}^n h(\mathbf{x}_i, g_i)$, $\mathbf{w} = (w_1, \dots, w_n)$, and

$$\mathcal{W} = \left\{ \mathbf{w} : w_i \geq 0, \quad \sum_{i=1}^n w_i = 1, \quad \sum_{i=1}^n w_i \left(\frac{\exp[\boldsymbol{\alpha}' \mathbf{x}_i + \beta c(g_i)]}{1 + \exp[\boldsymbol{\alpha}' \mathbf{x}_i + \beta c(g_i)]} - \rho_0 \right) = 0 \right\}.$$

A typical technique to eliminate these nuisance parameters is through profiling, i.e., for fixed parameter values $(\boldsymbol{\alpha}, \beta)$, maximize $h(\cdot, \cdot)$ in the joint likelihood to get the profile empirical likelihood, under the constraints given in $\mathcal{W}$, as

$$\begin{aligned} \tilde{P}(\mathbf{y}, \mathbf{X} | \mathbf{g}, \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta, \hat{h}_n(\boldsymbol{\alpha}, \beta)) &= P(\mathbf{y} | \mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta) \hat{h}_n(\mathbf{X}, \mathbf{g} | \boldsymbol{\alpha}, \beta) \\ &\propto P(\mathbf{y} | \mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta) \prod_{i=1}^n \hat{w}_i(\boldsymbol{\alpha}, \beta) := \max_{\mathbf{w} \in \mathcal{W}} [P(\mathbf{y} | \mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta) \prod_{i=1}^n w_i] \end{aligned}$$

and inference for $(\boldsymbol{\alpha}, \beta)$ is then based on the profile likelihood $\tilde{P}(\mathbf{y}, \mathbf{X} | \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta, \hat{h}_n(\boldsymbol{\alpha}, \beta))$. In the above the side information of ρ_0 is implemented into the joint likelihood via $h(\mathbf{X}, \mathbf{g})$ through the constraints in $\mathcal{W}$.

Denote $(\hat{\boldsymbol{\alpha}}_n, \hat{\beta}_n)$ as the maximum profile empirical likelihood estimator of $(\boldsymbol{\alpha}, \beta)$, i.e.

$$(\hat{\boldsymbol{\alpha}}_n, \hat{\beta}_n) = \arg \sup_{(\boldsymbol{\alpha}, \beta)} \tilde{P}(\mathbf{y}, \mathbf{X} | \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta, \hat{h}_n(\boldsymbol{\alpha}, \beta)).$$

The profile empirical likelihood is a special case of the profile likelihood, in which the nuisance parameter is the unknown distribution h . Let (α_0, β_0, h_0) be the true parameters generating the data under the specified model. It is known (Severini and Wong, 1992) that if there is a *least favorable curve* $h(\cdot, \cdot | \alpha, \beta)$ such that $P(\mathbf{y}, \mathbf{X} | \mathbf{g}, \mathbf{r}, \mathbf{s}; \alpha_0, \beta_0, h(\alpha_0, \beta_0)) = P(\mathbf{y}, \mathbf{X} | \mathbf{g}, \mathbf{r}, \mathbf{s}; \alpha_0, \beta_0, h_0)$, then under mild regularity conditions (as those in Proposition 2, Severini and Wong, 1992) $(\hat{\alpha}_n, \hat{\beta}_n)$ is consistent for (α_0, β_0) and asymptotical normal. The curve $h(\cdot, \cdot | \alpha, \beta)$ is least favorable in the sense the corresponding likelihood P has the least Fisher information at (α_0, β_0) , but it is the only curve along which estimates of (α_0, β_0) is unbiased, and for this reason we may also call it the *most favorable curve*. More basic properties of maximum profile estimators can be found, for example, in Murphy and Van der Vaart(2000). After evaluations (Appendix) we get

$$\hat{w}_i = \hat{w}_i(\alpha, \beta, \lambda) = n^{-1} \left(1 - \lambda \rho_0 + \frac{\lambda \exp[\alpha' \mathbf{x}_i + \beta c(g_i)]}{1 + \exp[\alpha' \mathbf{x}_i + \beta c(g_i)]} \right)^{-1}, \quad (3)$$

with λ determined by the equation

$$\sum_{i=1}^n \hat{w}_i(\alpha, \beta, \lambda) \frac{\exp[\alpha' \mathbf{x}_i + \beta c(g_i)]}{1 + \exp[\alpha' \mathbf{x}_i + \beta c(g_i)]} = \rho_0, \quad \text{or} \quad \sum_{i=1}^n \frac{\hat{w}_i(\alpha, \beta, \lambda)}{1 + \exp[\alpha' \mathbf{x}_i + \beta c(g_i)]} = 1 - \rho_0.$$

Plugging in the expression for $\hat{w}_i$ in (3) into the second equation above we get

$$\sum_{i=1}^n \frac{1}{1 - \lambda \rho_0 + (1 + \lambda(1 - \rho_0)) \exp[\alpha' \mathbf{x}_i + \beta c(g_i)]} = n(1 - \rho_0). \quad (4)$$

Rearrange the components of $\hat{\mathbf{w}}$ as $\{\hat{w}_{ij}\}$ according the order of the $\mathbf{x}_{ij}$'s, and let $L(\alpha, \beta, h(\alpha, \beta)) = \log P(\mathbf{y}, \mathbf{X} | \mathbf{g}, \mathbf{r}, \mathbf{s}; \alpha, \beta, h(\alpha, \beta))$. Note that using the $\hat{w}_i$'s, a density/mass function $\hat{h}_n = \hat{h}_n(\alpha, \beta)$ for $(\mathbf{x}_i, g_i)$ can be constructed such that, up to a quantity independent of (α, β) ,

$$\begin{aligned} \tilde{L}(\alpha, \beta, \hat{h}_n(\alpha, \beta)) &= \sum_{i=1}^n \tilde{l}(y_i, \mathbf{x}_i, g_i | \alpha, \beta, \hat{h}_n(\alpha, \beta)) \\ &:= \sum_{i=1}^n \left(y_i [\alpha' \mathbf{x}_i + \beta c(g_i)] - \log[1 - \lambda \rho_0 + (1 + \lambda(1 - \rho_0)) \exp[\alpha' \mathbf{x}_i + \beta c(g_i)]] \right) \\ &= \sum_{i=0}^2 \sum_{j=1}^{r_i} \alpha' \mathbf{x}_{ij} + \beta \sum_{i=0}^2 r_i c(G_i) - \sum_{i=0}^2 \sum_{j=1}^{n_i} \log[1 - \lambda \rho_0 + (1 + \lambda(1 - \rho_0)) \exp[\alpha' \mathbf{x}_{ij} + \beta c(G_i)]], \end{aligned} \quad (5)$$

with $\lambda = \lambda(\alpha, \beta)$ determined by (4), and $(\hat{\alpha}_n, \hat{\beta}_n)$ is the maximizer of the above profile empirical likelihood (5). In the above, the expression for $\tilde{l}(y_i, \mathbf{x}_i, g_i | \alpha, \beta, \hat{h}_n(\alpha, \beta))$ is up to a quantity independent of (α, β) . Here we treat $(\hat{\alpha}_n, \hat{\beta}_n)$ as a column vector. Let $\tilde{l}^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha, \beta, h) = \tilde{l}(y_i, \mathbf{x}_i, g_i | \alpha, \beta, h) / (\alpha, \beta)$, $\tilde{\mathbf{I}}^{(1)} = E[\tilde{l}^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, h(\alpha_0, \beta_0))] \times (\tilde{l}^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, h(\alpha_0, \beta_0)))'$; $\tilde{l}^{(2)}(y_i, \mathbf{x}_i, g_i | \alpha, \beta, h) = \tilde{l}(y_i, \mathbf{x}_i, g_i | \alpha, \beta, h) / [(\alpha, \beta) (\alpha, \beta)']$, and $\tilde{\mathbf{I}}^{(2)} = -E[\tilde{l}^{(2)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, h(\alpha_0, \beta_0))]$, where the expectations are under the true joint distribution. Denote $\xrightarrow{D}$ for convergence in distribution, we have

Theorem 1—Assume the least favorable curve $h(\cdot, \cdot)$ exists, and the Hessian matrix $\tilde{\mathbf{I}}^{(2)}$ is non-singular in a neighborhood of (α_0, β_0) . Then as $n \rightarrow \infty$, we have $(\hat{\alpha}_n, \hat{\beta}_n) \xrightarrow{a.s.} (\alpha_0, \beta_0)$, and

$$\sqrt{n}(\hat{\alpha}_n - \alpha_0, \hat{\beta}_n - \beta_0) \xrightarrow{D} N(\mathbf{0}, \Omega),$$

where $\Omega = (\tilde{\mathbf{I}}^{(2)})^{-1} \tilde{\mathbf{I}}^{(1)} (\tilde{\mathbf{I}}^{(2)})^{-1}$, here $(\hat{\alpha}_n - \alpha_0, \hat{\beta}_n - \beta_0)$ is treated as a column vector.

Remark—The existence of the least favorable curve is a very mild condition, which holds when there is a $h(\cdot|\alpha, \beta)$ such that $h(\cdot|\alpha_0, \beta_0) = h_0(\cdot)$. This is more relaxed than the usual assumption that some prespecified distribution $h^*(\cdot|\alpha, \beta)$ satisfies $h^*(\cdot|\alpha_0, \beta_0) = h_0(\cdot)$.

