Shrinkage Estimators for Robust and Efficient Inference in Haplotype-Based Case-Control Studies

Yi-Hau Chen, **Nilanjan Chatterjee**, and **Raymond J. Carroll**

Yi-Hau Chen is Associate Research Member with the Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, Republic of China (E-mail: yhchen@stat.sinica.edu.tw). Nilanjan Chatterjee is Chief and Principle Investigator with the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institute of Health, Department of Health and Human Services, Rockville Maryland 20852 (E-mail: chattern@mail.nih.gov). Raymond J. Carroll is Distinguished Professor with the Department of Statistics, Texas A&M University, College Station, Texas 77843−3143 (E-mail: carroll@stat.tamu.edu).

Abstract

Case-control association studies often aim to investigate the role of genes and gene-environment interactions in terms of the underlying haplotypes (i.e., the combinations of alleles at multiple genetic loci along chromosomal regions). The goal of this article is to develop robust but efficient approaches to the estimation of disease odds-ratio parameters associated with haplotypes and haplotype-environment interactions. We consider "shrinkage" estimation techniques that can adaptively relax the model assumptions of Hardy-Weinberg-Equilibrium and gene-environment independence required by recently proposed efficient "retrospective" methods. Our proposal involves first development of a novel retrospective approach to the analysis of case-control data, one that is robust to the nature of the gene-environment distribution in the underlying population. Next, it involves shrinkage of the robust retrospective estimator toward a more precise, but model-dependent, retrospective estimator using novel empirical Bayes and penalized regression techniques. Methods for variance estimation are proposed based on asymptotic theories. Simulations and two data examples illustrate both the robustness and efficiency of the proposed methods.

Keywords

Empirical Bayes; Genetic epidemiology; LASSO (Least Absolute Shrinkage and Selection Operator); Model averaging; Model robustness; Model selection

1. INTRODUCTION

Haplotypes, the combinations of alleles at multiple loci along individual homologous chromosomes, define the functional units of a gene through which the underlying protein product is made (Clark 2004). Association studies based on haplotypes, which can capture interloci interactions as well as "indirect association" due to linkage disequilibrium (LD) with unobserved causal variants, can be a powerful approach to the discovery and characterization of the genetic basis of complex diseases (Schaid 2004). Thus, in recent years, there has been tremendous interest in developing methods for haplotype-based regression analysis of genetic epidemiologic data. A technical problem has been that traditional epidemiologic studies only collect locus-specific genotype data, which does not provide the "phase information", that is, which alleles appear at multiple loci along the individual chromosomes. Statistically, the lack of phase information can be viewed as a special missing data problem.

For logistic regression analysis of unmatched case-control studies, two classes of methods have evolved. The "prospective" methods (Schaid, Rowland, Tines, Jacobson, and Polalnd 2002; Zhao, Li, and Khalid 2003; Lake et al. 2003) ignore the underlying retrospective nature of the case-control design. These methods are considered robust in the sense that they depend very weakly on the underlying assumptions of Hardy-Weinberg equilibrium (HWE) and gene-environment (*G-E*) independence, although the assumptions cannot be totally avoided because of the phase ambiguity problem. In contrast, "retrospective" methods (Epstein and Satten 2003; Stram et al. 2003; Satten and Epstein 2004; Spinka, Carroll, and Chatterjee 2005; Lin and Zeng 2006), which properly account for case-control sampling, can fully exploit the assumptions of HWE and *G-E* independence to gain major efficiency over the prospective methods. It is often debatable which of the two types of methods is more suitable for a particular study. Prospective estimates of haplotype-effects and haplotype-environment interactions involving relatively rare haplotypes often tend to be very imprecise. Retrospective methods can produce much more precise estimates of those parameters, but concern often remains about the potential for bias because of the possible violation of the underlying assumptions, a potential we see in our simulations.

The potential for bias in retrospective methods can be reduced by flexible modeling approaches that relax the underlying assumptions. Alternative population genetic models that can relax the HWE assumptions have been used for retrospective haplotype analysis of case-control data (Satten and Epstein 2004; Lin and Zeng 2006). It has been also shown that the assumption of *G-E* independence can be relaxed to a large extent by assuming haplotypes are independent of *E* given unphased genotypes, but allowing the conditional distribution of *E* given the unphased genotypes to remain completely unrestricted (Lin and Zeng 2006). These solutions, which can alleviate the concern about bias, are not completely satisfactory. First, models for relaxing the HWE assumption can capture only certain types of departures from the underlying constraints for the diplotype distribution and may not be able to model phenomena such as excess heterozygosity. Second, even if a completely nonparametric model for the *G-E* distribution is available, we may still be able to gain efficiency in analysis of case-control data by exploiting the fact that HWE and G-E independence often do hold, approximately if not exactly. In the existing methods, if one uses a very general model for the distribution of *G-E*, then the concern about bias will be minimized, but inevitably efficiency will be lost.

Our main objective is to develop methods for haplotype-based analysis of case-control studies, which can gain efficiency by exploiting model assumptions of HWE and *G-E* independence for the underlying population and yet are resistant to bias when those model assumptions are violated. The basic idea involves shrinkage of a "model-free" estimator that is robust to HWE and *G-E* independence toward a "model-based" estimator that directly exploits those assumptions. The amount of "shrinkage" is sample size and data adaptive so that in large samples the method has no bias whether the assumptions of HWE and *G-E* independence hold, and yet the method can gain efficiency by shrinking the analysis toward HWE and *G-E* independence, but only to the extent the data validates the assumptions.

There are several novel aspects of our proposal. First, in Section 2.2, we propose a novel retrospective likelihood approach to haplotype analysis of case-control data that is robust to the nature of the gene-environment distribution in the underlying population. Second, in Section 3, we develop an empirical Bayes (EB)-type shrinkage estimation approach and a rigorous asymptotic theory for it. The key difficulty is that the problem is semiparametric, in that there are infinite-dimensional nuisance parameters associated with the joint distribution of the gene and the environment. Our method overcomes this difficulty by focusing only on the parameters of interest. In Section 4, we develop a penalized likelihood approach and asymptotic theory for it. The penalized likelihood involves shrinkage, not of a parameter or set of parameters to zero as is usually done, but to a model-based estimator, and also overcomes

the problem of infinite-dimensional nuisance parameters. Effectively, we try to shrink the difference of the model-free and model-based estimators toward zero. In Sections 5 and 6, we use simulation studies and two real data examples to illustrate that unlike the existing haplotype-based regression methods, whose utility depends crucially on specific model assumptions, the proposed shrinkage methods adapt themselves to a wide range of situations.

Finally, although our scientific focus of this article is haplotype-based case-control studies, this article makes a far more general contribution. Using modern shrinkage and penalization techniques to combine assumption-laden and assumption-free methods in semiparametric problems with infinite dimensional nuisance parameters is an idea that transcends genetic association studies. We hope that our article will lead to further research in this more general area.

2. A MODEL-BASED AND A MODEL-FREE ESTIMATOR

Let $\mathbf{H} = (\mathbf{H}_a, \mathbf{H}_b)$ denote the diplotype status (haplotype pair) for a subject at M loci of interest within a genomic region. Given $\mathbf{H}$ and a set of environmental covariates, $\mathbf{X}$, assume that the risk of a binary disease outcome D is given by the logistic regression model

$$\text{logit} \left\{ \text{pr}_\beta (D=1|\mathbf{H},\mathbf{X}) \right\} = \beta_0 + m(\mathbf{H},\mathbf{X};\beta_1), \tag{1}$$

where $m(\cdot)$ is a known but arbitrary function that specifies the log-odds-ratio of the disease as a function of $\mathbf{H}$ and $\mathbf{X}$ in terms of a set of regression parameters $\boldsymbol{\beta}_1$. In (1), the effect of a diplotype can be further specified in terms of the effect of the constituent haplotypes assuming *dominant, recessive*, or *additive* modes of penetrance (Wallenstein, Hodge, and Weston 1998). Let $\mathbf{G} = (g_1, \ldots, g_M)$ denote the unphased genotype data for the M loci. As explained earlier, the genotype data $\mathbf{G}$ could be consistent with multiple diplotypes because of phase ambiguity. We denote $\mathcal{H}_\mathbf{G}$ to be the set of all possible diplotypes that are consistent with the genotype data $\mathbf{G}$. Let $F(\mathbf{X}, \mathbf{G})$ be the cumulative distribution function for $\mathbf{X}$ and $\mathbf{G}$ in the underlying population.

Assume data on $\mathbf{G}$ and $\mathbf{X}$ are collected in a case-control study for N_0 controls ($D = 0$) and N_1 cases ($D = 1$). Let $N = N_0 + N_1$. The fundamental likelihood for case-control data, known as the "retrospective" likelihood, is given by

$$
L_{\text{Haplo}}^R = \prod_{i=1}^N \text{pr} (\mathbf{G}_i, \mathbf{X}_i | D_i)
$$

$$
= \prod_{i=l}^N \frac{\left\{ \sum_{\mathbf{H} \in \mathcal{H}_{\mathbf{G}_i}} \text{pr}_\beta(D_i|\mathbf{H}_i, \mathbf{X}_i) \text{pr}(\mathbf{H}_i|\mathbf{X}_i, \mathbf{G}_i) \right\} dF(\mathbf{X}_i, \mathbf{G}_i)}{\int_X \sum_G \left\{ \sum_{\mathbf{H} \in \mathcal{H}_{\mathbf{G}_i}} \text{pr}_\beta(D_i|\mathbf{H}, \mathbf{X}) \text{pr}(\mathbf{H}|\mathbf{X}, \mathbf{G}) \right\} dF(\mathbf{X}, \mathbf{G})}, \tag{2}
$$

where the last expression follows by Bayes theorem and the identity that

$$
\text{pr} (D|\mathbf{X}, \mathbf{G}) = \sum_{\mathbf{H} \in \mathcal{H}_G} \text{pr}_\beta (D|\mathbf{H}, \mathbf{X}) \, \text{pr} (\mathbf{H}|\mathbf{X}, \mathbf{G}),
$$

2.1 A Model-Based Framework

First, let us consider obtaining a "model-based" estimator for β_1 under the assumption that **H** and **X** are independent and that the distribution of **H** follows HWE in the underlying population. Under these assumptions, the joint density function for **X** and **G** is

$$dF(\mathbf{X},\mathbf{G}) = dF(\mathbf{X}) \times q(\mathbf{G}),$$ (3)

where $F(\mathbf{X})$ is the marginal distribution function for $\mathbf{X}, q(\mathbf{G}) = \sum_{\mathbf{H} \in \mathcal{H}_G} \mathrm{pr}(\mathbf{H})$, and pr(**H**) is the population frequency of the diplotype **H**. Under HWE for haplotypes, we have

$$\mathrm{pr}_\theta\{\mathbf{H} = (\mathbf{h}_a, \mathbf{h}_b)\} \quad \begin{aligned} &= \theta_\alpha^2 \quad \text{if} \quad h_a = h_b \\ &= 2\theta_a \theta_b \quad \text{if} \quad h_a \neq h_b, \end{aligned}$$

where θ_s denotes the population frequency for the haplotype $\mathbf{h}_s$. Under the hapotype-environment (*H-X*) independence, we also have

$$\mathrm{pr}(\mathbf{H}|\mathbf{X},\mathbf{G}) = \mathrm{pr}(\mathbf{H}|\mathbf{G}) = I(\mathbf{H} \in \mathcal{H}_\mathbf{G}) \, \mathrm{pr}_\theta(\mathbf{H}) / \sum_{\mathbf{H}_* \in \mathcal{H}_G} \mathrm{pr}_\theta(\mathbf{H}_*).$$ (4)

Spinka et al. (2005) showed how to estimate β and θ by maximization of the retrospective likelihood (2) under HWE and the *H-X* independence, while allowing $F(\mathbf{X})$ to remain completely unrestricted, using a computationally simple "profile-likelihood" approach, see also (6) later. This forms the "model-based" approach.

2.2 A Model-Free Framework

Now consider obtaining a "model-free" estimator. Unfortunately, in the presence of phase ambiguity, β and θ are not identifiable from the retrospective likelihood (2) if the joint distribution of **X** and **H** is left completely unrestricted (e.g., see Epstein and Satten 2003;Lin and Zeng 2006). We propose to resolve this identifiability issue by making minimal distributional assumptions. We note that, given that **X** and **G** are directly observed, the joint distribution function $dF(\mathbf{X}, \mathbf{G})$ should be estimable nonparametrically, even though the joint distribution of **X** and **H** is not. Thus, β_1 should be identifiable from (2) with some constraints on the conditional distribution of **H** given **G** and **X**. We propose to use HWE and *H-X* independence constraints to specify the *conditional distribution* pr(**H**|<u>**G**</u>, **X**) [i.e., instead of (3) we assume only that (4) holds]. Lin and Zeng (2006) used a similar approach to allow the conditional distribution of pr(**X**|**G**) to be completely unrestricted, but they essentially imposed HWE or related population genetics model constraints not only on pr(**H**|**G**), but also on pr (**G**). Our method can allow the marginal distribution pr(**G**) to remain completely unrestricted and thus is even more robust.