Note Ω is consistently estimated by $\Omega_n = (\tilde{\mathbf{I}}_n^{(2)})^{-1} \tilde{\mathbf{I}}_n^{(1)} (\tilde{\mathbf{I}}_n^{(2)})^{-1}$, where

$$\begin{aligned} \tilde{\mathbf{I}}_n^{(1)} &= \frac{1}{n} \sum_{i=1}^n \tilde{l}^{(1)}(y_i, \mathbf{x}_i, g_i | \hat{\alpha}_n, \hat{\beta}_n, \hat{h}_n(\hat{\alpha}_n, \hat{\beta}_n)) \left(\tilde{l}^{(1)}(y_i, \mathbf{x}_i, g_i | \hat{\alpha}_n, \hat{\beta}_n, \hat{h}_n(\hat{\alpha}_n, \hat{\beta}_n)) \right)' \\ &= \frac{1}{n} \sum_{i=1}^n \left(y_i - \frac{(1+\lambda(1-\rho_0)) \exp[\hat{\alpha}_n' \mathbf{x}_i + \hat{\beta}_n c(g_i)]}{1-\lambda\rho_0 + (1+\lambda(1-\rho_0)) \exp[\hat{\alpha}_n' \mathbf{x}_i + \hat{\beta}_n c(g_i)]} \right)^2 \begin{pmatrix} \mathbf{x}_i \mathbf{x}_i' & c(g_i) \mathbf{x}_i' \\ c(g_i) \mathbf{x}_i & c^2(g_i) \end{pmatrix}, \end{aligned}$$

and

$$\begin{aligned} \tilde{\mathbf{I}}_n^{(2)} &= \frac{1}{n} \sum_{i=1}^n \tilde{l}^{(2)}(y_i, \mathbf{x}_i, g_i | \hat{\alpha}_n, \hat{\beta}_n, \hat{h}_n(\hat{\alpha}_n, \hat{\beta}_n)) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{(1-\lambda\rho_0)(1+\lambda(1-\rho_0)) \exp[\hat{\alpha}_n' \mathbf{x}_i + \hat{\beta}_n c(g_i)]}{(1-\lambda\rho_0 + (1+\lambda(1-\rho_0)) \exp[\hat{\alpha}_n' \mathbf{x}_i + \hat{\beta}_n c(g_i)])^2} \begin{pmatrix} \mathbf{x}_i \mathbf{x}_i' & c(g_i) \mathbf{x}_i' \\ c(g_i) \mathbf{x}_i & c^2(g_i) \end{pmatrix} \end{aligned}$$

and λ is determined by (4) with (α, β) replaced by $(\hat{\alpha}_n, \hat{\beta}_n)$. The expressions for $\tilde{l}^{(1)}(y_i, \mathbf{x}_i, g_i | \hat{\alpha}_n, \hat{\beta}_n, \hat{h}_n(\hat{\alpha}_n, \hat{\beta}_n))$ and $\tilde{l}^{(2)}(y_i, \mathbf{x}_i, g_i | \hat{\alpha}_n, \hat{\beta}_n, \hat{h}_n(\hat{\alpha}_n, \hat{\beta}_n))$ are given in the Appendix.

In our model, $h(\mathbf{X}, g)$ handles any dependence among $(\mathbf{X}, g)$ automatically via the data and the side information constraint, while in many parametric models, such as in Chatterjee and Carroll (2005) and Dai et al (2009), additional dependence parameters need to be specified to account for this.

In practice, ρ_0 is easily estimated from the data, but it is possible with some bias, i.e., $\rho_0 = \rho + O(1/n)$, where ρ is the true disease prevalence rate in the selected population. To investigate such bias in the effect of estimation, let $L_n(\alpha, \beta, \hat{h}_n(\alpha, \beta), \rho_0) = L(\alpha, \beta, \hat{h}_n(\alpha, \beta)) / n$ be the averaged profile semiparametric empirical likelihood, here we spelled out ρ_0 to emphasize its dependence this parameter. Then $(\hat{\alpha}_n, \hat{\beta}_n) = \arg \max_{(\alpha, \beta)} L_n(\alpha, \beta, \hat{h}_n(\alpha, \beta), \rho_0)$. It is easy to see that

$$\begin{aligned} \tilde{L}_n(\alpha, \beta, \hat{h}_n(\alpha, \beta), \rho_0) &= \tilde{L}_n(\alpha, \beta, \hat{h}_n(\alpha, \beta), \bar{\rho}) \\ &+ O\left(\frac{1}{n}\right) \frac{1}{n} \sum_{i=0}^2 \sum_{j=1}^{n_i} \frac{\lambda(1+\exp[\alpha' \mathbf{x}_{ij} + \beta c(G_i)])}{1-\lambda\bar{\rho} + (1+\lambda(1-\bar{\rho})) \exp[\alpha' \mathbf{x}_{ij} + \beta c(G_i)]}. \end{aligned}$$

Asymptotically $L_n(\alpha, \beta, \hat{h}_n(\alpha, \beta), \rho_0)$ and $L_n(\alpha, \beta, \hat{h}_n(\alpha, \beta), \rho)$ are the entropy valued at ρ_0 and ρ respectively. As the entropy is a continuous function of ρ , we see that the effect of a

bias $O(1/n)$ on $\hat{\rho}$ induces a bias of order $O(1/n)$ on the parameter estimation and the corresponding variance estimations.

Next we introduce the ABF for the association test.

2.2 Approximate Bayes factor

In Bayesian hypothesis testing, the Bayes factor (BF) is widely used, which is the posterior odds against the null, defined as

$$BF = \frac{\int_{\Theta_0} Pr(data|\theta, H_0) Pr(\theta|H_0) d\theta}{\int_{\Theta_1} Pr(data|\theta, H_1) Pr(\theta|H_1) d\theta}.$$

As an evidence against H_0 , Jeffreys (1935) gave a scale for interpretation of the values of BF: 1–1/3, barely worth mentioning; 1/3–1/10, substantial; 1/10–1/30, strong; 1/30–1/100, very strong; < 1/100, decisive.

Apart from some special cases, the BF has no closed form and can be computationally very demanding. Various numerical methods are used to approximate it, such as the Laplace approximation and the Markov chain Monte Carlo methods, and a comprehensive review can be found in Kass and Raftery (1995). These methods again are computationally intensive and even prohibitive in studies which involve large scale computations such as in genome-wide studies in which millions of markers are to be investigated. The approximate Bayes factor (Wakefield, 2007) is to evaluate the BF from a different view, and has a simple closed form, which makes it a valuable tool in our situation. The ABF can be viewed as a hybrid BF which replaces the likelihood of data, $Pr(data|\theta)$, with the asymptotic distribution of the MLEs of the parameters of interest in the full model as

$$ABF = \frac{Pr(\hat{\theta}_n|H_0)}{Pr(\hat{\theta}_n|H_1)}.$$

Thus, the ABF measures evidence in the data summary $\hat{\theta}_n$ rather than the full data. Since the MLEs are asymptotically normally distributed, without higher order terms, $\hat{\theta}_n|\theta \sim N(\theta, \mathbf{V}_n)$, where $\mathbf{V}_n$ is the asymptotic covariance matrix for $\hat{\theta}_n$ and n is the sample size. In practice, $\mathbf{V}_n$ can be approximated by the inverse of the observed Fisher information matrix with plug-in estimates. Normal priors are often used, i.e. $\pi(\theta) \sim N(\mathbf{0}, \mathbf{W})$ for known $\mathbf{W}$. By simple algebraic calculation, we have

$$ABF = \frac{|\mathbf{V}_n + \mathbf{W}|^{1/2}}{|\mathbf{V}_n|^{1/2}} \exp \left[-\frac{1}{2} \hat{\theta}_n' \left\{ \mathbf{V}_n^{-1} - (\mathbf{V}_n + \mathbf{W})^{-1} \right\} \hat{\theta}_n \right].$$

When testing the null $H_0: \beta = \beta_0$, it is worthy to note that as in Wakefield (2009a), the ABF can be alternatively obtained by replacing the likelihood of data, $Pr(data|\alpha, \beta)$, with the asymptotic distribution of the MLEs of the parameters in the full model. Thus,

$$ABF = \frac{\int \Pr(\hat{\alpha}_n, \hat{\beta}_n | \alpha, \beta_0) \pi(\alpha) d\alpha}{\int \int \Pr(\hat{\alpha}_n, \hat{\beta}_n | \alpha, \beta) \pi(\alpha, \beta) d\alpha d\beta}.$$

Let $\begin{pmatrix} V_{n,11} & V_{n,12} \\ V_{n,21} & V_{n,22} \end{pmatrix}$ be the asymptotic covariance matrix of $(\hat{\alpha}_n, \hat{\beta}_n)$. Reparameterize (α, β) to (α, γ) where

$$\gamma = \beta - V_{n,21} V_{n,22}^{-1} \alpha,$$

and we have

$$\begin{bmatrix} \hat{\alpha}_n \\ \hat{\gamma}_n \end{bmatrix} \sim N \left(\begin{bmatrix} \alpha \\ \gamma \end{bmatrix}, \begin{bmatrix} V_{n,11} & \mathbf{0} \\ \mathbf{0}' & V_{n,22} - V_{n,21} V_{n,22}^{-1} V_{n,12} \end{bmatrix} \right)$$

and the previous ABF becomes

$$ABF = \frac{\int \Pr(\hat{\alpha}_n, \hat{\beta}_n | \alpha, \gamma_0) \pi(\alpha) d\alpha}{\int \int \Pr(\hat{\alpha}_n, \hat{\gamma}_n | \alpha, \gamma) \pi(\alpha, \gamma) d\alpha d\gamma},$$

where $\Pr(\hat{\alpha}_n, \hat{\gamma}_n | \alpha, \gamma)$ is the asymptotic distribution of $(\hat{\alpha}_n, \hat{\gamma}_n)$. Let $\pi(\alpha, \gamma) = \pi(\alpha) \pi(\gamma)$ and $\pi(\alpha) \sim N(\mathbf{0}, \mathbf{W}_n)$. Then the resulting ABF is identical to the previous one. It follows that the Bayesian false discovery probability (BFDP) $\pi_1 = \Pr(H_0 | \hat{\alpha}_n) = PO \times ABF / (1 + PO \times ABF)$, where $PO = \pi_0 / (1 - \pi_0)$ is the posterior odds. If claiming a false non-discovery is four times as costly as a false discovery, the Bayesian “power” used by Wakefield (2007) is defined as $\Pr(\Pr(H_0 | \hat{\beta}_n) \leq 0.8 | \hat{\beta}_n)$. The posterior odds of H_0 given $\hat{\theta}_n$ is $PO \times ABF$. Since ABF can be justified as a hybrid BF replacing data with the MLEs, it enjoys interpretations similar to BF. Wakefield (2007) also showed high concordance between BF and ABF for genetic association studies using case-control samples.

2.3 Testing the hypothesis of no association via the regression model

The null hypothesis under this model is $H_{0,1}: \beta = 0$ and the alternative is $H_{1,1}: \beta \neq 0$. Denote $\theta = (\alpha, \beta)$ and $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n)$. By Theorem 1,

$$\hat{\theta}_n \sim N(\theta, \Omega/n).$$

The asymptotic variance for $\hat{\beta}_n$ is $V_n = \omega_{22}/n$, where ω_{22} is the last entry in the matrix Ω . Using the prior $\beta \sim N(0, W_\beta)$ under $H_{1,1}$, we obtain the ABF for association with covariates (Wakefield, 2009a)

$$ABF(\hat{\beta}) = \sqrt{\frac{W_{\beta} + V_n}{V_n}} \exp\left(-\frac{\hat{\beta}^2}{2} \frac{W_{\beta}}{V_n(W_{\beta} + V_n)}\right), \quad (6)$$

where $\hat{\beta}^2/V_n$ is Wald statistic for the genetic effect and W_{β} is pre-specified. In the above, the ABF only depends on the estimate of log odds ratio, its asymptotic variance and the prior distribution for the log odds ratio. It does not directly depend on the estimates and prior distributions of any nuisance parameters. Choosing W_{β} in the prior distribution for the log odds ratio has been discussed by Wakefield (2007) Wakefield (2009a) and WTCCC (2007). It can be chosen as a function of the log odds ratio, or the minor allele frequency, or a test statistic. If we assume that the prior probability of genetic effect of at least β_0 is 5%, W_{β} can be obtained as $W_{\beta} = \{\log(\beta_0)/1.96\}^2$ (Wakefield, 2009a). The larger W_{β} the higher upper bound for the genotype relative risk. For $\beta_0 = 1.2, 1.5$, and 2.0 , $W_{\beta} \approx 0.093^2, 0.21^2$, and 0.35^2 , respectively.

2.4 Unknown genetic mode

For many diseases, the genetic mode is unknown. We arbitrarily code the modes as (M_0, M_1, M_2) : (REC, DOM, ADD). For many diseases, the genetic mode is unique for all individuals. Denote $P(\mathbf{y}|\mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta, M_j)$ as $P(\mathbf{y}|\mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s}; \boldsymbol{\alpha}, \beta)$ in (1) with the specification of the $c(\cdot)$'s under genetic mode M_j denoted as $c_j(\cdot)$, the corresponding profile empirical likelihood (5) is now $L(\boldsymbol{\alpha}, \beta, \hat{h}_n(\boldsymbol{\alpha}, \beta)|M_j)$. Denote $(\hat{\boldsymbol{\alpha}}_{nj}, \hat{\beta}_{nj})$ be the MLE of $(\boldsymbol{\alpha}_0, \beta_0)$ under $L(\boldsymbol{\alpha}, \beta, \hat{h}_n(\boldsymbol{\alpha}, \beta)|M_j)$, then genetic mode M_j is chosen for the data if

$$\tilde{L}(\hat{\boldsymbol{\alpha}}_{nj}, \hat{\beta}_{nj}, \hat{h}_n(\hat{\boldsymbol{\alpha}}_{nj}, \hat{\beta}_{nj})|M_j) = \max_l \tilde{L}(\hat{\boldsymbol{\alpha}}_{nl}, \hat{\beta}_{nl}, \hat{h}_n(\hat{\boldsymbol{\alpha}}_{nl}, \hat{\beta}_{nl})|M_l).$$

For some other diseases, the genetic mode is not unique, it can vary among different groups of individuals. In this case let $\delta_j = P(M_j)$ ($j = 0, 1, 2$), $\boldsymbol{\delta} = (\delta_0, \delta_1, \delta_2)$, they are unknown parameters. Then the regression model is a mixture

$$P(\mathbf{y}|\mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s}; \boldsymbol{\theta}) h(\mathbf{X}, \mathbf{g}) = \prod_{i=1}^n \left(\sum_{j=0}^2 \delta_j \frac{(\exp[\boldsymbol{\alpha}'_j \mathbf{x}_i + \beta_j c_j(g_i)])^{y_i}}{1 + \exp[\boldsymbol{\alpha}'_j \mathbf{x}_i + \beta_j c_j(g_i)]} \right) h(\mathbf{x}_i, g_i).$$

The population rate in this case is

$$\rho_0 = \sum_{j=0}^2 \int \frac{\delta_j \exp[\boldsymbol{\alpha}'_j \mathbf{x}_i + \beta_j c_j(g_i)]}{1 + \exp[\boldsymbol{\alpha}'_j \mathbf{x}_i + \beta_j c_j(g_i)]} h(\mathbf{x}_i, g_i) d\mathbf{x}_i dg_i$$

the set $\mathcal{W}$ is now

$$\mathcal{W} = \left\{ \mathbf{w}: w_i \geq 0, \quad \sum_{i=1}^n w_i = 1, \quad \sum_{i=1}^n \sum_{j=0}^2 w_i \delta_j \left(\frac{\exp[\boldsymbol{\alpha}'_j \mathbf{x}_i + \beta_j c_j(g_i)]}{1 + \exp[\boldsymbol{\alpha}'_j \mathbf{x}_i + \beta_j c_j(g_i)]} - \rho_0 \right) = 0 \right\},$$

and, with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_j: j = 0, 1, 2)$ and $\beta = (\beta_j: j = 0, 1, 2)$,

$$\hat{w}_i = \hat{w}_i(\alpha, \beta, \lambda) = n^{-1} \left(1 - \lambda \rho_0 + \lambda \sum_{j=0}^2 \frac{\delta_j \exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]}{1 + \exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]} \right)^{-1},$$

where, as in Qin and Lawless (1992), λ is determined by

$$\sum_{i=1}^n \frac{\sum_{j=0}^2 \delta_j \left(\frac{\exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]}{1 + \exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]} - \rho_0 \right)}{1 + \lambda \sum_{j=0}^2 w_i \delta_j \left(\frac{\exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]}{1 + \exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]} - \rho_0 \right)} = 0.$$

Note $\sum_{j=0}^2 \delta_j = 1$, the above is re-written as

$$\sum_{i=1}^n \left(1 - \lambda \rho_0 + \lambda \sum_{j=0}^2 \frac{\delta_j \exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]}{1 + \exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]} \right)^{-1} \sum_{j=0}^2 \frac{\delta_j \exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]}{1 + \exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]} = n \rho_0. \quad (7)$$

Let $\theta = (\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \delta_0, \delta_1, \delta_2)$ be the set of all the parameters, and $\tilde{h}(\cdot, \cdot | \theta)$ be the least favorable curve in this case. Define

$$\tilde{P}(\mathbf{y}, \mathbf{X} | \mathbf{g}, \mathbf{r}, \mathbf{s}; \theta, \tilde{h}(\mathbf{X}, \mathbf{g} | \theta)) = \prod_{i=1}^n \left(\sum_{j=0}^2 \delta_j \frac{(\exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)])^{y_i}}{1 + \exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]} \right) \tilde{h}(\mathbf{x}_i, g_i | \theta).$$

The profile empirical likelihood is, up to a quantity independent of θ

$$\tilde{L}(\theta, \hat{h}_n(\theta)) = \sum_{i=1}^n \left[\log \left(\sum_{j=0}^2 \delta_j \frac{(\exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)])^{y_i}}{1 + \exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]} \right) - \log \left(1 - \lambda \rho_0 + \lambda \sum_{j=0}^2 \frac{\delta_j \exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]}{1 + \exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]} \right) \right], \quad (8)$$

In this case, the computation of the MLE $\hat{\theta}_n$ of θ_0 based on (8) may not easy. A typical alternative way is to use the EM-algorithm (Dempster et al. 1977). For this, code the genetic mode status for subject i as v_{ij} , with $v_{ij} = 1$ if subject i is of genetic mode j , otherwise $v_{ij} = 0$, denote $\mathbf{v}_i = (v_{i0}, v_{i1}, v_{i2})$, and $\mathbf{v} = (\mathbf{v}_1, \dots, \mathbf{v}_n)$. Treating $\mathbf{v}$ as missing data, and $\mathbf{z} = (\mathbf{y}, \mathbf{X}, \mathbf{v})$ as the “complete data”, we propose the following semiparametric empirical likelihood for the complete data

$$\tilde{P}_c(\mathbf{z} | \mathbf{g}, \mathbf{r}, \mathbf{s}; \theta, \hat{h}_n(\theta)) = \prod_{i=1}^n \prod_{j=0}^2 \left(\delta_j \frac{(\exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)])^{y_i}}{1 + \exp[\alpha'_j \mathbf{x}_i + \beta_j c_j(g_i)]} \right)^{v_{ij}} \hat{h}_n(\mathbf{x}_i, g_i | \alpha, \beta)$$

and the corresponding log-likelihood is, up to a quantity independent of θ ,

$$\begin{aligned} \tilde{L}_c(\boldsymbol{\theta}, \hat{h}_n(\boldsymbol{\theta})) = & \sum_{i=1}^n \sum_{j=0}^2 v_{ij} \left(\log \delta_j + y_i [\boldsymbol{\alpha}'_j \mathbf{x}_i + \beta_j c_j(g_i)] - \log(1 + \exp[\boldsymbol{\alpha}'_j \mathbf{x}_i + \beta_j c_j(g_i)]) \right) \\ & - \sum_{i=1}^n \log \left(1 - \lambda \rho_0 + \lambda \sum_{j=0}^2 \frac{\delta_j \exp[\boldsymbol{\alpha}'_j \mathbf{x}_i + \beta_j c_j(g_i)]}{1 + \exp[\boldsymbol{\alpha}'_j \mathbf{x}_i + \beta_j c_j(g_i)]} \right). \end{aligned} \quad (9)$$

The algorithm needs starting values $\boldsymbol{\theta}^{(0)}$, especially we set $(\delta_0^{(0)}, \delta_1^{(0)}, \delta_2^{(0)}) = (1/3, 1/3, 1/3)$. Given the t -th step value $(\boldsymbol{\theta}^{(t)}, \mathbf{v}^{(t)})$, the $(t+1)$ -step values $(\boldsymbol{\theta}^{(t+1)}, \mathbf{v}^{(t+1)})$ are given by (Appendix)

$$v_{ij}^{(t+1)} = \delta_j^{(t)} \frac{(\exp[\boldsymbol{\alpha}'_j \mathbf{x}_i + \beta_j c_j(g_i)])^{y_i}}{1 + \exp[\boldsymbol{\alpha}'_j \mathbf{x}_i + \beta_j c_j(g_i)]} / \sum_{l=0}^2 \delta_l^{(t)} \frac{(\exp[\boldsymbol{\alpha}'_l \mathbf{x}_i + \beta_l c_l(g_i)])^{y_i}}{1 + \exp[\boldsymbol{\alpha}'_l \mathbf{x}_i + \beta_l c_l(g_i)]} \quad (10)$$

$(i=1, \dots, n; j=0, 1, 2),$

$$\delta_j^{(t+1)} = \frac{\sum_{i=1}^n v_{ij}^{(t+1)}}{\sum_{i=1}^n \sum_{l=0}^2 v_{il}^{(t+1)}}, \quad (l=0, 1, 2,), \quad (11)$$

Substitute $\mathbf{v}^{(t+1)}$ and $(\delta_0^{(t+1)}, \delta_1^{(t+1)}, \delta_2^{(t+1)})$ into (9), maximizing the rest parameters in (9), including λ , along with the constraint for λ in (7), to get the $(t+1)$ -th iteration estimates $\boldsymbol{\theta}^{(t+1)}$. Then, by the property of EM-algorithm, as $t \rightarrow \infty$, we have $\boldsymbol{\theta}^{(t)} \rightarrow \hat{\boldsymbol{\theta}}_n$. Convergence of the algorithm can be evaluated by the relative difference of successive parameter estimates, for some pre-specified $\varepsilon > 0$.