To see why (4) involves very mild assumptions, note that if there were no phase ambiguity (i.e., **G** = **H**), then this formulation does not impose any restriction on the population distribution of the covariates of the logistic regression model (1). In this case, it follows from classical theory (Prentice and Pyke 1979) that the retrospective maximum-likelihood estimate of β_1 could be obtained using standard prospective logistic regression analysis, the validity of which does not depend on assumptions for the covariate distribution. In contrast, the validity of Lin and Zeng's estimator does depend on the assumed diplotype distribution.

In the presence of phase ambiguity, violation of HWE and *H-X* independence for the underlying population would imply that the corresponding constraint for the distribution pr(**H**|**X, G**) will also not hold. However, in typical association studies involving tightly linked loci, the problem of phase ambiguity tends to be modest and the misspecification of the conditional distribution in such situations will have a fairly small influence on inference on the regression parameters in our model-free framework, see Section 5 for simulation results.

2.3 ESTIMATORS

2.3.1 The Model-Free Estimator—We next develop an algorithm for obtaining the semiparametric maximum likelihood estimator of β_1, one that maximizes the retrospective likelihood (2) under the assumption (4), allowing $F(\mathbf{G}, \mathbf{X})$ to remain completely nonparametric. We consider a profile-likelihood approach analogous to that described in Chatterjee and Carroll (2005),Spinka et al. (2005) and Chatterjee and Chen (2007). In particular, assuming that the nonparametric maximum likelihood estimator of $F(\mathbf{X}, \mathbf{G})$ allows masses only on the observed data points, following arguments analogous to Spinka et al. we can show that the estimator of β_1, which maximizes the retrospective likelihood (2), can be obtained from an alternative pseudo-likelihood of the data where the contribution of each subject is given by

$$L_{\text{free}}(D,\mathbf{G},\mathbf{X},\Omega) = \frac{\sum\limits_{\mathbf{h}\in\mathcal{H}_G} q_{\text{free}}(\mathbf{h}|\mathbf{G},\theta)\,\mathcal{M}(D,\mathbf{h},\mathbf{X},\mathbf{G},\Omega)}{\sum\limits_{s=0}^{1}\sum\limits_{\mathbf{h}\in\mathcal{H}_G} q_{\text{free}}(\mathbf{h}|\mathbf{G},\theta)\,\mathcal{M}(s,\mathbf{h},\mathbf{X},\mathbf{G},\Omega)},$$

(5)

where $q_{\text{free}}(\mathbf{h}|\mathbf{G}, \theta)$ denotes pr(**H**|**G**), computed according to (4), and

$$\mathcal{M}(d,\mathbf{h},\mathbf{x},\mathbf{g},\Omega) = \frac{\exp\left[d\{\kappa+m(\mathbf{h},\mathbf{X},\beta_1)\}\right]}{1+\exp\{\beta_0+m(\mathbf{h},\mathbf{X},\beta_1)\}},$$

$p_d = N_d/N$, $\pi_d = \text{pr}(D=d)$, $\kappa = \beta_0 + \log(p_1/p_0) - \log(\pi_1/\pi_0)$, and $\Omega = (\beta_0, \beta_1, \theta, \kappa)$. Under a rare disease assumption, $\mathcal{M}(d,\mathbf{h},\mathbf{x},\mathbf{g},\Omega) \approx \exp\left[d\{\kappa+m(\mathbf{h},\mathbf{X},\beta_1)\}\right]$, β_0 is not identifiable and Ω no longer contains β_0.

Note that $L_{\text{free}}(D, \mathbf{G}, \mathbf{X}, \Omega)$ will contain little information on θ, because it conditions on **G**. Thus, when implementing methods based on this likelihood, we replace the score function for θ by the estimating function for θ based on the genotype data from the controls and the HWE assumption. It can be seen that, when using such an estimating function for θ and the rare disease approximation mentioned previously, the estimator obtained from the retrospective likelihood $L_{\text{free}}(D, \mathbf{G}, \mathbf{X}, \Omega)$ is equivalent to that from the prospective approach proposed by Zhao et al. (2003).

2.3.2 The Model-Based Estimator—We note that the profile-likelihood estimator Spinka et al. (2005) derived for maximization of (2) under the assumptions of HWE and *H-X* independence corresponds to a pseudo-likelihood where the contribution of each subject is given by

$$L_{\text{model}}(D,\mathbf{G},\mathbf{X},\Omega) = \frac{\sum\limits_{\mathbf{h}\in\mathcal{H}_G} q(\mathbf{h};\theta)\,\mathcal{M}(D,\mathbf{h},\mathbf{X},\mathbf{G},\Omega)}{\sum\limits_{s=0}^{1}\sum\limits_{\mathbf{h}} q(\mathbf{h};\theta)\,\mathcal{M}(s,\mathbf{h},\mathbf{X},\mathbf{G},\Omega)},$$

(6)

where $q(\mathbf{h}; \boldsymbol{\theta})$ denotes $\mathrm{pr}(\mathbf{H} = \mathbf{h})$ computed according to HWE.

2.3.3 An Alternative Characterization—Interestingly, the pseudo-likelihoods for both the "model-based" and the "model-free" estimators can be derived as a proper likelihood under an alternative sampling design, wherein a case-control study can be viewed as a prospective study with missing data. Consider a sampling scenario where each subject from the underlying population is selected into the case-control study using a Bernoulli sampling scheme where the selection probability for a subject given his or her disease status $D = d$ is proportional to $\mu_d = N_d/\mathrm{pr}(D = d)$. Let $R = 1$ denote the indicator of whether a subject is selected in the case-control sample under this Bernoulli sampling scheme and hence has been observed. Under this sampling scheme, it is easy to show that

$$\begin{aligned} L_{\mathrm{model}}(D,\mathbf{G},\mathbf{X},\Omega) &= \mathrm{pr}(D,\mathbf{G}|\mathbf{X},R=1) \quad \text{and} \\ L_{\mathrm{free}}(D,\mathbf{G},\mathbf{X},\Omega) &= \mathrm{pr}(D|\mathbf{G},\mathbf{X},R=1). \end{aligned}$$

The proof for the former identity can be found in Spinka et al. (2005) and that for the latter identity follows similarly. Thus, in this alternative sampling scheme, the difference between the model-based and model-free estimators corresponds to whether the genotype data has been conditioned out of the likelihood or not.

As discussed in Chatterjee and Carroll (2005), in a case-control sample, it may be difficult to estimate the intercept parameter b_0 even when the haplotype-environment independence assumption is imposed. However, the estimation of b_0 can be avoided by imposing the rare-disease assumption so that the parameters in effect are $(\boldsymbol{k}, \boldsymbol{\beta}_1, \boldsymbol{\theta})$. In the following discussion, we will adopt this convention and redefine the regression parameters $\boldsymbol{\beta} = (\boldsymbol{k}, \boldsymbol{\beta}_1)$.

3. EMPIRICAL BAYES-TYPE SHRINKAGE ESTIMATORS

Once we obtain two estimators of $\boldsymbol{\beta}$, we propose to combine them using EB-type weighting in the spirit of Mukherjee and Chatterjee (2008). Previously, we have developed a general theory for obtaining such weighted estimators when the departure of the population distribution of the risk-factors from the underlying models can be indexed by a finite set of parameters. In the current setting, however, the departure of the nonparametric density $dF(\mathbf{X}, \mathbf{G})$ from the restricted density $dF_0(\mathbf{X},\mathbf{G}) = dF(\mathbf{X})\sum_{\mathbf{H}\in\mathcal{H}_G}\mathrm{pr}_\theta(\mathbf{H})$ cannot be indexed by a fixed set of parameters. Thus, we propose constructing the EB-type shrinkage estimator directly in terms of the focus parameters of interest, namely $\boldsymbol{\beta}$, rather than both $\boldsymbol{\beta}$ and the nuisance parameters $\boldsymbol{\theta}$.

Let $\boldsymbol{\beta}_{\mathrm{free}}$ and $\boldsymbol{\beta}_{\mathrm{model}}$ denote the asymptotic limit of model-free and model-based estimators proposed previously. Note that when HWE and $(G\text{-}E)$ independence hold, we have $\boldsymbol{\psi} = \boldsymbol{\beta}_{\mathrm{model}} - \boldsymbol{\beta}_{\mathrm{free}} = 0$. Thus, if we want to relax this assumption, we can use a stochastic framework where we assume $\psi \sim \mathrm{Normal}(0,Y)$ and note that $\widehat{\psi\psi^\mathsf{T}} = (\widehat{\beta}_{\mathrm{model}} - \widehat{\beta}_{\mathrm{free}})(\widehat{\beta}_{\mathrm{model}} - \widehat{\beta}_{\mathrm{free}})^\mathsf{T}$ is a conservative estimate of $\mathbf{Y}$, conservative in the sense that its mean is greater than $\mathbf{Y}$ (a matrix $\mathbf{A}$ is defined as greater than a matrix $\mathbf{B}$ when $\mathbf{A} - \mathbf{B}$ is semi-positive definite). Define a shrinkage factor given by the matrix

$$\mathbf{K} = \mathbf{V}\left(\mathbf{V} + \widehat{\psi\psi^\mathsf{T}}\right)^{-1},$$

where $\mathbf{V}$ is the (estimated) variance-covariance matrix of $\widehat{\psi}$. By this logic, we can construct an EB-type estimator

$$\widehat{\beta}_{\text{EB}} = \widehat{\beta}_{\text{free}} + \mathbf{K}\left(\widehat{\beta}_{\text{model}} - \widehat{\beta}_{\text{free}}\right), \tag{7}$$

We observe that Equation (7) suggests a general way of constructing simple shrinkage estimators. The shrinkage factor $\mathbf{K}$ determines the amount of shrinkage of the model-free estimator toward its model-based counterpart, with the two extremes being $\mathbf{K} = \mathbf{I}$ (identity matrix) implying $\widehat{\beta}_{\text{EB}} = \widehat{\beta}_{\text{model}}$ and $\mathbf{K} = 0$ implying $\widehat{\beta}_{\text{EB}} = \widehat{\beta}_{\text{free}}$. If the estimator is to be approximately consistent in large samples, whether the HWE and (*G-E*) independence assumptions hold or not, the matrix $\mathbf{K}$ should go to zero, at least when $\boldsymbol{\beta}_{\text{model}} \neq \boldsymbol{\beta}_{\text{free}}$, as sample size increases. Moreover, if $\mathbf{K}$ goes to zero at a suitable rate, $\widehat{\beta}_{\text{EB}}$ can be asymptotically equivalent to the model-free estimator, but in finite samples and when the difference between $\boldsymbol{\beta}_{\text{model}}$ and $\boldsymbol{\beta}_{\text{free}}$ is small, which might often be the case in practice, then $\widehat{\beta}_{\text{EB}}$ can still have better finite sample performance in terms of the bias-variance trade-off. Simulation results in Section 5 will illustrate this feature. It is intuitive that $\mathbf{K}$ should be such that more weight should be given to $\widehat{\beta}_{\text{model}}$ or $\widehat{\beta}_{\text{free}}$ depending on the bias of the model-based estimator, a quantity that can be estimated empirically as $\widehat{\psi} = \widehat{\beta}_{\text{model}} - \widehat{\beta}_{\text{free}}$. In Sections 5 and 6, we will also consider one such alternative EB-type shrinkage estimator where we choose $\mathbf{K}$ to be a diagonal matrix with the ith diagonal element being $k_i = v_i / \left(v_i + \widehat{\psi}_i^2\right)$, where v_i is the ith diagonal element of $\mathbf{V}$ and $\widehat{\psi}_i$ is the ith component of $\widehat{\psi}$.

3.1 Asymptotic Theory

When the assumptions of HWE or *H-X* independence or both are violated so that $\boldsymbol{\beta}_{\text{model}} \neq \boldsymbol{\beta}_{\text{free}}$, it is easy to show that the EB estimator is asymptotically equivalent to $\widehat{\beta}_{\text{free}}$, and hence is consistent for $\boldsymbol{\beta}$ (see Appendix C). Nevertheless, utilizing the bias-variance trade-offs between $\widehat{\beta}_{\text{model}}$ and $\widehat{\beta}_{\text{free}}$, the EB estimator we propose can have substantially better finite sample properties than the model-free estimator $\boldsymbol{\beta}_{\text{free}}$ (see Tables 2 and 3; see also Section 5 for simulation details). Moreover, using the δ-method, in Appendix C we derive an approximate covariance matrix estimator for the EB estimator, which is found to be more accurate in finite samples than the "naïve" covariance estimator obtained by the covariance matrix of the asymptotically equivalent model-free estimator.

Now we consider the asymptotic theory for the EB-type estimator (7) when HWE and (*G-E*) independence hold, which implies $\boldsymbol{\beta}_{\text{model}} = \boldsymbol{\beta}_{\text{free}}$ (i.e., the model-based estimator is consistent). Let $\boldsymbol{\Psi}_{\text{model}}(D, \mathbf{G}, \mathbf{X}, \boldsymbol{\Omega}_{\text{model}})$ and $\boldsymbol{\Psi}_{\text{free}}(D, \mathbf{G}, \mathbf{X}, \boldsymbol{\Omega}_{\text{free}})$ be the individual score/estimating functions for the model-based and model-free estimators. These are the derivatives of the logarithms of (5) and (6), respectively, with respect to the parameters, with the exception that in the model-free case, the score for $\boldsymbol{\theta}$ is as described in Section 2.3.1.

Let $\mathcal{I}_{\text{model}}$ be minus the expectation of $N^{-1}\sum_{i=1}^{N}\partial\Psi_{\text{model}}(D_i, \mathbf{G}_i, \mathbf{X}_i, \boldsymbol{\Omega}_{\text{model}})/\partial\boldsymbol{\Omega}_{\text{model}}^{\mathsf{T}}$ and let $\mathcal{I}_{\text{free}}$ be defined analogously. Let $\Rightarrow$ mean convergence in distribution to a random variable having the same distribution as the right hand side. It is a consequence of Theorems 2 and 3 in the Appendix B that

$$\left\{N^{1/2}\left(\widehat{\beta}_{\text{model}} - \beta_{\text{model}}\right)^{\mathsf{T}}, N^{1/2}\left(\widehat{\beta}_{\text{free}} - \beta_{\text{free}}\right)^{\mathsf{T}}\right\}^{\mathsf{T}} \Rightarrow \left(\mathcal{Z}_{\text{model}}^{\mathsf{T}}, \mathcal{Z}_{\text{free}}^{\mathsf{T}}\right)^{\mathsf{T}} \sim \text{Normal}\left(0, \Sigma_{\beta}\right), \text{ where}$$

$$\Sigma_\beta = \mathrm{cov} \left\{ \begin{array}{l} \left(\mathbf{I}_p \; 0\right) \mathcal{I}_{\mathrm{model}}^{-1} N^{-1/2} \sum_{i=1}^{N} \Psi_{\mathrm{model}} \left(D_i, \mathbf{G}_i, \mathbf{X}_i, \Omega_{\mathrm{model}}\right) \\ \left(\mathbf{I}_p \; 0\right) \mathcal{I}_{\mathrm{free}}^{-1} N^{-1/2} \sum_{i=1}^{N} \Psi_{\mathrm{free}} \left(D_i, \mathbf{G}_i, \mathbf{X}_i, \Omega_{\mathrm{free}}\right) \end{array} \right\},$$

with $\mathbf{I}_p$ the identity matrix of size $p = \mathrm{dim}(\boldsymbol{\beta})$, and $\mathbf{0}$ the matrix of zeros of size $p \times q$, $q = \mathrm{dim}$ $(\boldsymbol{\theta})$. Define $\mathcal{V} = \left(\mathbf{I}_p, -\mathbf{I}_p\right) \Sigma_\beta \left(\mathbf{I}_p, -\mathbf{I}_p\right)^\mathsf{T}$ and define

$\mathcal{M}(\mathcal{Z}_{\mathrm{model}}, \mathcal{Z}_{\mathrm{free}}) = \mathcal{V} \{ \mathcal{V} + (\mathcal{Z}_{\mathrm{model}} - \mathcal{Z}_{\mathrm{free}})(\mathcal{Z}_{\mathrm{model}} - \mathcal{Z}_{\mathrm{free}})^\mathsf{T} \}^{-1}$. Then, it follows immediately that when HWE and (G-E) independence hold,

$$N^{1/2} \left(\widehat{\beta}_{\mathrm{EB}} - \beta_{\mathrm{free}}\right) \Rightarrow [\mathcal{M}(\mathcal{Z}_{\mathrm{model}}, \mathcal{Z}_{\mathrm{free}}),$$
$$\mathbf{I} - \mathcal{M}(\mathcal{Z}_{\mathrm{model}}, \mathcal{Z}_{\mathrm{free}})] \left(\mathcal{Z}_{\mathrm{model}}^\mathsf{T}, \mathcal{Z}_{\mathrm{free}}^\mathsf{T}\right)^\mathsf{T}. \qquad (8)$$

Of course, the limit distribution in (8) is not normally distributed, a phenomenon that is expected for many model-average estimators at the null or reduced model (Hjort and Claeskens 2003; Claeskens and Carroll 2007). However, simulation studies show that the lack of normality, while real, is not serious in practice in our context. The quantile-quantile plots shown in Figure 1 for comparing the distribution of the EB-type estimates from a set of simulations with the normal distribution illustrate this fact (see Section 5 for details about simulations). Moreover, the EB estimator in this situation can be more efficient than the model-free estimator not only in finite samples, but also in large samples (Tables 2 and 3, the blocks with $f = 0$, $\gamma_{1,3} = 0$). Such a phenomenon of "super-efficiency", first observed by Hodges (e.g., see Lehmann 1983, pp. 405−406), is expected for many shrinkage estimators. Finally, we note that variance estimators we propose in the case of $\boldsymbol{\beta}_{\mathrm{model}} \neq \boldsymbol{\beta}_{\mathrm{free}}$ provide remarkably good approximations for the variance of the EB estimators even when $\boldsymbol{\beta}_{\mathrm{model}} = \boldsymbol{\beta}_{\mathrm{free}}$, although in the latter case the derivation of the asymptotic variances based on the δ-method is not strictly accurate, but it is acceptable in practice.

4. PENALIZED LIKELIHOODS

A further consideration of (7) suggests an entirely different approach to combining assumption-laden and assumption-free methods in semiparametric problems with infinite dimensional nuisance parameters. Specifically, setting $\mathbf{K}^* = \mathbf{I} - \mathbf{K}$, we can rewrite (7) as

$$\widehat{\beta}_{\mathrm{EB}} = \widehat{\beta}_{\mathrm{model}} + \mathbf{K}^* \left(\widehat{\beta}_{\mathrm{free}} - \widehat{\beta}_{\mathrm{model}}\right). \qquad (9)$$

Because $\mathbf{K}^* = 0$ leads to the model-based estimator and $\mathbf{K}^* = \mathbf{I}$ leads to the model-free estimator, one interpretation of (9) is that one is shrinking the difference between the model-free and model-based estimators toward zero. With $\mathbf{K}^* = 0$ being interpreted as full shrinkage, a more useful interpretation of (9) is that one is shrinking the model-free estimator toward the model-based estimator when it is appropriate to do so.

With this idea in mind, we now turn to penalized likelihoods, which are also used for constructing shrinkage estimators. However, most of these proposals shrink parameters to zero. Based on the intuition of the EB-type estimators described in the previous paragraph, we instead propose shrinking the model-free estimator toward the model-based estimator (or equivalently, shrinking their difference to zero) by maximization of a penalized log-likelihood of the form

$$l_p = \sum_{i=1}^{N} \log \ \{L_{\mathrm{free}}(D_i, \mathbf{G}_i, \mathbf{X}_i, \Omega)\}$$
$$- \sum_{j=1}^{p} \lambda_j P\left(\beta_j, \widehat{\beta}_{\mathrm{model},j}\right),$$

(10)

where $\widehat{\beta}_{\mathrm{model},j}$ denotes the estimate of β_j from the model-based likelihood L_{model}, $P(\cdot)$ denotes a suitable penalty function and the λ-values denote parameter-specific penalties that determine the degree of shrinkage.

There are two unique features of (10). First, we propose shrinkage directly with respect to the focus parameters of interest. Given that we are trying to exploit *G-E* independence and HWE, which are assumptions about nuisance parameters, it may seem more natural that one maximizes a penalized likelihood where the penalties are given to the nuisance parameters. However, in a semiparametric model involving infinite dimensional nuisance parameters, this could become a challenging task. By formulating the problem with respect to the focus parameters themselves, however, we can simply work with the profile likelihoods that are completely free of the nuisance parameters. A second interesting feature of our proposal is that we are shrinking parameters, not toward some fixed values as is usually done, but toward the estimates obtained from an alternative model.

One can use both L_1 (LASSO) and L_2 (ridge) penalty functions, with the former given by $P\left(\beta_j, \widehat{\beta}_{\mathrm{model},j}\right) = |\beta_j - \widehat{\beta}_{\mathrm{model},j}|$ and the latter by $P\left(\beta_j, \widehat{\beta}_{\mathrm{model},j}\right) = \left(\beta_j - \widehat{\beta}_{\mathrm{model},j}\right)^2$. For logistic regression models, L_2 penalized likelihoods can be maximized by iteratively reweighted ridge regressions and hence can be implemented conveniently. An interesting feature of the L_1 penalty is that it can produce "sparse" solutions [i.e., we can have certain estimates exactly equal to those obtained from the HWE and *G-E* independence model (Tibshirani 1996)]. More detailed discussion of the respective properties for the two types of penalty functions can be found in Hastie, Tibshirani, and Friedman (2001).

Two issues are important for practical implementation of the penalized likelihood estimation: (a) computation and (b) the choice of the penalty parameters λ_j. In the Appendix D we deal with issue (a) and in what follows we deal with issue (b).

4.1 Choice of the Penalty Parameters

Both L_1 and L_2 penalized estimation can be interpreted as Bayesian posterior mode estimation (Tibshirani 1996), with the prior given by the Laplace and the normal distributions, respectively. Note that the penalty parameter λ_j in the penalized log-likelihood (10) has a $1-1$ correspondence to the variance τ_j of the prior distribution for $\psi_j = \beta_{\mathrm{model},j} - \beta_{\mathrm{free},j}$; the L_1 penalty corresponds a Laplace prior with variance $2/\lambda_j^2$, whereas the L_2 penalty corresponds to a normal prior with variance $1/(2\lambda_j)$. As in Section 3, we use $\widehat{\psi}_j^2 = \left(\widehat{\beta}_{\mathrm{model},j} - \widehat{\beta}_{\mathrm{free},j}\right)^2$ as a conservative estimate of τ_j. These facts in turn suggest that the choice of λ_j should be inversely proportional to $|\widehat{\psi}_j|$ and $\widehat{\psi}_j^2$ for the L_1 and L_2 penalized regression, respectively. Moreover, we would like to have the penalty parameters converge to zero in large samples, so that the resulting penalized estimators are asymptotically equivalent to the model-free estimator and hence are consistent, even when $\boldsymbol{\beta}_{\mathrm{model}} \neq \boldsymbol{\beta}_{\mathrm{free}}$ (i.e., HWE/*H-X* independence does not hold). Accordingly, we thus propose the following choices of the penalty parameters. For L_2-penalized regression we suggest using $\lambda_j = v_j / \left(2\widehat{\psi}_j^2\right)$, where v_j is the variance of $\widehat{\psi}_j$; for L_1-penalized regression, we

suggest $\lambda_j = \sqrt{v_j/2}/|\widehat{\psi}_j|$. It is readily seen that these choices of penalty parameters satisfy both the desired properties: they yield more shrinkage toward the model-based estimator when the magnitude of the estimated bias $\widehat{\psi}$ is smaller, and will converge to zero in large samples when $\psi = \boldsymbol{\beta}_{\text{model}} - \boldsymbol{\beta}_{\text{free}} \neq 0$ because $\boldsymbol{v}_j \to 0$ as $N \to \infty$.