$$\frac{\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\|}{\|\boldsymbol{\theta}^{(t)}\|} \leq \varepsilon.$$

Let $l(y_i, \mathbf{x}_i, g_i | \boldsymbol{\theta}, h(\boldsymbol{\theta})) = \log P(y_i, \mathbf{x}_i, g_i | \boldsymbol{\theta}, h(\boldsymbol{\theta}))$ be the i -th terms in the right hand side of the log-likelihood (8), and define $l^{(1)}(y_i, \mathbf{x}_i, g_i | \boldsymbol{\theta}, h(\boldsymbol{\theta}))$ and $l^{(2)}(y_i, \mathbf{x}_i, g_i | \boldsymbol{\theta}, h(\boldsymbol{\theta}))$ similarly as before. Let $\mathbf{J}^{(1)} = E[l^{(1)}(y_i, \mathbf{x}_i, g_i | \boldsymbol{\theta}_0, h(\boldsymbol{\theta}_0)) | l^{(1)}(y_i, \mathbf{x}_i, g_i | \boldsymbol{\theta}_0, h(\boldsymbol{\theta}_0))']$ and $\mathbf{J}^{(2)} = -E[l^{(2)}(y_i, \mathbf{x}_i, g_i | \boldsymbol{\theta}_0, h(\boldsymbol{\theta}_0))]$. Parallel to Theorem 1, we have

Theorem 2—Under conditions of Theorem 1, as $n \rightarrow \infty$, we have $(\hat{\boldsymbol{\theta}}_n) \xrightarrow{a.s.} (\boldsymbol{\theta}_0)$, and

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Lambda}),$$

where $\boldsymbol{\Lambda} = (\mathbf{J}^{(2)})^{-1} \mathbf{J}^{(1)} (\mathbf{J}^{(2)})^{-1}$.

Similarly, $\mathbf{J}^{(1)}$ and $\mathbf{J}^{(2)}$ are consistently estimated by their empirical versions

$$\mathbf{J}_n^{(1)} = \frac{1}{n} \sum_{i=1}^n \tilde{l}^{(1)}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) [\tilde{l}^{(1)}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n))]'$$

$$\text{and } \mathbf{J}_n^{(2)} = \frac{1}{n} \sum_{i=1}^n \tilde{l}^{(2)}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)),$$

see the Appendix for details.

2.5 The gene frequency model

Now we turn to the gene frequency model part $P(\mathbf{r}, \mathbf{s} | \mathbf{p}, \mathbf{q}, f)$. Denote frequency of genotype G_i as p_i for the case population and q_i for the control population ($i = 0, 1, 2$), $\mathbf{p} = (p_0, p_1, p_2)$ and $\mathbf{q} = (q_0, q_1, q_2)$. A genetic model is referred to as the underlying mode of inheritance defined through penetrances $k_i = P(\text{case} | G_i)$ for $i = 0, 1, 2$. In terms of penetrances, the null hypothesis of no association is formulated as $H_{0,2}$: $k_0 = k_1 = k_2 = k$, where $k = P(\text{case})$ is the prevalence of the disease. The genotype relative risks are given by $\lambda_i = k_i/k_0$ using $G_0 = AA$ as the reference genotype. Under the alternative hypothesis $H_{1,2}$, REC, ADD, and DOM models correspond to $\lambda_1 = 1$, $\lambda_1 = (1 + \lambda_2)/2$, and $\lambda_1 = \lambda_2$, respectively. The genotype counts $\mathbf{r}$ and $\mathbf{s}$ have multinomial distributions as $\mathbf{r} \sim \text{Mul}(\mathbf{r}; \mathbf{p})$ and $\mathbf{s} \sim \text{Mul}(\mathbf{s}; \mathbf{q})$, with $p_i = k_i g_i / k$ and $q_i = (1 - k_i) g_i / (1 - k)$, and $g_i = P(G_i)$ is the frequency of genotype G_i in the whole population.

When there is no covariate except for the intercept, e.g. in a quick genome-wide scan, a closed form of $\hat{\beta}$ can be obtained for a given genetic model and this further simplifies the calculation of ABF. Assume B is the risk allele. Under REC model, genotypes G_0 and G_1 can be pooled so that the 2×3 table becomes a 2×2 table with counts $(r_0 + r_1, r_2)$ for cases and $(s_0 + s_1, s_2)$ for controls. Then β is the log odds ratio for the 2×2 table and $\beta = r_2(s_0 + s_1) / \{s_2(r_0 + r_1)\}$ with $V_n = 1/s_2 + 1/(r_0 + r_1) + 1/r_2 + 1/(s_0 + s_1)$. Similarly, under DOM model, $\beta = s_0(r_1 + r_2) / \{r_0(s_1 + s_2)\}$ with $V_n = 1/r_0 + 1/(s_1 + s_2) + 1/s_0 + 1/(r_1 + r_2)$.

Under ADD model, the MLE of β has no closed form. However, a 2×2 table can also be considered by counting number of alleles. This approach seems to be valid under Hardy-Weinberg equilibrium in the population (Sasieni, 1997; Duedj, Nuel, and Prum, 2008; Zheng, 2008a). Thus, under ADD, we have $(2r_0 + r_1, 2r_2 + r_1)$ for cases and $(2s_0 + s_1, 2s_2 + s_1)$ for controls. Hence, $\hat{\beta} = (2s_0 + s_1)(2r_2 + r_1) / \{(2s_2 + s_1)(2r_0 + r_1)\}$ with $V_n = 1/(2s_2 + s_1) + 1/(2r_0 + r_1) + 1/(2s_0 + s_1) + 1/(2r_2 + r_1)$. The performance of the ABF based on counting alleles will be examined later.

Deviation from HWE—Suppose a random sample of n individuals is drawn from the population. The genotype counts for (G_0, G_1, G_2) are $(n_0, n_1, n_2) \sim \text{Mul}(n; g_0, g_1, g_2)$. Under HWE, $(g_0, g_1, g_2) = (p_A^2, 2p_A p_B, p_B^2)$, where p_A is the frequency of allele A and $p_B = 1 - p_A$ is that for allele B . Several measures can be used to express departure from HWE. One of the commonly used is Wright's coefficient f of inbreeding (Weir, 1996), as $g_0 = p_A^2 + f p_A p_B$, $g_1 = 2p_A p_B(1 - f)$, and $g_2 = p_B^2 + f p_A p_B$. HWE holds if and only if $f = 0$. The MLE for f is

$$\hat{f} = \frac{4n_0 n_2 - n_1^2}{(2n_0 + n_1)(2n_2 + n_1)} \quad (12)$$

and by delta method, its asymptotic variance is (Weir, 1996)

$$V(f) = \frac{1-f}{2np_Ap_B} \{2p_Ap_B(1-f)(1-2f) + f(2-f)\}.. \quad (13)$$

Since the gene frequency model and the regression model are separate components, test for association and HWE in this case is the same as in Zheng and Ng (2008).

3 Simulation studies and real data analysis

In this section we demonstrate the use of the proposed methods by a simulated example, and then apply them to a real dataset from a genome-wide association analysis.

3.1 Simulation study

We use the example in Zheng and Ng (2008), except that here the data are simulated with a given rate ρ_0 of the disease in the selected population. We compare the results using the proposed method with those from Zheng and Ng (2008). The asymptotic variance matrices from these two methods may not be comparable as they are computed under different model assumptions, but we can compare the accuracies of the estimates.

We consider the following setup.

$$P(Y=1|X, G) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X - \beta_2 C(G))},$$

$(\pi_1, \pi_2, \pi_3) = (P(X = X_1 = (0,0)), P(X = X_2 = (0,1)), P(X = X_3 = (1,0))) = (0.2, 0.3, 0.5)$, $p = 0.1, 0.2, 0.3, 0.4, 0.5$; $F = 0.1$, ρ_0 ranges from 0.12 to 0.38; $(k_1, k_2, k_3) = (0.05, 0.1, 0.15)$

$$\begin{aligned} p(G=AA|X_i) &= (1-p_i)^2, p_i \sim \text{Beta}(p(1-F)/F, (1-p)(1-F)/F), \\ \beta_0 &= \log(k_1/(1-k_1)), \quad \beta_{11} = \log k_2(1-k_1)/\{k_1(1-k_2)\}, \\ \beta_{12} &= \log k_3(1-k_1)/\{k_1(1-k_3)\}, \quad \beta_2 = \log 1.15, \log 1.3, \log 1.4, \log 1.5, \quad n=1000 \end{aligned}$$

Dominant, recessive or additive model.

We see from Table 1 that, for dominant model, with side information the estimate for the intercept is improved, but the other estimates are almost the same. For the recessive model in Table 2, without side information the estimate on β_2 has serious bias, and with side information the estimate improved significantly. But estimates on the other parameters seem not sensitive to the auxiliary information. An interpretation is that, for recessive model, only genotype G_2 enters the model, and the number of G_2 individual is relatively sparse. This makes the estimation of β_2 difficult, and adding side information helps the estimation a lot. For the additive model as show in Table 3, the estimate of the intercept are very accurate without side information, and with side information it becomes a little bit worse, while all the other estimates are basically the same. However, the estimated mean squared errors for intercept with and without side information are close. In this case with side information, there is no much improvement. To mimic the case of no genetic effect ($\beta_2 = 0$), results are shown in the last rows of Table 1 and Table 3. We see some improvements with side information.

To summarize, with side information, the performances of estimates generally improves the performance of the estimates.