5. SIMULATIONS

5.1 Set Up

We conducted simulation studies to examine the performance of the EB-type and penalized likelihood shrinkage estimators. We implemented two EB estimators, termed EB1 and EB2, one corresponding to "multivariate shrinkage" with $\mathbf{K} = \mathbf{V}\left(\mathbf{V} + \widetilde{\psi}\widetilde{\psi}^{\mathsf{T}}\right)^{-1}$, where $\mathbf{V}$ is the estimated variance-covariance of $\widehat{\psi} = \widehat{\boldsymbol{\beta}}_{\text{model}} - \widehat{\boldsymbol{\beta}}_{\text{free}}$, and the other corresponding to "componentwise shrinkage" with $\mathbf{K} = \text{diag}\left\{v_i/\left(v_i + \widehat{\psi}_i^2\right)\right\}$, where $\widehat{\psi}_i$ is the ith component of $\widehat{\psi}$ and $\boldsymbol{v}_i$ is the ith diagonal component of $\mathbf{V}$. We also implemented two penalized likelihood methods, PL1 and PL2, corresponding respectively to the L_1 and L_2 penalties, with the penalty parameters chosen as described in Section 4.1. Haplotype data were simulated using the haplotype patterns and frequencies (see Table 1) for five Single Nucleotide Polymorphisms (SNPs) along a diabetes susceptibility region on chromosome 22, reported in the Finland-United States Investigation of Non-Insulin-Dependent Diabetes Mellitus (FUSION) study (Epstein and Satten 2003). The environmental covariate X is a binary variable, with $\text{pr}(X = 1) = 0.3$.

In our simulations, given the environmental covariate X, we generated a 5-SNP diplotype for a subject according to the model

$$\log\left[\frac{\text{pr}\left\{\mathbf{H} = \left(\mathbf{h}_{j_1}, \mathbf{h}_{j_2}\right) | X\right\}}{\text{pr}\left\{\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_1) | X\right\}}\right] = \gamma_{0j_1j_2} + \left(\gamma_{1,j_1} + \gamma_{1,j_2}\right)X,$$

(11)

where $j_1, j_2 = 1, \ldots, 7$ index haplotypes, and the diplotype (h_1, h_1) is chosen as the reference. The parameters $\{\gamma_{0j_1j_2}\}$ are specified so that the marginal (unconditional on X) diplotype frequencies are given as

$$\text{pr}\left\{\mathbf{H} = \left(\mathbf{h}_{j_1}, \mathbf{h}_{j_2}\right)\right\} = \begin{cases} 2(1-f)\theta_{j_1}\theta_{j_2} & j_1 \neq j_2 \\ f\theta_{j_1} + (1-f)\theta_{j_1}^2 & j_1 = j_2 \end{cases}$$

(12)

where the θ_j are the haplotype frequencies given in Table 1, and f is the fixation index quantifying the departure from HWE, with $f = 0$ indexing HWE (Satten and Epstein 2004). The parameters $\{\gamma_{1,j}\}$ are all zero for $j \neq 3$, and $\gamma_{1,3}$ is set to 0 or -0.4, which quantifies the departure from haplotype-environment independence, with $\gamma_{1,3} = 0$ indexing independence. Hence, $f = 0$ and $\gamma_{1,3} = 0$ corresponds to an "ideal" model where both HWE and H-X independence hold. Given X and $\mathbf{H}$, disease status is generated from the model

$$\text{logit} \quad \text{pr}(D = 1 | \mathbf{H}, X) = \beta_0 + \beta_H \mathcal{H} + \beta_X X + \beta_{HX} \mathcal{H}X,$$

(13)

where $\mathcal{H}$ is coded as indicating whether a subject carries at least one copy of the causal haplotype "01100" ($j = 3$) (i.e., a dominant genetic model), or indicating whether a subject carries two copies of this haplotype (i.e., a recessive model). The parameter values $(\beta_0, \beta_H, \beta_X, \beta_{HX}) = (-3.0, 0.3, 0.3, 0.3)$. A case-control sample with N_1 cases and N_0 controls was

sampled from the simulated population. Once we generated the data in the previous fashion, we deleted the phase information.

In the simulations we compare the shrinkage estimators with the model-free and model-based estimators. We also consider the estimator $\widehat{\beta}_{\mathrm{HWD}}$ accounting for the Hardy-Weinberg disequilibrium of the form (12) but not accounting for *H-X* dependence, which is proposed in Lin and Zeng (2006).

5.2 Efficiency and Bias

Tables 2 and 3 display simulation results for the dominant and recessive models, respectively. We take the sample sizes to be $N_1 = N_2 = 150$, 300, or 600 for the dominant model and $N_1 = N_0 = 300$, 600 or 1,000 for the recessive model. To save space, we report only the results for β_H and β_{HX} for the causal haplotype "01100" ($j = 3$), although during the analysis we fitted a "full" model containing all haplotypes and their interactions with the environmental covariate. We make the following key observations.

When the assumptions of HWE and *H-X* independence are met ($f = 0$, $\gamma_{1,3} = 0$), all the different estimators are essentially unbiased. In this situation, the model-based estimator can be remarkably more efficient than the model-free estimator, especially when the genetic effect is recessive. The different shrinkage estimators give up some efficiency relative to the model-based estimator, but all of them, with the occasional exception of EB1 (multivariate empirical-Bayes shrinkage), achieve major gains in efficiency over the model-free estimators in small samples ($N_1 = N_0 = 150$ or 300). Moreover, EB2 (component-wise empirical-Bayes shrinkage) consistently retains substantial gains over the model-free estimator even for larger sample sizes such as $N_1 = N_0 = 600$ or $N_1 = N_0 = 1,000$. The estimator $\widehat{\beta}_{\mathrm{HWD}}$ assuming Hardy-Weinberg disequilibrium performs similarly to the model-based estimator.

When the assumptions of HWE or *H-X* independence or both are violated, the model-based estimator yields very large bias for the genetic main effect or the gene-environment interaction parameter or both. The model-free estimator, although it depends weakly on these same assumptions, has negligible bias. As a result, in these situations, the mean squared error (MSE) for the model-based estimator is often much larger than that of the model-free estimator when the violation of model assumptions is severe or the sample size is large. Note that when only the assumption of HWE is violated but *H-X* independence still holds ($f = 0.05$, $\gamma_{1,3} = 0$), the estimator $\widehat{\beta}_{\mathrm{HWD}}$ performs best among all the estimators considered, which is as expected because $\widehat{\beta}_{\mathrm{HWD}}$ assumes exactly the true model in this setting. When the *H-X* independence is violated, $\widehat{\beta}_{\mathrm{HWD}}$ performs similarly to the model-based estimator and has large bias and MSE. All of the shrinkage estimators adapt to the situation quite nicely and reduce the bias dramatically. The MSE of the shrinkage estimators usually remains much smaller than that of the model-free estimator when the sample size is small, whereas the performances of the shrinkage and model-free estimators become quite similar in large samples. Overall, among the various shrinkage estimators, the PL1 produces the smallest MSE in most cases under the dominant model, whereas the EB2 and PL2 produce the smallest MSE in most cases under the recessive model. The magnitude of the bias for all shrinkage estimators is similar, with that for the EB2 and PL1 estimators being the largest in most cases.

5.3 Variance Estimators

Variance estimators for the EB and penalized estimators are given in the Appendices C and F. In Tables given in the supplementary material and available from the first author, we have studied the performance of these variance estimators. We observed for each shrinkage estimator that the mean of the estimated variances over simulations was quite close to the

empirical variance of the shrinkage estimator. The variance estimators work remarkably well even when HWE and (*G-E*) independence assumptions are met, although in this situation the application of the δ-method is indeed not technically correct due to nonnormality of the limiting distribution.

6. CASE-CONTROL STUDIES OF COLORECTAL ADENOMA AND PROSTATE CANCER

In this section, we discuss results from two case-control data examples. The examples were chosen in such a way that from a priori biological grounds one would expect the gene-environment independence assumption to hold in one example, but probably not in the other. The two examples taken together illustrate how the different shrinkage estimators adapt to alternative scenarios for the gene-environment distribution.

6.1 Background of the Examples

The first application involves a case-control study of colorectal adenoma, a precursor of colorectal cancer. The study sample includes 628 prevalent advanced adenoma cases and 635 gender-matched controls, selected from the screening arm of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial at the National Cancer Institute, USA (Gohagan, Prorok, Hayes, and Kramer 2000; Moslehi et al. 2006). One of the main objectives of this study is to assess whether the smoking-related risk of colorectal adenoma may be modified by certain haplotypes in *NAT2*, a gene known to be important in the metabolism of smoking related carcinogens. In addition, because *NAT2* is involved in the smoking metabolism pathway, potentially it can influence an individual's addiction to smoking itself, causing the gene-environment independence assumption to be violated.

The second application involves a case-control study of prostate cancer. The sample includes 749 prostate cancer cases and 781 controls, also selected from the screening arm of the PLCO trial described earlier. The main objective of the study is to examine the relationship between risk of prostate cancer and [25(OH)D], a serum level biomarker of vitamin D, that reflects both dietary and sunlight exposures. The anticancer effect of vitamin D is hypothesized due to the ability of prostate cells to convert [25(OH)D]] into 1,25-dihydroxy-vitamin D [1,25(OH)2D], the most active form of this vitamin, which regulates the gene transcription of many proteins involving cellular differentiation, proliferation, and apoptosis via the vitamin D receptor (VDR). It is thus of interest to see whether genetic polymorphisms in the VDR gene modify the relationship between [25(OH)D]] and risk of prostate cancer. Given the "downstream" role of the VDR gene in the vitamin-D pathway, it is very unlikely that these polymorphisms actually could influence the level of the [25(OH)D] itself. Thus, the gene-environment independence assumption in this application is likely to be valid.

In the colorectal adenoma study, genotype data were available on six SNPs, for which seven common haplotypes (with estimated frequency > 0.5%) are inferred by the EM algorithm of Li, Khalid, Carlson, and Zhao (2003). In the prostate cancer study, genotype data were available on three SNPs, for which four common haplotypes (with estimated frequency > 0.5%) are inferred by the expectation maximization (EM) algorithm. Our association analysis is performed by fitting the logistic regression model (13). The haplotype covariate vector $\mathcal{H}$ is coded by a multiplicative rule (i.e., the *h*th component of $\mathcal{H}$ is the number of haplotypes *h* carried by a subject). The most frequent haplotype is chosen as the reference haplotype, and those with haplotype frequency < 2% are grouped into the reference.

For the colorectal adenoma study, the environmental covariate vector **X** includes the variables Age (age in years), Female (indicator for female gender), Smk1 and Smk2 (indicators

respectively for former and current smokers). All the environmental variables are centered to have mean zero. The haplotype-environment interaction terms include the interactions of the haplotype "101010", a key haplotype of interest because of its known role in the metabolism of smoking related carcinogens (Chang-Claude et al. 2002), with Smk1 and Smk2. For the prostate cancer study, the environmental covariate vector **X** includes the age level (categorized into four groups), center (nine groups), and the [25(OH)D)] level (nmol/L). We consider the interaction of the haplotypes "000" and "001" with the standardized [25(OH)D)] level (standardized by its sample standard deviation), which are the most promising interactions according to preliminary analysis.

6.2 Results

Table 4 shows the results for the two applications, based on the association analyses introduced in Sections 2, 3, and 4. For simplicity, we report only results for the interaction parameters of main interest (the full models we fitted to the two datasets are described in Section 6.1). Also, we report only results from the model-free and model-based analyses, the component-wise empirical Bayes shrinkage analysis, and the L_1 penalized likelihood analysis (with the penalty parameters chosen by matching them with the estimated prior variances, as discussed in Section 4.1). The results from the multivariate empirical Bayes shrinkage analysis are quite similar to those from the component-wise empirical Bayes shrinkage analysis, and a similar relationship holds between the L_1 and the L_2 penalized likelihood analyses. The standard errors are obtained by the formulas given in the Appendices C and F, and the p-values are obtained by normal approximation.

The major conclusions we draw from Table 4 are as follows. For the colorectal adenoma study, where we expect that the haplotype-environment independence may be violated and hence the model-based analysis may be biased, we do observe a large discrepancy between the point estimates from the model-free and the model-based analyses. The empirical Bayes and the penalized likelihood methods we propose nicely adapt to the situation and produce results much closer to the model-free analysis than the model-based analysis. In summary, carriers of the haplotype "101010" tend to have lower smoking related risk than noncarriers. This agrees with previous laboratory and epidemiologic studies that have identified the haplotype "101010", known as *NAT2*4*, as a rapid metabolizer for smoking related carcinogens (e.g., see Moslehi et al. 2006).

For the prostate cancer study, where we expect haplotype-environment independence to hold, the model-free and the model-based analyses do produce very similar point estimates, revealing that both of them may be unbiased. However, the model-based analysis is more efficient than the model-free analysis. We note that in this example the component-wise empirical Bayes analysis and the L_1 penalized likelihood analysis produce results quite close to the model-based analysis, especially in the interaction of the [25(OH)D)] level with haplotype "000", and thus retain the efficiency gain. In summary, we can conclude from the model-based and the proposed shrinkage methods that the haplotype "000" may significantly modify the risk of prostate cancer associated with the [25(OH)D)] level.

7. DISCUSSION

In this work we first examined retrospective likelihood methods for haplotype-based case-control studies. A "model-free" retrospective likelihood was proposed, which, in contrast to the "model-based" retrospective likelihood (Spinka et al. 2005) that assumes HWE and haplotype-environment independence to specify the joint distribution of diplotypes and environmental exposures, utilizes the same assumptions only to specify the *conditional* diplotype distribution given environmental exposures and unphased genotypes. Our simulation studies show that such retrospective likelihood analysis leads to inference that is very robust

to violation of the HWE and (*G-E*) independence assumptions. With the rare disease assumption, the proposed "model-free" retrospective likelihood is closely related to the "prospective likelihood" used by Zhao et al. (2003) and Lake et al. (2003).

We further considered a number of alternative shrinkage estimation techniques for adaptive relaxation of the HWE and *G-E* independence assumptions from the model-based retrospective likelihood. These different shrinkage estimators were constructed by different ways of shrinking the model-free estimator toward the model-based estimator, with the aim of achieving better finite sample properties in terms of the bias-efficiency trade-off. Our simulation studies reveal that the proposed shrinkage estimators can dramatically reduce the bias of "model-based" retrospective methods in the presence of violation of model assumptions. On the other hand, when the model assumptions are satisfied, exactly or approximately, the shrinkage estimators can retain a considerable gain in efficiency over the "model-free" estimators. Thus, overall, the proposed techniques provide a natural way of trading off between bias and efficiency in the type of problems we studied in this article. Two empirical illustrations were described; one where the assumption of *G-E* independence is likely violated and one where it likely holds, and analysis results from these examples are consistent with those from the simulations.

We also have studied asymptotic theory for the EB and penalized shrinkage estimators. We have shown that the proposed EB estimator is approximately $\sqrt{N}$−consistent whether the underlying assumptions of HWE and *G-E* independence hold or not. Further, when the assumptions of HWE or *G-E* independence assumptions or both are violated, we have shown that the EB estimator has an asymptotic normal distribution, the variance of which can be reliably estimated in a simple closed form using the δ method. On the other hand, when the assumptions of HWE and *G-E* independence are met, the proposed EB estimator can be viewed as a model-average estimator that converges to a nonnormal distribution. In practice, however, we found that this limiting nonnormal distribution can be well approximated by a normal distribution, the variance of which can still be reliably estimated using the same estimator derived for the case when the model assumptions fail. All the previously mentioned properties are also satisfied by the penalized estimators for suitable choice of the penalty parameters.

In this article, we have compared the performance of the alternative shrinkage estimators in terms of mean squared errors of the estimators. In large genetic association studies, however, one is often first interested in testing as opposed to estimation. It is important to note that the misspecification of HWE or (*G-E*) independence or both causes bias even under the null hypothesis of no association or no interaction. Thus, from both the testing and estimation points of view, it is important to account for the impact of model misspecification. Although in this article we do not study the properties of shrinkage estimators from a testing point of view, a recent study has reported that EB-type shrinkage estimators indeed performs very well for large scale association testing (Mukherjee et al. 2008).

The advantages in bias, efficiency, computational simplicity, and availability of a unified approach to inference make the proposed shrinkage procedures very appealing in genetic association studies. Further, using modern shrinkage and penalization techniques to combine assumption-laden and assumption-free methods in semiparametric problems is an idea that transcends genetic association studies. We hope that this work will lead to further research in this general area.

Acknowledgements

APPENDIX A: LINKING THE CASE-CONTROL AND ALTERNATIVE SAMPLING SCHEMES

In this section, we use the subscripts "alt" and "cc" to denote the expectation operators under the alternative (see Section 2.3.3) and the case-control sampling schemes. We drop the subscript "free" for the parameters.

Let $\mathbf{\Psi}_{\text{free}}(D, \mathbf{G}, \mathbf{X}, \mathbf{\Omega})$ be the estimating function we consider for the "model-free" estimation, where the estimating function for $\boldsymbol{\beta}$ is obtained by the score $\partial \log L_{\text{free}}(D, \mathbf{G}, \mathbf{X}, \mathbf{\Omega})/\partial \boldsymbol{\beta}$, and the estimating function for $\boldsymbol{\theta}$ is obtained by the EM-type estimating function based on the controls ($D = 0$) (Zhao et al. 2003). Because L_{free} is a legitimate likelihood function in the alternative sampling scheme and the rare disease assumption is used, $\mathbf{\Psi}_{\text{free}}(D, \mathbf{G}, \mathbf{X}, \mathbf{\Omega})$ is an unbiased estimating function for $\mathbf{\Omega}$ in the alternative sampling scheme. If $\bullet = (D, \mathbf{G}, \mathbf{X})$, then this means that $0 = E_{\text{alt}}\{\mathbf{\Psi}_{\text{free}}(\bullet, \mathbf{\Omega})|R = 1\}$ at the true value of $\mathbf{\Omega}$.

The following result will be used to justify using $\mathbf{\Psi}_{\text{free}}(D, \mathbf{G}, \mathbf{X}, \mathbf{\Omega})$ as an unbiased estimating function in the case-control sampling scheme.

Theorem 1. For any function $K(D, \mathbf{G}, \mathbf{X})$,

$$E_{\text{alt}}\{K(D,\mathbf{G},\mathbf{X})|R=1\} = E_{\text{cc}}\{K(D,\mathbf{G},\mathbf{X})\}. \tag{A.1}$$

A.1 PROOF OF THEOREM 1

By definition,

$$\begin{aligned} E_{\text{cc}}\{K(D,\mathbf{G},\mathbf{X})\} &= N^{-1} \sum_{i=1}^{N} E\{K(D_i,\mathbf{G}_i,\mathbf{X}_i)|D_i\} \\ &= \sum_{d=0}^{1} p_d E\{K(d,\mathbf{G},\mathbf{X})|D=d\}. \end{aligned}$$

Further, note that $\text{pr}_{\text{alt}}(D = d|R = 1) p_d$. Recall that the distribution of R depends only on D. It then follows that

$$\begin{aligned} E_{\text{alt}}\{K(D,\mathbf{G},\mathbf{X})|R=1\} &= E_{\text{alt}}[E_{\text{alt}}\{K(D,\mathbf{G},\mathbf{X})|D,R=1\}|R=1] \\ &= E_{\text{alt}}[E_{\text{alt}}\{K(D,\mathbf{G},\mathbf{X})|D\}|R=1] \\ &= E_{\text{alt}}[E\{K(D,\mathbf{G},\mathbf{X})|D\}|R=1] \\ &= \sum_{d=0}^{1} p_d E\{K(D,\mathbf{G},\mathbf{X})|D=d\}, \end{aligned}$$

completing the proof. In the previous argument, the second line uses the fact that the distribution of R is independent of $(\mathbf{H}, \mathbf{G}, \mathbf{X})$ given D, whereas the third line notes that probability calculations about the distribution of $(\mathbf{H}, \mathbf{G}, \mathbf{X})$ given D are the same in either sampling scheme.

APPENDIX B: ASYMPTOTIC THEORY FOR (5) AND (6)

Theorem 1 in Appendix A shows that expectations in the case-control sampling scheme are the same as expectations in the alternative sampling scheme, and hence that $\Psi_{\text{free}}(\cdot)$ is an unbiased estimating function [i.e., $N^{-1}\sum_{i=1}^{N} E\{\Psi_{\text{free}}(D_i,\mathbf{G}_i,\mathbf{X}_i,\Omega_{\text{free}})|D_i\}=0$]. Let $\mathbf{I}_{\text{free}}$ be the information matrix,

$$\mathcal{I}_{\text{free}} = -N^{-1}\sum_{i=1}^{N} E\left\{\partial\Psi_{\text{free}}(D_i,\mathbf{G}_i,\mathbf{X}_i,\Omega_{\text{free}})/\partial\Omega_{\text{free}}^{\mathsf{T}}|D_i\right\},$$

which can be estimated as

$$\widehat{\mathcal{I}}_{\text{free}} = -N^{-1}\sum_{i=1}^{N} \partial\Psi_{\text{free}}\left(D_i,\mathbf{G}_i,\mathbf{X}_i,\widehat{\Omega}_{\text{free}}\right)/\partial\Omega_{\text{free}}^{\mathsf{T}}.$$

Define $\Lambda_{\text{free}} = \sum_d p_d E\{\Psi_{\text{free}}(d,\mathbf{G},\mathbf{X},\Omega_{\text{free}})|D=d\} E\left\{\Psi_{\text{free}}^{\mathsf{T}}(d,\mathbf{G},\mathbf{X},\Omega_{\text{free}})|D=d\right\}$ where the expectations are estimated as $N_d^{-1}\sum_{i=1}^{N} I(D_i=d)\Psi_{\text{free}}\left(d,\mathbf{G}_i,\mathbf{X}_i,\widehat{\Omega}_{\text{free}}\right)$. Also define $\mathcal{I}_{\text{free}}^{*} = \sum_d p_d E\left\{\Psi_{\text{free}}(d,\mathbf{G},\mathbf{X},\Omega_{\text{free}})\Psi_{\text{free}}^{\mathsf{T}}\left(d,\mathbf{G},\mathbf{X},\widehat{\Omega}_{\text{free}}\right)|D=d\right\}$, which can be estimaed by

$$\widehat{\mathcal{I}}_{\text{free}}^{*} = N^{-1}\sum_{i=1}^{N} \Psi_{\text{free}}\left(D_i,\mathbf{G}_i,\mathbf{X}_i,\widehat{\Omega}_{\text{free}}\right)\Psi_{\text{free}}^{\mathsf{T}}\left(D_i,\mathbf{G}_i\mathbf{X}_i,\widehat{\Omega}_{\text{free}}\right),$$

Then because of Theorem 1, we have the following result.

Theorem 2. As $N_0, N_1 \to \infty$,

$$N^{1/2}\left(\widehat{\Omega}_{\text{free}} - \Omega_{\text{free}}\right) = N^{-1/2}\sum_{i=1}^{N} \mathcal{I}_{\text{free}}^{-1}\Psi_{\text{free}} \\ \times(D_i,G_i,X_i,\Omega_{\text{free}}) + o_p(1). \tag{A.2}$$

In particular this means that $N^{1/2}\left(\widehat{\Omega}_{\text{free}} - \Omega_{\text{free}}\right) \Rightarrow \text{Normal}\left\{0, \mathcal{I}_{\text{free}}^{-1}\left(\mathcal{I}_{\text{free}}^{*} - \Lambda_{\text{free}}\right)\mathcal{I}_{\text{free}}^{-1}\right\}$. In addition, the estimates $\widehat{\mathcal{I}}_{\text{free}}, \widehat{\mathcal{I}}_{\text{free}}^{*}$, and $\widehat{\Lambda}_{\text{free}}$ are consistent for $\mathcal{I}_{\text{free}}, \mathcal{I}_{\text{free}}^{*}$ and Λ_{free}.

For the sake of completeness, we note that a similar expansion was derived by Spinka, et al. (2005) when HWE and (*G-E*) independence hold.

Theorem 3. As $N_0, N_1 \to \infty$,

$$N^{1/2}\left(\widehat{\Omega}_{\text{model}} - \Omega_{\text{model}}\right) = N^{-1/2}\sum_{i=1}^{N} \mathcal{I}_{\text{model}}^{-1}\Psi_{\text{model}} \\ \times(D_i,G_i,X_i,\Omega_{\text{free}}) + o_p(1). \tag{A.3}$$

In particular, this means that $N^{1/2}\left(\widehat{\Omega}_{\text{model}} - \Omega_{\text{model}}\right) \Rightarrow \text{Normal}\left\{0, \mathcal{I}_{\text{model}}^{-1}\left(\mathcal{I}_{\text{model}} - \Lambda_{\text{model}}\right)\mathcal{I}_{\text{model}}^{-1}\right\}$. In addition, the estimates

$\widehat{\mathcal{I}}_{\text{model}}$ and $\widehat{\Lambda}_{\text{model}}$ are consistent for $\mathbf{I}_{\text{model}}$ and Λ_{model} (all the matrices here are defined analogously to their counterparts defined in Theorem 2).

APPENDIX C: ASYMPTOTIC THEORY AND VARIANCE ESTIMATION FOR EB SHRINKAGE ESTIMATORS WHEN βmodel ≠ βfree

Here we focus on $\widehat{\beta}_{\text{EB1}}$; the asymptotic theory for $\widehat{\beta}_{\text{EB2}}$ can be derived analogously (the definitions for the EB1 and EB2 empirical Bayes-type estimators is given in Section 5.1). The asymptotic theory is simple, because when $\boldsymbol{\beta}_{\text{model}} \neq \boldsymbol{\beta}_{\text{free}}$, the EB shrinkage estimators are asymptotically equivalent to the model-free estimator (see later). Our main goal here is to develop for the EB shrinkage estimator an approximate covariance matrix estimator that is more accurate in finite samples than the covariance matrix for the model-free estimator. Note that, although the latter also serves as an estimator for variance of the EB estimator due to the asymptotic equivalence, it is often too conservative in finite samples (simulation data not shown).