3.2 Real data analysis

For illustration, we analyzed a real GWAS dataset (Chen et al., 2011) which was conducted in the Howard University Family Study (HUFS) to identify genetic loci influencing serum total bilirubin levels in African Americans. This data set contains measurements of 270 SNPs, with covariates age, gender, and bmi for 619 healthy unrelated individuals. Total serum bilirubin, an important byproduct of heme metabolism, has been associated with several clinical outcomes, including cardiovascular disease, diabetes and drug metabolism. The primary goal of this study was to discover genetic loci influencing serum total bilirubin levels in African ancestry populations. The identification of such genes might lead to a further investigation for a causal link to high bilirubin development. The prevalence of the disease of “high bilirubin” is known to be 3.8% in the population and thus conveniently incorporated into the estimation via empirical likelihood. For simplicity, we assume a specific genetic model for a given SNP and the p-value for the significance of the SNP after adjusting for age, gender and bmi is calculated. For the additive model, the results for the 10 most significant SNPs are summarized in Table 4. For comparison, the results without incorporating disease prevalence information are also given. For the method without utilizing side information, none of the genes are significant with Bonferroni correction (the rejection level 0.05/270). However, with side information, two significant SNPs (rs887829 and rs10929302) are found. In Table 4, we displayed $-\log(\text{p-value})$ of the 10 most significant genes. We can see that the p-values with side information is generally smaller (or $-\log(\text{p-value})$ is bigger) than those without side information, indicating that by incorporating side information, our method is more efficient and enhances the chance of detecting SNPs which are associated with the phenotype. The conclusions based on the results for the dominant and recessive models are similar and unreported here. Based on our findings, it is evident that some genes play an important role in the determination of bilirubin level in human populations with African ancestral backgrounds. This also contributes to the understanding of the etiology of hyperbilirubinaemia and the pharmacogenomics role. Although some genes are not significant with Bonferroni correction, the SNPs with small p-values deserve further investigation.

4 Concluding remarks

We considered a semiparametric model for case-control genome-wide association studies with covariates. This is a joint approach and the nonparametric modeling of the distribution of the covariates is robust and further allows the easy incorporation of the side information to improve the estimation accuracy. The estimation procedure uses the empirical likelihood which profiles out the nonparametric component. The obtained estimates of the parameters are then used for hypothesis testing by the approximate Bayes factor which can also take into account the information available from the Hardy-Weinberg disequilibrium in cases. The proposed method is computationally convenient and can be readily used in genome-wide association studies where hundreds of thousands to a million genetic markers are tested. Simulation studies and a real application illustrates the practical utilities of the proposed methodology.

Acknowledgments

This work is supported in part by the National Center for Research Resources at NIH grant 2G12RR003048.

References

- Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*. 2005; 92:399–418.

- Chen G, Ramos E, Adeyemo A, et al. UGT1A1 is a Major Locus Influencing Bilirubin Levels in African Americans. *European Journal of Human Genetics*. 2011 To appear.
- Dai JY, LeBlanc M, Kooperberg C. Semiparametric estimation exploring covariate independence in two-phase randomized trials. *Biometrics*. 2009; 65:178–187. [PubMed: 18479485]
- Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 1977; 39:1–38.
- Efromovich S. On nonparametric regression for IID observations in a general setting. *Annals of Statistics*. 1996; 24:1126–1144.
- Fan J, Gijbels I. Censored regression: local linear approximations and their applications. *Journal of American Statistical Association*. 1994; 89:560–570.
- Jeffreys H. Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*. 1935; 31:203–222.
- Joo J, Kwak M, Zheng G. Improving power for testing genetic association in case-control studies by reducing the alternative space. *Biometrics*. 2009 in press. 10.1111/j.1541-0420.2009.01241.x
- Kass R, Raftery A. Bayes factors. *Journal of the American Statistical Association*. 1995; 90:773–795.
- Leung H, Qin J. Semi-parametric inference in a bivariate (multivariate) mixture model. *Statistica Sinica*. 2006; 16:153–163.
- Li QZ, Yu K, Li Z, Zheng G. MAX-rank: a simple and robust genome-wide scan for case-control association studies. *Human Genetics*. 2008; 123:617–623. [PubMed: 18491142]
- Murphy SA, Van der Vaart AW. On profile likelihood. *Journal of the American Statistical Association*. 2000; 93:1461–1474. (with discussions).
- Owen AB. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*. 1988; 75:237–249.
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.
- Qin J, Lawless JL. Empirical likelihood and general estimating equations. *Annals of Statistics*. 1994; 22:300–325.
- Schaid DJ, McDonnell SK, Hebring SJ, Cunningham JM. Nonparametric tests of association of mutation genes with human disease. *American Journal of Human Genetics*. 2005; 76:780–793. [PubMed: 15786018]
- Severini TA, Wong WH. Profile likelihood and conditionally parametric models. *Annals of Statistics*. 1992; 20:1768–1802.
- Song K, Elston RC. A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. *Statistics in Medicine*. 2006; 25:105–126. [PubMed: 16220513]
- Tan Z. A note on profile likelihood for exponential tilt mixture models. *Biometrika*. 2009; 96:229–236.
- Tzeng JY, Devlin B, Wasserman L, Roeder K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *American Journal of Human Genetics*. 2003; 72:891–902. [PubMed: 12610778]
- Weir B. *Genetic Data Analysis II*. Sinauer; Sunderland: MA: 1996.
- Yuan A, Yue Q, Apprey V, Bonney G. An extension of the weighted dissimilarity test to association study in families. *Human Genetics*. 2007; 122:83–94. [PubMed: 17530290]
- Zheng G. Can the allelic test be retired from analysis of case-control association studies? *Annals of Human Genetics*. 2008a; 72:848–851. [PubMed: 18652599]
- Zheng G. Analysis of ordered categorical data: Two score-independent approaches. *Biometrics*. 2008b; 64:1276–1279. [PubMed: 18266890]
- Zheng G, Ng HKT. Genetic model selection in two-phase analysis for case-control association studies. *Biostatistics*. 2008; 9:391–399. [PubMed: 18003629]
- Zheng G, Joo J, Zaykin DV, Wu CO, Geller NL. Robust tests in genome-wide scans under two-locus model. *Statistical Science*. 2009 Invited contribution (under review).

Appendix

Computation of (3)

For fixed (α, β) , $\hat{\mathbf{w}}$ is the maximizer of $[P(\mathbf{y}|\mathbf{X}, \mathbf{g}, \mathbf{r}, \mathbf{s}; \alpha, \beta) \prod_{i=1}^n w_i]$ subject to the constraints in $\mathcal{W}$. Taking logarithm of the joint likelihood, and using Lagrange multiplier for the constraints in $\mathcal{W}$, equivalently, $\hat{\mathbf{w}}$ is the maximizer of $J(\mathbf{w})$, with

$$J(\mathbf{w}) = \sum_{i=1}^n \log w_i - \gamma \left(\sum_{i=1}^n w_i - 1 \right) - n\lambda \sum_{i=1}^n w_i \left(\frac{\exp[\alpha' \mathbf{x}_i + \beta c(g_i)]}{1 + \exp[\alpha' \mathbf{x}_i + \beta c(g_i)]} - \rho_0 \right).$$

Set

$$0 = \frac{\partial J(\mathbf{w})}{\partial w_i} = \frac{1}{w_i} - \gamma - n\lambda \left(\frac{\exp[\alpha' \mathbf{x}_i + \beta c(g_i)]}{1 + \exp[\alpha' \mathbf{x}_i + \beta c(g_i)]} - \rho_0 \right)$$

we get $\gamma = n$ and

$$\hat{w}_i = \hat{w}_i(\alpha, \beta, \lambda) = n^{-1} \left(1 + \lambda \left(\frac{\exp[\alpha' \mathbf{x}_i + \beta c(g_i)]}{1 + \exp[\alpha' \mathbf{x}_i + \beta c(g_i)]} - \rho_0 \right) \right)^{-1},$$

with λ determined by the equation

$$\sum_{i=1}^n w_i(\alpha, \beta, \lambda) \frac{\exp[\alpha' \mathbf{x}_i + \beta c(g_i)]}{1 + \exp[\alpha' \mathbf{x}_i + \beta c(g_i)]} = \rho_0, \quad \text{or} \quad \sum_{i=1}^n \frac{w_i(\alpha, \beta, \lambda)}{1 + \exp[\alpha' \mathbf{x}_i + \beta c(g_i)]} = 1 - \rho_0.$$

Proof of Theorem 1

Let

$$\tilde{L}_n^{(1)}(\alpha, \beta, h(\alpha, \beta)) = \frac{1}{n} \sum_{i=1}^n \tilde{l}^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha, \beta, h(\alpha, \beta)),$$

and

$$\tilde{L}_n^{(2)}(\alpha, \beta, h(\alpha, \beta)) = \frac{1}{n} \sum_{i=1}^n \tilde{l}^{(2)}(y_i, \mathbf{x}_i, g_i | \alpha, \beta, h(\alpha, \beta)).$$

Since $\tilde{L}_n^{(1)}(\hat{\alpha}_n, \hat{\beta}_n, \hat{h}_n(\hat{\alpha}_n, \hat{\beta}_n)) = \mathbf{0}$, we have

$$\begin{aligned} -\tilde{L}_n^{(1)}(\alpha_0, \beta_0, \hat{h}_n(\alpha_0, \beta_0)) &= \tilde{L}_n^{(1)}(\hat{\alpha}_n, \hat{\beta}_n, \hat{h}_n(\hat{\alpha}_n, \hat{\beta}_n)) - \tilde{L}_n^{(1)}(\alpha_0, \beta_0, \hat{h}_n(\alpha_0, \beta_0)) \\ &= \tilde{L}_n^{(2)}(\alpha_n, \beta_n, h_n)(\hat{\alpha}_n - \alpha_0, \hat{\beta}_n - \beta_0), \end{aligned}$$

where (α_n, β_n, h_n) is an intermediate point between $(\hat{\alpha}_n, \hat{\beta}_n, \hat{h}_n(\hat{\alpha}_n, \hat{\beta}_n))$ and (α_0, β_0, h_0) . This gives

$$(\hat{\alpha}_n - \alpha_0, \hat{\beta}_n - \beta_0) = -[\tilde{L}_n^{(2)}(\alpha_n, \beta_n, h_n)]^{-1} \tilde{L}_n^{(1)}(\alpha_0, \beta_0, \hat{h}_n(\alpha_0, \beta_0)). \quad (\text{A.1})$$

By the assumption, $\tilde{L}_n^{(2)}(\alpha_n, \beta_n, h_n)$ is non-singular for large n as long as (α_n, β_n) stay in a neighborhood of (α_0, β_0) .

Let $h_0(\cdot, \cdot)$ be the true density/mass function of $(\mathbf{x}_i, g_i)$. Note $\hat{h}_n(\cdot, \cdot | \alpha, \beta)$ can be constructed from the set $\{\hat{w}_i(\alpha, \beta)\}$, of the form $\hat{h}_n(\mathbf{x}, g | \alpha, \beta) = \hat{w}_i(\alpha, \beta) / V_i(\mathbf{x}_i, g_i)$ for $(\mathbf{x}, g) \in V_i(\mathbf{x}_i, g_i)$, here we use $V_i(\mathbf{x}_i, g_i)$ to denote both a volumn element and a neighborhood, depending on the observation $(\mathbf{x}_i, g_i)$ and some its neighboring observations, but independent of the parameters. Since the least favorable curve h minimizes the Kullback-Leibler divergence between the joint model and the one under the true h_0 , and for fixed (α, β) , $\hat{h}_n(\cdot, \cdot | \alpha, \beta)$ is the nonparametric maximum likelihood estimator, it maximizes the empirical Kullback-Leibler divergence between the joint model and the one under the true h_0 , subject to the constraints in $\mathcal{W}$. So under mild conditions, it is known (cf. Serverini and Wong, 1992) that $\hat{h}_n \rightarrow h$ (a.s.) (a.e.), and $h(\cdot, \cdot | \alpha_0, \beta_0) = h_0(\cdot, \cdot)$. The full form of $l(\tilde{y}_i, \mathbf{x}_i, g_i | \alpha, \beta, \hat{h}_n(\alpha, \beta))$ is

$$(y_i[\alpha' \mathbf{x}_i + \beta c(g_i)] - \log[1 - \lambda \rho_0 + (1 + \lambda(1 - \rho_0)) \exp[\alpha' \mathbf{x}_i + \beta c(g_i)]] - \log V_i(\mathbf{x}_i, g_i).$$

Since the last term is independent of (α, β) , the partial derivatives $l^{(k)}(y_i, \mathbf{x}_i, g_i | \alpha, \beta, \hat{h}_n)$ ($k = 1, 2$) are the same when computed using the expression for l given in (5).

We have

$$\begin{aligned} \tilde{L}_n^{(1)}(\alpha_0, \beta_0, \hat{h}_n(\alpha_0, \beta_0)) &= \tilde{L}_n^{(1)}(\alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0)) \\ &+ \langle \frac{\partial}{\partial h} \tilde{L}_n^{(1)}(\alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0)), \hat{h}_n(\alpha_0, \beta_0) - \tilde{h}(\alpha_0, \beta_0) \rangle + o(n^{-1/2}) \quad (\text{a.s.}) \end{aligned} \quad (\text{A.2})$$

where

$$\begin{aligned} &\langle \frac{\partial}{\partial h} \tilde{L}_n^{(1)}(\alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0)), \hat{h}_n(\alpha_0, \beta_0) - \tilde{h}(\alpha_0, \beta_0) \rangle \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial h} \tilde{l}^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0)) (\hat{h}_n(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0) - \tilde{h}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0)). \end{aligned}$$

In the above the partial derivative $\frac{\partial}{\partial h} \tilde{l}^{(1)}$ is in the Fréchet sense, as h is a function.

Let $P(y_i, \mathbf{x}_i, g_i | \alpha, \beta, h_0)$ be the true joint distribution of $(y_i, \mathbf{x}_i, g_i)$ corresponding to h_0 , $P(\tilde{y}_i, \mathbf{x}_i, g_i | \alpha, \beta, h(\alpha, \beta))$ be that corresponding to h , and $P^{(\tilde{I})}(\tilde{y}_i, \mathbf{x}_i, g_i | \alpha, \beta, h(\alpha, \beta))$ be its first order

partial derivatives with respect to (α, β) . Since P is a density for all (α, β) , we have $\int P^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha, \beta, h(\alpha, \beta)) d\mathbf{x}_i dg_i = \mathbf{0}$, and so

$$\begin{aligned} & E[\tilde{l}^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0))] \\ &= \int P(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, h_0) \frac{\tilde{P}^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0))}{\tilde{P}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0))} d\mathbf{x}_i dg_i \\ &= \int P(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, h_0) \frac{\tilde{P}^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0))}{P(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, h_0)} d\mathbf{x}_i dg_i \\ &= \int \tilde{P}^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0)) d\mathbf{x}_i dg_i = \mathbf{0}. \end{aligned}$$

Now $E[\frac{\partial}{\partial h} l^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0))] = \frac{\partial}{\partial h} E[l^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0))] = \mathbf{0}$, thus by the strong law of large numbers,

$$\tilde{L}_n^{(1)}(\alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0)) \xrightarrow{a.s.} E[l^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0))] = \mathbf{0}$$

and

$$\frac{\partial}{\partial h} \tilde{L}_n^{(1)}(\alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0)) \xrightarrow{a.s.} E[\frac{\partial}{\partial h} l^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0))] = \mathbf{0}.$$

Also, $\hat{h}_n(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0) - h(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0) = o(1)$ for all i . So by (A.2)

$$\begin{aligned} & \tilde{L}_n^{(1)}(\alpha_0, \beta_0, \hat{h}(\alpha_0, \beta_0)) \xrightarrow{a.s.} E[\tilde{l}^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0))] \\ & + E[\frac{\partial}{\partial h} l^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0))] o(1) = \mathbf{0}. \end{aligned}$$

This and (A.1) gives the a.s. consistency of $(\hat{\alpha}_n, \hat{\beta}_n)$.

In addition, by central limit theorem,

$$\sqrt{n} \frac{\partial}{\partial h} \tilde{L}_n^{(1)}(\alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0)) \xrightarrow{D} N(\mathbf{0}, \mathbf{V})$$

for some $\mathbf{V}$, or $\frac{\partial}{\partial h} \tilde{L}_n^{(1)}(\alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0)) = O_p(n^{-1/2})$ and so

$\langle \frac{\partial}{\partial h} \tilde{L}_n^{(1)}(\alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0)), \hat{h}_n(\alpha_0, \beta_0) - \tilde{h}(\alpha_0, \beta_0) \rangle = o_p(n^{-1/2})$. Thus by (A.2) and the above argument,

$$\sqrt{n}(\hat{\alpha}_n - \alpha_0, \hat{\beta}_n - \beta_0) = -[\tilde{L}_n^{(2)}(\alpha_n, \beta_n, h_n)]^{-1} \sqrt{n} \tilde{L}_n^{(1)}(\alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0)) + o(1).$$

Since $\tilde{L}_n^{(1)}(\alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0))$ is an average of i.i.d. random vectors each with mean zero vector and variance matrix $\tilde{\mathbf{I}}^{(1)}$, by the central limit theorem,

$$\sqrt{n}\tilde{L}_n^{(1)}(\alpha_0, \beta_0, \tilde{h}(\alpha_0, \beta_0)) \xrightarrow{D} N(\mathbf{0}, \tilde{\mathbf{I}}^{(1)}).$$

Also, by the consistency of $\hat{h}_n(\cdot, \cdot)$ and $(\hat{\alpha}_n, \hat{\beta}_n)$, we have $\hat{h}(\cdot, \cdot | \hat{\alpha}_n, \hat{\beta}_n) \rightarrow h(\cdot, \cdot | \alpha_0, \beta_0) = h_0(\cdot, \cdot)$, and $\tilde{L}_n^{(2)}(\alpha_n, \beta_n, h_n) \xrightarrow{a.s.} \tilde{\mathbf{I}}^{(2)}$, and this completes the proof using Slutsky's theorem.

Note although $E[l^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, h(\alpha_0, \beta_0))] = E[l^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, h_0)] = \mathbf{0}$, but generally $l^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, h(\alpha_0, \beta_0)) \neq l^{(1)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, h_0)$, and $l^{(2)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, h(\alpha_0, \beta_0)) \neq l^{(2)}(y_i, \mathbf{x}_i, g_i | \alpha_0, \beta_0, h_0)$. So $\tilde{\mathbf{I}}^{(1)}$ and $\tilde{\mathbf{I}}^{(2)}$ are not the usual Fisher information matrix.

Derivation of (10)–(11)

It is obvious that $P(v_{ij}=1 | \boldsymbol{\theta}^{(t)} = \delta_j^{(t)})$, so

$$\begin{aligned} v_{ij}^{(t+1)} &= E(v_{ij} | \mathbf{z}_i, \boldsymbol{\theta}^{(t)}) = P(v_{ij}=1 | \mathbf{z}_i, \boldsymbol{\theta}^{(t)}) = \frac{P(\mathbf{z}_i | \boldsymbol{\theta}^{(t)}, v_{ij}=1) P(v_{ij}=1 | \boldsymbol{\theta}^{(t)})}{\sum_{l=0}^2 P(\mathbf{z}_i | \boldsymbol{\theta}^{(t)}, v_{il}=1) P(v_{il}=1 | \boldsymbol{\theta}^{(t)})} \\ &= \delta_j^{(t)} \frac{(\exp[\alpha_j^{(t)'} \mathbf{x}_i + \beta_j^{(t)} c_j(g_i)])^{y_i}}{1 + \exp[\alpha_j^{(t)'} \mathbf{x}_i + \beta_j^{(t)} c_j(g_i)]} / \sum_{l=0}^2 \delta_l^{(t)} \frac{(\exp[\alpha_l^{(t)'} \mathbf{x}_i + \beta_l^{(t)} c_l(g_i)])^{y_i}}{1 + \exp[\alpha_l^{(t)'} \mathbf{x}_i + \beta_l^{(t)} c_l(g_i)]}. \end{aligned}$$

This gives (10).

For (11), we need to maximize $L(\boldsymbol{\theta}, \hat{h}_n(\boldsymbol{\theta}))$, given by (9), over $\boldsymbol{\theta}$. It is easy to see that this can be done by maximizing $(\delta_0, \delta_1, \delta_2)$ subject to the constraint $\sum_{j=0}^2 \delta_j = 1$, separately from the rest parameters. This gives (11).