Let $\widehat{\beta} = \left(\widehat{\beta}_{\text{model}}^{\mathsf{T}}, \widehat{\beta}_{\text{free}}^{\mathsf{T}}\right)^{\mathsf{T}}$, and $\widehat{\Sigma}_{\widehat{\beta}}$ denote the estimated variance-covariance of $\widehat{\beta}$. Let $\widehat{\Psi}_{\text{model},i} = \Psi_{\text{model}}\left(D_i, \mathbf{G}_i, \mathbf{X}_i, \widehat{\Omega}_{\text{model}}\right)$, and $\widehat{\Psi}_{\text{free},i}$ be defined analogously. The block-diagonal terms of $\widehat{\Sigma}_{\widehat{\beta}}$ are obtained directly from Theorems 2 and 3, and, by Theorems 2 and 3, the off block-diagonal terms are given by

$$N^{-2}\begin{pmatrix}\mathbf{I}_p & 0\end{pmatrix}\widehat{\mathcal{I}}_{\text{model}}^{-1}\left(\sum_{i=1}^{N}\widehat{\Psi}_{\text{model},i}\widehat{\Psi}_{\text{free},i}^{\mathsf{T}} - \widehat{\Lambda}\right)\widehat{\mathcal{I}}_{\text{free}}^{-1}\begin{pmatrix}\mathbf{I}_p & 0\end{pmatrix}^{\mathsf{T}}$$

and its transpose, where $\mathbf{I}_p$ is the identity matrix of size $p = \dim(\boldsymbol{\beta})$, $\mathbf{0}$ is a $p \times q$ zero matrix ($q = \dim(\boldsymbol{\theta})$), and the matrix $\widehat{\Lambda} = \sum_d p_d \widehat{E}\left(\widehat{\Psi}_{\text{model}}|D=d\right)\widehat{E}\left(\widehat{\Psi}_{\text{free}}^{\mathsf{T}}|D=d\right)$, with $\widehat{E}\left(\widehat{\Psi}_*|D=d\right) \equiv N_d^{-1}\sum_{i=1}^{N}I(D_i=d)\widehat{\Psi}_{*,i}$ $d=0,1$. Let $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_{\text{model}}, \boldsymbol{\beta}_{\text{free}})$, ψ and $\boldsymbol{\beta}_{\text{EB1}}$ be the asymptotic limits of $\widehat{\beta}, \widehat{\psi}$ and $\widehat{\beta}_{\text{EB1}}$, respectively. Note that when $\boldsymbol{\beta}_{\text{model}} \neq \boldsymbol{\beta}_{\text{free}}$, it is easy to see that $\widehat{\beta}_{\text{EB1}} \to \beta_{\text{free}}$ in probability when N goes to ∞ because in this case $V \to 0$ and $\widehat{\psi} \to \psi \neq 0$. Thus, $\boldsymbol{\beta}_{\text{EB1}} = \boldsymbol{\beta}_{\text{free}}$.

To get an approximate estimated covariance matrix, which is more accurate than that of the model-free estimator in small samples, we use the following calculations. Define

$$\widetilde{\beta}_{\text{EB1}} = \beta_{\text{free}} + V(V + \psi\psi^{\mathsf{T}})^{-1}\psi$$

and note that by a first order Taylor expansion, we can write

$$\sqrt{N}\left(\widehat{\beta}_{\text{EB1}} - \widetilde{\beta}_{\text{EB1}}\right) = \mathcal{G}_1 \times \sqrt{N}\left(\widehat{\beta} - \beta_*\right) + o_p(1),$$

where the $p \times 2p$ matrix $\mathbf{G}_1$ is given as

$$\mathcal{G}_1 = \left(\begin{array}{c} -\dfrac{2\psi\psi^{\mathsf{T}}\mathbf{V}^{-1}}{(1+\psi^{\mathsf{T}}\mathbf{V}^{-1}\psi)^2} + \dfrac{1}{(1+\psi^{\mathsf{T}}\mathbf{V}^{-1}\psi)}\mathbf{I}_p \\[3mm] \dfrac{\psi^{\mathsf{T}}\mathbf{V}^{-1}\psi}{(1+\psi^{\mathsf{T}}\mathbf{V}^{-1}\psi)}\mathbf{I}_p + \dfrac{2\psi\psi^{\mathsf{T}}\mathbf{V}^{-1}}{(1+\psi^{\mathsf{T}}\mathbf{V}^{-1}\psi)^2} \end{array} \right).$$

Of course, because $\mathbf{V} = O_p(N^{-1})$, it follows that $\sqrt{N}\left(\widetilde{\beta}_{\mathrm{EB1}} - \beta_{\mathrm{EB1}}\right) = o_p(1)$. Thus, $\widehat{\beta}_{\mathrm{EB1}}$ is approximately $\sqrt{N}$−consistent and asymptotically normal. Moreover, the asymptotic variance of $\widehat{\beta}_{\mathrm{EB1}}$ can be estimated as $\widehat{\mathcal{G}}_1 \sum_{\widehat{\beta}} \widehat{\mathcal{G}}_1^{\mathsf{T}}$, where $\widehat{\mathcal{G}}_1$ is the plug-in estimate of $\mathbf{G}_1$. The analogous asymptotic variance-covariance matrix estimator for $\widehat{\beta}_{\mathrm{EB2}}$ is given by $\widehat{\mathcal{G}}_2 \sum_{\widehat{\beta}} \widehat{\mathcal{G}}_2^{\mathsf{T}}$, where $\widehat{\mathcal{G}}_2$ is the plug-in estimate for

$$\mathcal{G}_2 = \left(\mathrm{diag}\left\{ \frac{v_j\left(v_j - \psi_j^2\right)}{\left(v_j + \psi_j^2\right)^2} \right\} \mathbf{I}_p - \mathrm{diag}\left\{ \frac{v_j\left(v_j - \psi_j^2\right)}{\left(v_j + \psi_j^2\right)^2} \right\} \right).$$

APPENDIX D: COMPUTATION FOR PENALIZED LIKELIHOOD ESTIMATION

The penalized likelihood estimators $\widehat{\beta}_{\mathrm{PL1}}$ and $\widehat{\beta}_{\mathrm{PL2}}$ are obtained from (10) with L_1 (LASSO) and L_2 (ridge) penalties, respectively. Their implementation involves solving the corresponding score functions for (10); note that the score function used for θ in $L_{\mathrm{free}}(\cdot)$ is modified as described in Section 2.3.1.

Following Fan and Li (2001, Section 3.3), we use a unified Newton-Raphson algorithm for implementing these estimators, where the penalty function is approximated locally by a quadratic function, which is exact for the L_2 penalty. Specifically, write the penalty function $P(\beta_j, \beta_{\mathrm{model},j})$ as $P(|b_j|)$ with $b_j = \beta_j - \beta_{\mathrm{model},j}$, and take $b_j^* = \widehat{\beta}_j^* - \beta_{\mathrm{model},j}$, where $\widehat{\beta}_j^*$ is a current estimate for β_j. Approximate $P(|b_j|)$ by a quadratic function around b_j^* such that

$$\mathrm{d}P\left(|b_j^*|\right) / \mathrm{d}b_j \approx \left\{ \dot{P}\left(|b_j^*|\right) / |b_j^*| \right\} b_j^*,$$

where the superscript dot denotes derivative. Let

$$\Psi_{\mathrm{free}}(\Omega) = \sum_{i=1}^{N} \Psi_{\mathrm{free}}(D_i, \mathbf{G}_i, \mathbf{X}_i, \Omega),$$

where $\boldsymbol{\Psi}_{\mathrm{free}}(\cdot)$ is defined in Section 3.1, and

$$\overline{\mathcal{I}}_{\mathrm{free}}(\Omega) = -\partial\Psi_{\mathrm{free}}^{\mathsf{T}}(\Omega) / \partial\Omega.$$

Let $\widehat{\Omega}_{\mathrm{model}}$ be the "model-based" estimates for $\boldsymbol{\Omega}$ obtained from (6). Then at a current parameter estimate $\boldsymbol{\Omega}^* = (\beta^*, \theta^*)$, the score function for the penalized likelihood can be approximated by

$$\Psi_{\text{free}}(\Omega^*) - \Gamma(\Omega^*)\left(\Omega^* - \widehat{\Omega}_{\text{model}}\right),$$

where the matrix $\boldsymbol{\Gamma}(\boldsymbol{\Omega}^*)$ is diagonal whose diagonal elements are $\lambda_j \dot{P}\left(|b_j^*|\right)/|b_j^*|$ for those corresponding to $\boldsymbol{\beta}_j$, and are 0 for those corresponding to $\boldsymbol{\theta}$, becaue we do not penalize the estimates for haplotype frequency parameters $\boldsymbol{\theta}$. The updated parameter estimate is then

$$
\begin{aligned}
\widehat{\Omega}= \ & \widehat{\Omega}^* + \left\{\bar{\mathcal{I}}_{\text{free}}(\Omega^*) + \Gamma(\Omega^*)\right\}^{-1} \\
& \times \left\{\Psi_{\text{free}}(\Omega^*) - \Gamma(\Omega^*)\left(\Omega^* - \widehat{\Omega}_{\text{model}}\right)\right\}.
\end{aligned}
$$

In the case where the L_1 penalty is used, when $\widehat{\beta}_{*,j}$ becomes close to $\widehat{\beta}_{\text{model},j}$ (e.g., the absolute difference $< 10^{-5}$), we set $\widehat{\beta}_j = \widehat{\beta}_{\text{model},j}$ and set the corresponding diagonal element of $\boldsymbol{\Gamma}(\boldsymbol{\Omega})$ to a large value (e.g., 10^5). We have found that this algorithm is very stable and fast. A sandwich-type variance estimator is proposed in Appendix F for the resulting penalized estimator.

APPENDIX E: ASYMPTOTIC THEORY FOR PENALIZED ESTIMATORS

We first consider the case that $\boldsymbol{\beta}_{\text{model}} \neq \boldsymbol{\beta}_{\text{free}}$. In this case, if $\lambda_j \to 0$ as $N \to \infty$ for $\beta_{\text{model},j} \neq \beta_{\text{free},j}$, it is then obvious that the penalized estimator for such choices of penalty parameters is asymptotically equivalent to the model-free estimator. Note that the choices of λ_js presented in Section 4.1 satisfies this condition because $v_j = O_p(N^{-1})$, whereas

$$\widetilde{\psi}_j^2 \to \psi_j^2 = \left(\beta_{\text{model},j} - \beta_{\text{free},j}\right)^2 \neq 0.$$

Next, as in Section 3.1, we consider the case where HWE and G-E independence hold, so that $\boldsymbol{\beta}_{\text{model}} = \boldsymbol{\beta}_{\text{free}}$ and $\boldsymbol{\theta}_{\text{model}} = \boldsymbol{\theta}_{\text{free}}$, and hence $\boldsymbol{\Omega}_{\text{model}} = \boldsymbol{\Omega}_{\text{free}}$. In this case, we will show that the penalized estimator, though it has the same limit value as the model-free estimator, is a different estimator from the model-free estimator. In fact, from the asymptotic distribution for the penalized estimator derived later, it is seen that the penalized estimator is a model-average estimator, and hence is more efficient than the model-free estimator.

Consider the L_2-penalized estimator first. With a slightly expanded notation we have

$$
\begin{aligned}
& N^{1/2}\left(\widehat{\beta}_{\text{free}} - \beta_{\text{free}}, \widehat{\theta}_{\text{free}} - \theta_{\text{free}}, \widehat{\beta}_{\text{model}} - \beta_{\text{model}}, \widehat{\theta}_{\text{model}} - \Omega_{\text{model}}\right) \\
& = \begin{bmatrix} \mathcal{I}_{\text{free}}^{-1}(\Omega_{\text{free}}) N^{-1/2} \sum_{i=1}^{N} \Psi_{\text{free},i}(\Omega_{\text{free}}) \\ \mathcal{I}_{\text{model}}^{-1}(\Omega_{\text{model}}) N^{-1/2} \sum_{i=1}^{N} \Psi_{\text{model},i}(\Omega_{\text{model}}) \end{bmatrix} + o_p(1) \\
& \Rightarrow \left(\mathcal{Z}_{\text{free}}^{\mathsf{T}}, \mathcal{Q}_{\text{free}}^{\mathsf{T}}, \mathcal{Z}_{\text{model}}^{\mathsf{T}}, \mathcal{Q}_{\text{model}}^{\mathsf{T}}\right)^{\mathsf{T}} \sim \text{Normal}(0, \Sigma_{\text{all}}),
\end{aligned}
$$

say, where $\boldsymbol{\Psi}_{\text{free},i}(\boldsymbol{\Omega}_{\text{free}}) = \boldsymbol{\Psi}_{\text{free}}(D_i, \mathbf{G}_i, \mathbf{X}_i, \boldsymbol{\Omega}_{\text{free}})$ and similarly for $\boldsymbol{\Psi}_{\text{model},i}(\boldsymbol{\Omega}_{\text{model}})$.