Expressions for $\tilde{l}^{(1)}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n))$ and $\tilde{l}^{(2)}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n))$

We have, with $a_{ji} = \exp[\hat{\alpha}_{nj}' \mathbf{x}_i + \hat{\beta}_{nj} c_j(g_i)]$, for $j, k = 0, 1, 2$ ($j \neq k$); and

$$b_i = \sum_{l=0}^2 \hat{\delta}_{nl} a_{li}^{y_i} / (1 + a_{li}) \text{ and } d_i = 1 - \lambda \rho_0 + \lambda b_i, (i = 1, \dots, n),$$

$$\frac{\partial}{\partial \alpha_j} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) = \frac{\hat{\delta}_{nj} a_{ji}}{(1 + a_{ji})^2} \left(\frac{(-1)^{y_i-1}}{b_i} - \frac{\lambda}{d_i} \right) \mathbf{x}_i,$$

$$\frac{\partial}{\partial \beta_j} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) = \frac{\hat{\delta}_{nj} a_{ji}}{(1 + a_{ji})^2} \left(\frac{(-1)^{y_i-1}}{b_i} - \frac{\lambda}{d_i} \right) c_j(g_i), \frac{\partial}{\partial \delta_j} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) = \frac{a_{ji}}{1 + a_{ji}} \left(\frac{a_{ji}^{y_i-1}}{b_i} - \frac{\lambda}{d_i} \right),$$

$$\frac{\partial}{\partial \alpha_j \partial \alpha_j'} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) = \frac{\hat{\delta}_{nj} a_{ji}}{(1 + a_{ji})^2} \left[\frac{1 - a_{ji}}{1 + a_{ji}} \left(\frac{(-1)^{y_i-1}}{b_i} - \frac{\lambda}{d_i} \right) - \frac{a_{ji}}{(1 + a_{ji})^2} \left(\frac{\hat{\delta}_{nj}}{b_i^2} - \frac{\lambda^2}{d_i^2} \right) \right] \mathbf{x}_i \mathbf{x}_i',$$

$$\begin{aligned} \frac{\partial}{\partial \alpha_j \partial \alpha_k} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) \\ = \frac{\hat{\delta}_{nj} a_{ji}}{(1+a_{ji})^2} \frac{\hat{\delta}_{nk} a_{ki}}{(1+a_{ki})^2} \left(-\frac{1}{b_i^2} + \frac{\lambda^2}{d_i^2} \right) \mathbf{x}_i \mathbf{x}_k', \frac{\partial}{\partial \alpha_j \partial \beta_j} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) = \frac{\hat{\delta}_{nj} a_{ji}}{(1+a_{ji})^2} \left[\frac{1-a_{ji}}{1+a_{ji}} \left(\frac{(-1)^{y_i-1}}{b_i} - \frac{\lambda}{d_i} \right) - \frac{a_{ji}}{(1+a_{ji})} \right] \\ = \frac{\hat{\delta}_{nj} a_{ji}}{(1+a_{ji})^2} \frac{\hat{\delta}_{nk} a_{ki}}{(1+a_{ki})^2} \left(-\frac{1}{b_i^2} + \frac{\lambda^2}{d_i^2} \right) c_k(g_i) \mathbf{x}_i, \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \alpha_j \partial \delta_j} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) = \frac{a_{ji}}{(1+a_{ji})^2} \left[\frac{(-1)^{y_i-1}}{b_i} - \frac{\lambda}{d_i} + \frac{a_{ji}}{1+a_{ji}} \left(\frac{(-a_{ji})^{y_i-1}}{b_i} - \frac{\lambda^2}{d_i} \right) \right] \mathbf{x}_i, \frac{\partial}{\partial \alpha_j \partial \delta_k} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) \\ = \frac{\hat{\delta}_{nj} a_{ji}}{(1+a_{ji})^2} \frac{a_{ki}}{1+a_{ki}} \left(\frac{(-a_{ki})^{y_i-1}}{b_i} - \frac{\lambda^2}{d_i} \right) \mathbf{x}_i, \end{aligned}$$

$$\frac{\partial}{\partial \beta_j \partial \beta_j} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) = \frac{\hat{\delta}_{nj} a_{ji}}{(1+a_{ji})^2} \left[\frac{1-a_{ji}}{1+a_{ji}} \left(\frac{(-1)^{y_i-1}}{b_i} - \frac{\lambda}{d_i} \right) - \frac{a_{ji}}{(1+a_{ji})^2} \left(\frac{\hat{\delta}_{nj}}{b_i^2} + \frac{\lambda^2}{d_i^2} \right) \right] c_j^2(g_i),$$

$$\begin{aligned} \frac{\partial}{\partial \beta_j \partial \beta_k} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) \\ = \frac{\hat{\delta}_{nj} a_{ji}}{(1+a_{ji})^2} \frac{\hat{\delta}_{nk} a_{ki}}{(1+a_{ki})^2} \left(-\frac{1}{b_i^2} + \frac{\lambda^2}{d_i^2} \right) c_j(g_i) c_k(g_i), \frac{\partial}{\partial \beta_j \partial \delta_j} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) \\ = \frac{a_{ji}}{(1+a_{ji})^2} \left[\frac{(-1)^{y_i-1}}{b_i} - \frac{\lambda}{d_i} + \frac{a_{ji}}{1+a_{ji}} \left(\frac{(-a_{ji})^{y_i-1}}{b_i} - \frac{\lambda^2}{d_i} \right) \right] c_j(g_i), \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \beta_j \partial \delta_k} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) \\ = -\frac{\hat{\delta}_{nj} a_{ji}}{(1+a_{ji})^2} \frac{a_{ki}}{1+a_{ki}} \left(\frac{(-a_{ki})^{y_i-1}}{b_i} - \frac{\lambda^2}{d_i} \right) c_j(g_i), \frac{\partial}{\partial \delta_j \partial \delta_j} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) \\ = \frac{a_{ji}^2}{(1+a_{ji})^2} \left(-\frac{a_{ji}^{2(y_i-1)}}{b_i^2} + \frac{\lambda^2}{d_i^2} \right), \end{aligned}$$

$$\frac{\partial}{\partial \delta_j \partial \delta_k} \tilde{l}(y_i, \mathbf{x}_i, g_i | \hat{\boldsymbol{\theta}}_n, \tilde{h}(\hat{\boldsymbol{\theta}}_n)) = \frac{a_{ji} a_{ki}}{(1+a_{ji})(1+a_{ki})} \left(-\frac{(a_{ji} a_{ki})^{y_i-1}}{b_i^2} + \frac{\lambda^2}{d_i^2} \right),$$

and λ is determined by (9) with $\boldsymbol{\theta}$ replaced by $\hat{\boldsymbol{\theta}}_n$.

Table 1

Biases and Mean squared errors for β_0 , β_2 , β_{11} , β_{12} , for each p and β_2 , the first line is bias and the second line is the mean squared error

Model	p	β_2	without side information				with side information				
			β_0	β_2	β_{11}	β_{12}	β_0	β_2	β_{11}	β_{12}	
DOM	0.1	1.15	-0.0215	0.0317	0.0430	-0.0196	-0.0189	0.0316	0.0427	-0.0198	
			0.1140	0.1000	0.1299	0.1394	0.1038	0.1000	0.1300	0.1395	
		1.30	-0.0753	-0.0066	0.0728	0.0531	-0.0630	-0.0069	0.0731	0.0535	
			0.1520	0.0528	0.1626	0.1588	0.1379	0.0528	0.1626	0.1587	
	1.40	1.15	-0.0572	-0.0036	0.0212	0.0495	-0.0466	-0.0035	0.0214	0.0497	
			0.0895	0.0889	0.1238	0.1026	0.0856	0.0889	0.1239	0.1027	
		1.50	0.0133	-1.0462	-0.0310	-0.0078	0.0114	-0.1065	-0.0310	-0.0079	
			0.0887	106.66	0.1079	0.1068	0.0830	0.9057	0.1079	0.1068	
		0.2	1.15	-0.0267	-0.0089	0.0231	0.0241	-0.0270	-0.0090	0.0232	0.0243
				0.0892	0.0427	0.1142	0.0935	0.0817	0.0427	0.1142	0.0934
			1.30	-0.0209	-0.0286	0.0201	0.0332	-0.0261	-0.0286	0.0201	0.0331
				0.1306	0.0449	0.1726	0.1456	0.1256	0.0449	0.1727	0.1457
1.40		1.15	-0.0405	-0.0050	0.0352	0.0232	-0.0318	-0.0049	0.0352	0.0233	
			0.1311	0.0323	0.1406	0.1193	0.1137	0.0323	0.1406	0.1194	
		1.50	0.0161	0.0070	-0.0292	-0.0110	0.0037	0.0070	-0.0293	-0.0111	
			0.0849	0.0384	0.1015	0.0860	0.0755	0.0384	0.1015	0.0861	
0.3		1.15	-0.0650	0.0304	0.0389	0.0529	-0.0675	0.0304	0.0389	0.0529	
			0.0834	0.0450	0.0789	0.0728	0.0720	0.0450	0.0789	0.0728	
		1.30	-0.0410	0.0171	0.0077	0.0196	-0.0287	0.0170	0.0077	0.0196	
			0.0750	0.0379	0.0781	0.0800	0.0729	0.0379	0.0782	0.0801	
1.40	1.15	-0.0105	-0.0179	0.0073	0.0123	-0.0078	-0.0179	0.0073	0.0123		
		0.0767	0.0350	0.0779	0.0652	0.0776	0.0350	0.0779	0.0652		
	1.50	-0.0759	0.0331	0.0835	0.0255	-0.0654	0.0331	0.0835	0.0255		
		0.0647	0.0443	0.0735	0.0561	0.0547	0.0443	0.0735	0.0562		
	0.4	1.15	-0.0803	0.0547	0.0300	0.0451	-0.0775	0.0547	0.0301	0.0452	
			0.0786	0.0427	0.0806	0.0556	0.0776	0.0427	0.0806	0.0556	

Model	p	without side information				with side information				
		β_2	β_0	β_2	β_{11}	β_{12}	β_0	β_2	β_{11}	β_{12}
	1.30	1.30	-0.0147	-0.0009	0.0273	0.0313	-0.0311	-0.0010	0.0273	0.0313
			0.0797	0.0334	0.0665	0.0586	0.0722	0.0334	0.0666	0.0586
		1.40	-0.0083	0.0017	-0.0167	0.0262	-0.0186	0.0017	-0.0168	0.0261
	1.50	1.50	0.0790	0.0414	0.0742	0.0657	0.0773	0.0414	0.0743	0.0657
			-0.0479	0.0204	0.0127	0.0206	-0.0392	0.0204	0.0127	0.0206
		0.0991	0.0496	0.0824	0.0666	0.0864	0.0496	0.0825	0.0666	
	0.5	1.15	0.0020	0.0214	-0.0178	-0.0117	-0.0133	0.0214	-0.0178	-0.0117
			0.0772	0.0618	0.0606	0.0538	0.0762	0.0617	0.0607	0.0538
		1.30	-0.0257	0.0352	0.0084	-0.0008	-0.0370	0.0352	0.0085	-0.0007
1.40	1.40	0.0843	0.0689	0.0518	0.0432	0.0770	0.0689	0.0518	0.0432	
		-0.0445	0.0029	0.0296	0.0252	-0.0296	0.0029	0.0296	0.0252	
	0.0843	0.0461	0.0750	0.0686	0.0749	0.0461	0.0751	0.0687		
1.50	1.50	-0.0239	-0.0143	0.0247	0.0347	-0.0196	-0.0143	0.0247	0.0347	
		0.0679	0.0361	0.0561	0.0506	0.0613	0.0361	0.0561	0.0507	
	0.0	0.0	-0.0354	-0.0040	0.0316	0.0291	-0.0380	-0.0038	0.0313	0.0288
0.1296			0.0562	0.1285	0.1329	0.1105	0.0562	0.1284	0.1330	

Table 2

Biases and Mean squared errors for β_0 , β_2 , β_{11} , β_{12} , for each p and β_2 , the first line is bias and the second line is the mean squared error