Define $\mathcal{U}\left\{N^{1/2}\left(\widehat{B}_{\text{free}} - \widehat{B}_{\text{model}}\right)\right\} = \text{diag}\left[\left(v_j/2\right)\left\{N^{1/2}\left(\widehat{\beta}_{\text{free},j} - \widehat{\beta}_{\text{model},j}\right)\right\}^{-2}, 0, \dots, 0\right]$. Then the L_2-penalized estimator as described in Section 4.1, $\widehat{\Omega}_{\text{PL2}}$, solves

$$
\begin{aligned}
0= \;& N^{-1/2} \sum_{i=1}^{N} \Psi_{\text{free},i}\left(\widehat{\Omega}_{\text{PL2}}\right) - \mathcal{U}\left\{N^{1/2}\left(\widehat{\beta}_{\text{free}} - \widehat{\Omega}_{\text{model}}\right)\right\} N^{1/2} \\
& \times \left(\widehat{\Omega}_{\text{PL2}} - \widehat{\Omega}_{\text{model}}\right).
\end{aligned}
$$

By Taylor series, we see that

$$
\begin{aligned}
0= \;& N^{-1/2} \sum_{i=1}^{N} \Psi_{\text{free},i}\left(\Omega_{\text{free}}\right) \\
& - \mathcal{U}\left\{N^{1/2}\left(\widehat{\beta}_{\text{free}} - \widehat{\beta}_{\text{model}}\right)\right\} N^{1/2}\left(\Omega_{\text{free}} - \widehat{\Omega}_{\text{model}}\right) \\
& - \left[\mathcal{I}_{\text{free}} + \mathcal{U}\left\{N^{1/2}\left(\widehat{\beta}_{\text{free}} - \widehat{\beta}_{\text{model}}\right)\right\}\right] N^{1/2}\left(\widehat{\Omega}_{\text{PL2}} - \Omega_{\text{free}}\right) + o_p(1).
\end{aligned}
$$

Dropping the dependence of $\mathbf{I}_{\text{free}}$ on $\boldsymbol{\Omega}_{\text{free}}$, remembering that the leading term in the previous expression has the limit distribution $\mathcal{I}_{\text{free}}\left(\mathcal{Z}_{\text{model}}^{\mathsf{T}}, Q_{\text{free}}^{\mathsf{T}}\right)^{\mathsf{T}}$ and then solving, we see that

$$
\begin{aligned}
N^{1/2}\left(\widehat{\Omega}_{\text{PL2}} - \Omega_{\text{free}}\right) \;\Rightarrow\;& \left\{\mathcal{I}_{\text{free}} + \mathcal{U}\left(\mathcal{Z}_{\text{model}} - \mathcal{Z}_{\text{free}}\right)\right\}^{-1} \\
& \times \left[\mathcal{I}_{\text{free}}\left(\mathcal{Z}_{\text{free}}^{\mathsf{T}}, Q_{\text{free}}^{\mathsf{T}}\right)^{\mathsf{T}}\right. \\
& \left. + \mathcal{U}\left(\mathcal{Z}_{\text{model}} - \mathcal{Z}_{\text{free}}\right)\left(\mathcal{Z}_{\text{model}}^{\mathsf{T}}, Q_{\text{model}}^{\mathsf{T}}\right)^{\mathsf{T}}\right] \\
& + o_p(1).
\end{aligned}
$$

Note that this limit distribution, just as that for the EB estimators given in (8), is a type of distribution for a model-average estimator. It is then seen that, when HWE and *G-E* independence hold, the limit distribution for $\widehat{\beta}_{\text{PL2}}$ is in principle not a normal distribution. Our simulations, however, show that the departure from normality in this case is b not large.

Now consider the asymptotic distribution for the L_1-penalized estimator, $\widehat{\Omega}_{\text{PL1}}$, when HWE and *G-E* independence hold, with the penalty parameter λ being chosen as in Section 4.1. Let

$\mathcal{R}\left(N^{1/2}\widehat{\beta}_{\text{free}}, N^{1/2}\widehat{\beta}_{\text{model}}\right) = \text{diag}\left[\left(v_j/2\right)^{1/2}|N^{1/2}\left(\widehat{\beta}_{\text{free},j} - \widehat{\beta}_{\text{model},j}\right)\mathcal{B}_j^*|^{-1}, 0, \ldots, 0\right]$, where

$\mathcal{B}_j^* = N^{1/2}\left(\widehat{\beta}_{\text{PL1},j} - \widehat{\beta}_{\text{model},j}\right)$. By Taylor expansion we have

$$
\begin{aligned}
0= \;& N^{-1/2} \sum_{i=1}^{N} \Psi_{\text{free},i}\left(\Omega_{\text{free}}\right) - \mathcal{R}\left(\mathcal{Z}_{\text{model}}, \mathcal{Z}_{\text{free}}\right) N^{1/2} \\
& \left(\Omega_{\text{free}} - \widehat{\Omega}_{\text{model}}\right) \\
& - \left\{\mathcal{I}_{\text{free}} + \mathcal{R}\left(\mathcal{Z}_{\text{model}}, \mathbf{Z}_{\text{free}}\right)\right\} N^{1/2}\left(\widehat{\Omega}_{\text{PL1}} - \Omega_{\text{free}}\right) + o_p(1),
\end{aligned}
$$

which implies that

$$
\begin{aligned}
N^{1/2}\left\{\widehat{\Omega}_{\text{PL1}} - \Omega_{\text{free}}\right\} \;\Rightarrow\;& \left(\mathcal{I}_{\text{free}} + \mathcal{R}\left(\mathcal{Z}_{\text{model}}, \mathcal{Z}_{\text{free}}\right)\right)^{-1} \\
& \left[\mathcal{I}_{\text{free}}\left(\mathcal{Z}_{\text{free}}^{\mathsf{T}}, Q_{\text{free}}^{\mathsf{T}}\right)^{\mathsf{T}}\right. \\
& \left. + \mathcal{R}\left(\mathcal{Z}_{\text{model}}, \mathcal{Z}_{\text{free}}\right)\left(\mathcal{Z}_{\text{model}}^{\mathsf{T}}, Q_{\text{model}}^{\mathsf{T}}\right)^{\mathsf{T}}\right] \\
& + o_p(1),
\end{aligned}
$$

which is again a form of the distribution for a model-average estimator, and hence is in general not normally distributed.

APPENDIX F: VARIANCE ESTIMATION FOR PENALIZED ESTIMATORS

The major purpose of this section is to derive accurate and easily computed approximate variance estimates for the penalized estimators, using sandwich ideas. Here we assume that $\boldsymbol{\beta}_{\text{model}} \neq \boldsymbol{\beta}_{\text{free}}$, and the penalty parameters λ_js are chosen so that $\lambda_j \to 0$ as $N \to \infty$. As discussed in Appendix E, the penalized estimator in this case is asymptotically normal with mean $\boldsymbol{\Omega}_{\text{PL}} = \boldsymbol{\Omega}_{\text{free}}$.

Let

$$\bar{\Psi}_{\text{free}}(\Omega) = \sum_{i=1}^{N} \Psi_{\text{free}}(D_i, \mathbf{G}_i, \mathbf{X}_i, \Omega) \equiv \sum_{i=1}^{N} \Psi_{\text{free},i}(\Omega),$$

and $\bar{\mathcal{I}}_{\text{free}}(\Omega) = -\partial \bar{\Psi}_{\text{free}}(\Omega)/\partial \Omega^{\mathsf{T}}$, where $\boldsymbol{\Psi}_{\text{model},i}(\boldsymbol{\Omega}_{\text{model}})$ and $\boldsymbol{\Psi}_{\text{free},i}(\boldsymbol{\Omega}_{\text{free}})$ are defined as in the Appendix E. In Appendix D we approximated the score function by

$$\bar{\Psi}_{\text{PL}} \equiv \Psi_{\text{free}}(\Omega) - \Gamma(\Omega)\left(\Omega - \widehat{\Omega}_{\text{model}}\right),$$

(A.4)

where the matrix $\boldsymbol{\Gamma}(\boldsymbol{\Omega})$ is diagonal with diagonal elements $\lambda_j \dot{P}\left(|b_j|\right)/|b_j|, b_j = \beta_j - \widehat{\beta}_{\text{model,j}}$, for those corresponding to β_j, and 0 for those corresponding to $\boldsymbol{\theta}$.

Recall that in our proposal the penalty parameters λ may be functions of $\widehat{\psi} = \widehat{\beta}_{\text{model}} - \widehat{\beta}_{\text{free}}$. Let

$$\begin{aligned} \mathcal{Z}_{\text{model},i}(\Omega_{\text{model}}) &= \mathcal{I}_{\text{model}}^{-1} \Psi_{\text{model},i}(\Omega_{\text{model}}), \\ \mathcal{Z}_{\text{free},i}(\Omega_{\text{model}}) &= \mathcal{I}_{\text{free}}^{-1} \Psi_{\text{free},i}(\Omega_{\text{free}}). \end{aligned}$$

Accordingly, we can further write $\boldsymbol{\Psi}_{\text{PL}} = \Sigma_i \boldsymbol{\Psi}_{\text{PL},i}(\boldsymbol{\Omega}) + O_p(N^{1/2})$, where

$$\begin{aligned} \Psi_{\text{PL},i}(\Omega) = \ & \Psi_{\text{free},i}(\Omega) + N^{-1}\Gamma(\Omega)\{\mathcal{Z}_{\text{model},i} - (\Omega - \Omega_{\text{model}})\} \\ & - N^{-1}\mathcal{P}(\Omega, \Omega_{\text{model}}, \Omega_{\text{free}})\{\mathcal{Z}_{\text{model},i}(\Omega_{\text{model}}) \\ & - \mathcal{Z}_{\text{free},i}(\Omega_{\text{free}})\}, \end{aligned}$$

where the matrix $\mathbf{P}(\boldsymbol{\Omega}, \boldsymbol{\Omega}_{\text{model}}, \boldsymbol{\Omega}_{\text{free}})$ is diagonal with diagonal elements $\left(d\lambda_j/d\psi_j\right)\dot{P}\left(|b_j|\right)b_j/|b_j|, b_j = \beta_j - \beta_{\text{model, }j}$, for elements corresponding to β_j, and 0 for elements corresponding to $\boldsymbol{\theta}$. It is thus seen that we have $N^{-1/2}\boldsymbol{\Psi}_{\text{PL}} = N^{-1/2}\boldsymbol{\Psi}_{\text{free}}(\boldsymbol{\Omega}) + O_p(1)$ (i.e., the penalized estimator is asymptotically equivalent to the mode-free estimator when $\boldsymbol{\beta}_{\text{model}} \neq \boldsymbol{\beta}_{\text{free}}$, if the penalty parameters are chosen such that $\lambda\left(\widehat{\psi}_j\right) = o_p\left(N^{1/2}\right)$ and $d\lambda\left(\widehat{\psi}_j\right)/d\widehat{\psi}_j = o_p(N)$, and provided that the terms $\dot{P}\left(|b_j|\right)b_j/|b_j|$ are bounded). The choices of λ_j in Section 4.1 satisfy theses conditions.

Denote by $\widehat{\Omega}_{\text{PL}}, \widehat{\Omega}_{\text{model}}$, and $\widehat{\Omega}_{\text{free}}$ the penalized, model-based and model-free estimators, respectively, and $\widehat{\Lambda}_{\text{PL}} = \sum_d p_d \widehat{E}\left\{\Psi_{\text{PL}}|D=d\right\}\widehat{E}\left\{\Psi_{\text{PL}}^{\mathsf{T}}|D=d\right\}$, where $\widehat{E}\left(\Psi_{\text{PL}}|D=d\right) \equiv N_d^{-1}\sum_{i=1}^{N} I(D_i=d)\Psi_{\text{PL},i}\left(\widehat{\Omega}_{\text{PL}}\right)$. Then a sandwich-type variance estimator for the penalized likelihood estimator can be obtained as

$$\left\{ \bar{\mathcal{I}}_{\text{free}}\left(\widehat{\Omega}_{\text{PL}}\right) + \Gamma\left(\widehat{\Omega}_{\text{PL}}\right) \right\}^{-1} \left(\sum_{i=1}^{N} \Psi_{\text{PL},i}\left(\widehat{\Omega}_{\text{PL}}\right)^{\otimes 2} - \widehat{\Lambda}_{\text{PL}} \right)$$
$$\times \left\{ \bar{\mathcal{I}}_{\text{free}}\left(\widehat{\Omega}_{\text{PL}}\right) + \Gamma\left(\widehat{\Omega}_{\text{PL}}\right) \right\}^{-1},$$

where $\mathbf{A}^{\otimes 2}$ for a vector $\mathbf{A}$.