Model	p	β_2	without side information				with side information			
			β_0	β_2	β_{11}	β_{12}	β_0	β_2	β_{11}	β_{12}
REC	0.1	1.15	-0.0561	-9.1848	0.0197	0.0616	-0.0546	-0.6270	0.0203	0.0623
			0.1214	927.58	0.1384	0.1498	0.1120	6.3150	0.1392	0.1506
		1.30	-0.0364	-14.1590	0.0200	0.0201	-0.0335	-0.6553	0.0258	0.0256
			0.1482	2291.2	0.1888	0.1461	0.1296	10.1034	0.1889	0.1485
	1.40	1.40	-0.0974	-7.0092	0.0933	0.1131	-0.1093	-0.2362	0.0933	0.1132
			0.1648	1355.8	0.2237	0.1875	0.1594	6.1843	0.2242	0.1871
		1.50	-0.0531	-7.1643	0.0563	0.0595	-0.0585	-0.4003	0.0561	0.0594
			0.0997	1147.8	0.1491	0.1074	0.0912	5.2266	0.1495	0.1071
	0.2	1.15	-0.0217	-0.9955	0.0118	0.0318	-0.0337	-0.0745	0.0118	0.0317
			0.1126	103.71	0.1544	0.1114	0.0981	1.0521	0.1542	0.1111
		1.30	-0.0513	-1.9937	0.0316	0.0730	-0.0645	-0.0991	0.0323	0.0736
			0.1504	209.58	0.1883	0.1809	0.1356	1.3881	0.1883	0.1809
	1.40	1.40	-0.0035	-1.0677	0.0039	-0.0155	0.0011	-0.1180	0.0046	-0.0147
			0.1313	104.80	0.1686	0.1433	0.1223	0.7137	0.1682	0.1432
		1.50	0.0028	-2.0681	0.0019	0.0006	-0.0101	-0.1958	0.0016	0.0004
			0.1163	419.18	0.1319	0.1331	0.1048	2.6639	0.1318	0.1330
	0.3	1.15	-0.0149	-0.0369	0.0215	0.0040	-0.0150	-0.0353	0.0214	0.0040
			0.0974	0.0981	0.1309	0.1165	0.0941	0.0974	0.1312	0.1166
		1.30	-0.0773	0.0568	0.0252	0.0534	-0.0630	0.0563	0.0243	0.0526
			0.1645	0.1115	0.1905	0.1899	0.1521	0.1126	0.1900	0.1891
	1.40	1.40	-0.0619	-0.0015	0.0541	0.0662	-0.0635	-0.0015	0.0542	0.0664
			0.1194	0.0663	0.1501	0.1121	0.0987	0.0664	0.1501	0.1121
		1.50	-0.0144	0.0069	0.0177	-0.0039	-0.0074	0.0069	0.0176	-0.0040
			0.1052	0.0704	0.1365	0.1247	0.1008	0.0704	0.1366	0.1247
	0.4	1.15	0.0254	-0.0006	-0.0531	-0.0298	0.0204	-0.0006	-0.0530	-0.0297
			0.1187	0.0449	0.1515	0.1229	0.1013	0.0449	0.1515	0.1230

Model	p	without side information					with side information				
		β_2	β_0	β_2	β_{11}	β_{12}	β_0	β_2	β_{11}	β_{12}	
	1.30	1.15	-0.0933	0.0222	0.0632	0.0987	-0.0952	0.0221	0.0631	0.0985	
			0.1614	0.0566	0.2149	0.1950	0.1667	0.0565	0.2146	0.1948	
	1.40	1.40	-0.0287	-0.0147	0.0150	0.0372	-0.0345	-0.0146	0.0150	0.0372	
			0.0922	0.0588	0.1142	0.1020	0.0910	0.0587	0.1143	0.1021	
	1.50	1.50	-0.0661	0.0086	0.0708	0.0467	-0.0569	0.0083	0.0710	0.0467	
			0.0841	0.0452	0.1112	0.0871	0.0775	0.0452	0.1113	0.0871	
	0.5	1.15	-0.0528	0.0227	0.0401	0.0473	-0.0607	0.0227	0.0401	0.0472	
			0.1229	0.0417	0.1494	0.1353	0.1201	0.0417	0.1494	0.1354	
	1.30	1.30	-0.0863	0.0073	0.0933	0.0889	-0.0887	0.0072	0.0933	0.0889	
			0.1025	0.0414	0.1175	0.1164	0.0964	0.0414	0.1177	0.1165	
	1.40	1.40	-0.0583	0.0126	0.0543	0.0611	-0.0520	0.0125	0.0542	0.0611	
			0.1191	0.0402	0.1351	0.1180	0.1053	0.0402	0.1351	0.1179	
1.50	1.50	-0.0563	0.0229	0.0296	0.0319	-0.0501	0.0229	0.0297	0.0320		
		0.0957	0.0376	0.0947	0.0973	0.0852	0.0376	0.0948	0.0974		

Table 3

Biases and Mean squared errors for β_0 , β_2 , β_{11} , β_{12} , for each p and β_2 , the first line is bias and the second line is the mean squared error

Model	p	β_2	without side information				with side information			
			β_0	β_2	β_{11}	β_{12}	β_0	β_2	β_{11}	β_{12}
ADD	0.1	1.15	-0.0063	-0.1356	0.0300	-0.0689	0.0592	-0.1356	0.0301	-0.0688
			0.0782	0.1171	0.1216	0.1015	0.0780	0.1172	0.1215	0.1015
		1.30	-0.0216	-0.1788	0.0375	-0.0346	0.0468	-0.1789	0.0374	-0.0347
			0.1019	0.1000	0.1200	0.1172	0.0966	0.1001	0.1200	0.1172
	1.40	1.15	0.0088	-0.1823	0.0279	-0.1082	0.0877	-0.1823	0.0279	-0.1081
			0.0825	0.0856	0.1111	0.1174	0.0890	0.0856	0.1111	0.1174
		1.50	0.0083	-0.1765	0.0258	-0.0756	0.0818	-0.1766	0.0258	-0.0756
			0.0935	0.0816	0.1227	0.1297	0.1005	0.0816	0.1227	0.1296
	0.2	1.15	0.0380	-0.1005	-0.0121	-0.1751	0.1406	-0.1006	-0.0120	-0.1751
			0.0747	0.0345	0.1108	0.1314	0.0936	0.0345	0.1108	0.1314
		1.30	0.0838	-0.1532	-0.0057	-0.2015	0.1747	-0.1532	-0.0058	-0.2015
			0.0915	0.0599	0.1349	0.1384	0.1123	0.0599	0.1349	0.1384
1.40	1.15	0.0665	-0.1214	-0.0074	-0.2152	0.1641	-0.1214	-0.0075	-0.2153	
		0.1458	0.0471	0.1457	0.1635	0.1478	0.0471	0.1457	0.1636	
	1.50	0.0862	-0.1123	-0.0323	-0.2304	0.1695	-0.1124	-0.0323	-0.2303	
		0.1040	0.0513	0.1037	0.1470	0.1134	0.0513	0.1038	0.1471	
0.3	1.15	1.15	0.0294	-0.0681	0.0396	-0.2060	0.1259	-0.0681	0.0396	-0.2060
			0.0808	0.0335	0.0859	0.1170	0.0902	0.0335	0.0859	0.1171
		1.30	0.0149	-0.0605	0.0178	-0.2197	0.1288	-0.0605	0.0178	-0.2196
			0.0949	0.0272	0.1064	0.1594	0.1047	0.0272	0.1065	0.1595
	1.40	1.15	0.0377	-0.0788	0.0269	-0.2099	0.1389	-0.0788	0.0269	-0.2099
			0.0644	0.0263	0.0649	0.1056	0.0784	0.0263	0.0649	0.1056
		1.50	0.0476	-0.0608	0.0100	-0.2847	0.1602	-0.0609	0.0099	-0.2848
			0.0954	0.0353	0.0833	0.1629	0.1147	0.0353	0.0833	0.1630
	0.4	1.15	0.0082	0.0047	-0.0244	-0.2630	0.1169	0.0047	-0.0243	-0.2630
			0.0877	0.0245	0.0732	0.1369	0.0903	0.0245	0.0732	0.1369

Model	p	without side information				with side information					
		β_2	β_0	β_2	β_{11}	β_{12}	β_0	β_2	β_{11}	β_{12}	
	1.30	1.30	-0.0502	0.0161	0.0006	-0.2195	0.0653	0.0161	0.0006	-0.2195	
			0.0922	0.0211	0.0801	0.1227	0.0921	0.0211	0.0801	0.1227	
	1.40	1.40	0.0007	-0.0144	-0.0112	-0.2527	0.1070	-0.0143	-0.0111	-0.2527	
			0.0956	0.0280	0.0900	0.1383	0.0972	0.0280	0.0900	0.1384	
	1.50	1.50	-0.0002	-0.0238	0.0266	-0.2801	0.1248	-0.0238	0.0266	-0.2801	
			0.0649	0.0224	0.0754	0.1247	0.0729	0.0224	0.0754	0.1247	
0.5	1.15	1.15	-0.0595	0.0549	0.0036	-0.2409	0.0332	0.0549	0.0035	-0.2409	
			0.0953	0.0293	0.0636	0.1129	0.0956	0.0293	0.0636	0.1129	
	1.30	1.30	-0.0624	0.0382	0.0598	-0.2436	0.0321	0.0382	0.0598	-0.2436	
			0.0809	0.0257	0.0640	0.1267	0.0823	0.0257	0.0640	0.1267	
	1.40	1.40	-0.0855	0.0329	0.0340	-0.2094	0.0272	0.0329	0.0340	-0.2094	
			0.0795	0.0206	0.0664	0.0997	0.0673	0.0206	0.0664	0.0998	
	1.50	1.50	-0.0405	0.0348	-0.0154	-0.2811	0.0656	0.0348	-0.0154	-0.2811	
			0.0843	0.0308	0.0442	0.1234	0.0753	0.0308	0.0442	0.1234	
	0.0	0.0	0.0	-0.0473	-1.2340	0.0632	0.0101	-0.0006	-0.2996	0.0629	0.00098
				0.1342	100.75	0.1863	0.1446	0.1185	0.6247	0.1862	0.1443

Table 4

–log (p-value) of the 10 most significant SNPs for the serum total bilirubin GWAS dataset.

Gene	without side information	with side information
rs887829	8.408	8.752
rs10929302	8.267	8.614
rs6753320	8.052	8.320
rs6753569	7.947	8.228
rs10168155	7.773	7.931
rs6736508	7.624	7.825
rs6736743	7.503	7.694
rs11680450	7.495	7.607
rs10171367	7.481	7.561
rs12623271	7.476	7.540