REFERENCES

Chang-Claude J, Kropp S, Jäger B, Bartsch H, Risch A. Differential Effect of NAT2 on the Association between Active and Passive Smoke Exposure and Breast Cancer Risk. Cancer Epidemiology, Biomarkers & Prevention 2002;11:698–704.

Chatterjee N, Carroll RJ. Semiparametric Maximum Likelihood Estimation in Case-Control Studies of Gene-Environment Interactions. Biometrika 2005;92:399–418.

Chatterjee N, Chen Y-H. Maximum Likelihood Inference on a Mixed Conditionally and Marginally Specified Regression Model for Genetic Epidemiologic Studies with Two-Phase Sampling. Journal of the Royal Statistical Society 2007;69:123–142.Series B

Claeskens G, Carroll RJ. Post-Model Selection Inference in Semiparametric Models. Biometrika 2007;94:249–265.

Clark AG. The Role of Haplotypes in Candidate Gene Studies. Genetic Epidemiology 2004;27:321–333. [PubMed: 15368617]

Epstein MP, Satten GA. Inference on Haplotype Effects in Case-Control Studies Using Unphased Genotype Data. American Journal of Human Genetics 2003;73:1316–1329. [PubMed: 14631556]

Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. Journal of the American Statistical Association 2001;96:1348–1360.

Gohagan JK, Prorok PC, Hayes RB, Kramer BS. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, Organization, and Status. Controlled Clinical Trials 2000;21(Suppl):251S–272S. [PubMed: 11189683]

Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag; New York: 2001.

Hjort NL, Claeskens G. Frequentist Model Average Estimators. Journal of the American Statistical Association 2003;98:879–899.(with discussion)

Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ. Estimation and Tests of Haplotype-Environment Interaction When Linkage Phase Is Ambiguous. Human Heredity 2003;55:56–65. [PubMed: 12890927]

Lehmann, EL. Theory of Point Estimation. John Wiley and Sons; New York: 1983.

Li SS, Khalid N, Carlson C, Zhao LP. Estimating Haplotype Frequencies and Standard Errors for Multiple Single Nucleotide Polymorphisms. Biostatistics (Oxford, England) 2003;4:513–522.

Lin DY, Zeng D. Likelihood-Based Inference on Haplotype Effects in Genetic Association Studies. Journal of the American Statistical Association 2006;101:89–118.(with discussion)

Moslehi R, Chatterjee N, Church TR, Chen J, Yeager M, Weissfield J, Hein DW, Hayes RB. Cigarette Smoking, N-acetyltransferase Genes and the Risk of Advanced Colorectal Adenoma. Pharmacogenomics 2006;7:819–829. [PubMed: 16981843]

Mukherjee B, Ahn J, Gruber SB, Rennert G, Moreno V, Chatterjee N. Tests for Gene-Environment Interaction from Case-Control Data: A Novel Study of Type I Error, Power and Designs. Genetic Epidemiology. 2008Epub PMID: 18473390

Mukherjee B, Chatterjee N. Exploiting Gene-Environment Independence for Analysis of Case-Control Studies: A Shrinkage Approach to Trade Off Between Bias and Efficiency. Biometrics 2008;64:685–694. [PubMed: 18162111]

Prentice RL, Pyke R. Logistic Disease Incidence Models and Case-Control Studies. Biometrika 1979;66:403–411.

Satten GA, Epstein MP. Comparison of Prospective and Retrospective Methods for Haplotype Inference in Case-Control Studies. Genetic Epidemiology 2004;27:192–201. [PubMed: 15372619]

Schaid DJ. Evaluating Associations of Haplotypes with Traits. Genetic Epidemiology 2004;27:348–364. [PubMed: 15543638]

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score Tests for Association Between Traits and Haplotypes When Linkage Phase Is Ambiguous. American Journal of Human Genetics 2002;70:425–434. [PubMed: 11791212]

Spinka C, Carroll RJ, Chatterjee N. Analysis of Case-Control Studies of Genetic and Environmental Factors With Missing Genetic Information and Haplotype-Phase Ambiguity. Genetic Epidemiology 2005;29:108–127. [PubMed: 16080203]

Stram DO, Pearce CL, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC. Modeling and E-M Estimation of Haplotype-Specific Relative Risks from Genotype Data for a Case-Control Study of Unrelated Individuals. Human Heredity 2003;55:179–190. [PubMed: 14566096]

Tibshirani RJ. Regression Shrinkage and Selection via the LASSO. Journal of the Royal Statistical Society 1996;58:267–288.Series B

Wallenstein S, Hodge SE, Weston A. Logistic Regression Model for Analyzing Extended Haplotype Data. Genetic Epidemiology 1998;15:173–181. [PubMed: 9554554]

Zhao LP, Li SS, Khalid NA. Method for the Assessment of Disease Associations with Single Nucleotide Polymorphism Haplotypes and Environmental Variables in Case-Control Studies. American Journal of Human Genetics 2003;72:1231–1250. [PubMed: 12704570]
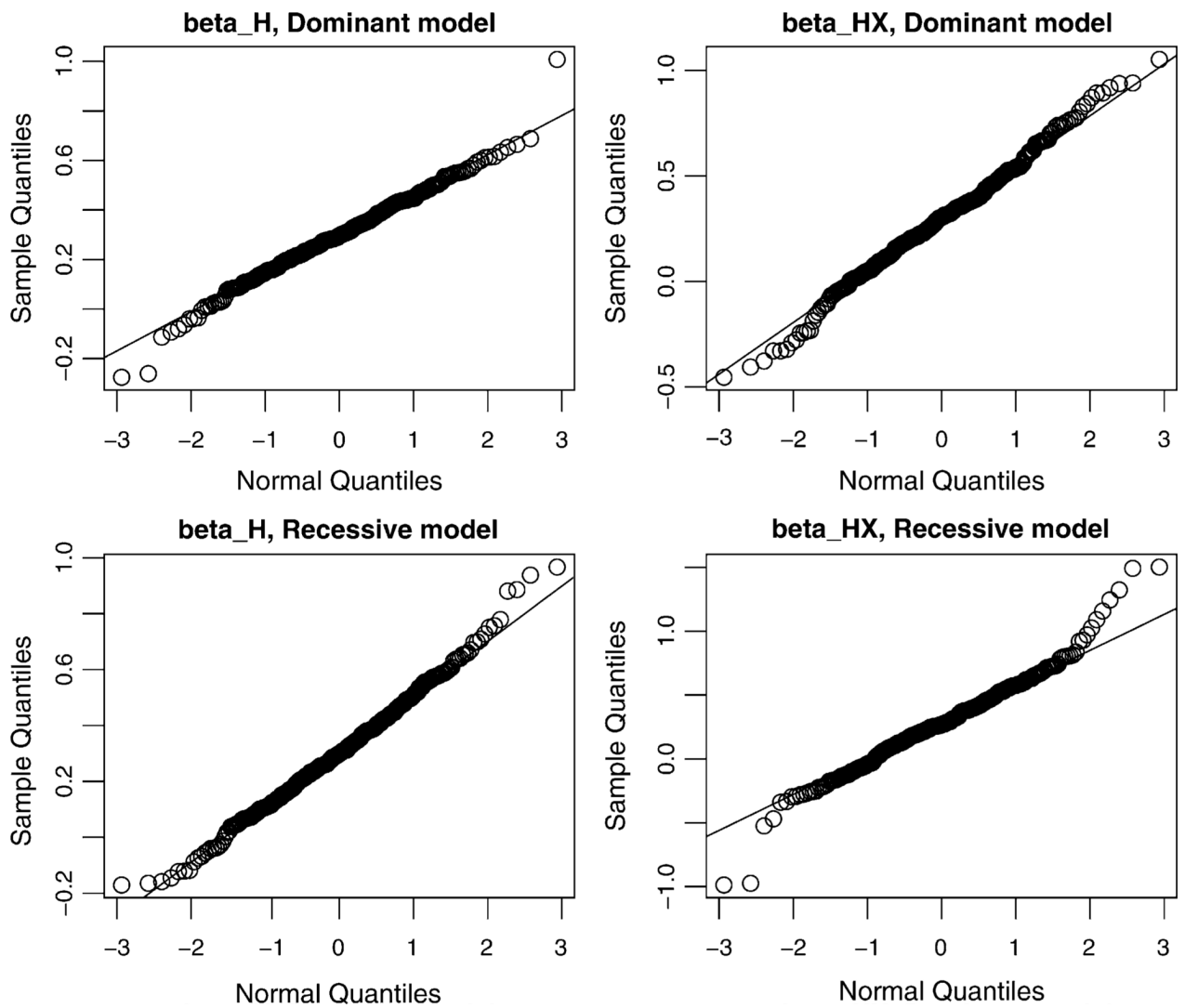
Figure 1.
Quantile-Quantile (Q-Q) plots for comparing the distribution of empirical Bayes estimates with

the normal distribution. The empirical Bayes estimates $\left(\widehat{\beta}_{EB1}\right)$ for β_H and β_{HX} are obtained from the simulation study under HWE and G-E independence, with sample size $N_1 = N_0 = 600$ (1,000) in the dominant (recessive) genetic model.

Table 1

Haplotype frequencies used in the simulation

j	Haplotype	Frequency
1	10011	0.3327
2	01011	0.1409
3	01100	0.2489
4	10000	0.0295
5	10100	0.0611
6	10110	0.0453
7	11011	0.1416

NIH-PA Author Manuscript

Table 2

MSE (bias in parenthesis) over 1,000 simulations: dominant model

		$N_1 = N_0 = 150$		$N_1 = N_0 = 300$		$N_1 = N_0 = 600$	
		MSE (bias)		MSE (bias)		MSE (bias)	
		H	$H \times X$	H	$H \times X$	H	$H \times X$
$f = 0$	$\hat{\beta}_{\text{free}}$	0.10 (0.00)	0.34 (0.01)	0.05 (0.00)	0.16 (0.05)	0.03 (0.00)	0.07 (0.00)
$\gamma_{1,3} = 0$	$\hat{\beta}_{\text{model}}$	0.07 (0.00)	0.13 (−0.04)	0.04 (0.00)	0.07 (−0.02)	0.02 (0.00)	0.04 (−0.04)
	$\hat{\beta}_{\text{HWD}}$	0.08 (0.00)	0.14 (−0.04)	0.04 (−0.01)	0.07 (−0.01)	0.02 (0.00)	0.04 (−0.04)
	$\hat{\beta}_{\text{EB1}}$	0.10 (0.00)	0.32 (0.00)	0.05 (0.00)	0.15 (0.04)	0.03 (0.00)	0.07 (0.00)
	$\hat{\beta}_{\text{EB2}}$	0.09 (−0.01)	0.23 (0.00)	0.04 (0.00)	0.11 (0.03)	0.02 (0.00)	0.05 (−0.01)
	$\hat{\beta}_{\text{PL1}}$	0.08 (0.00)	0.19 (−0.01)	0.04 (0.00)	0.10 (0.03)	0.02 (0.00)	0.05 (−0.02)
	$\hat{\beta}_{\text{PL2}}$	0.09 (0.00)	0.24 (0.00)	0.05 (0.00)	0.14 (0.04)	0.03 (0.00)	0.06 (0.00)
$f = 0.05$	$\hat{\beta}_{\text{free}}$	0.11 (−0.01)	0.34 (0.06)	0.05 (0.00)	0.17 (0.02)	0.02 (−0.02)	0.06 (0.03)
$\gamma_{1,3} = 0$	$\hat{\beta}_{\text{model}}$	0.09 (−0.12)	0.14 (−0.02)	0.05 (−0.12)	0.07 (−0.03)	0.03 (−0.13)	0.03 (−0.03)
	$\hat{\beta}_{\text{HWD}}$	0.08 (0.00)	0.14 (−0.02)	0.04 (0.01)	0.07 (−0.03)	0.02 (0.01)	0.03 (−0.03)
	$\hat{\beta}_{\text{EB1}}$	0.11 (−0.02)	0.31 (0.05)	0.05 (−0.01)	0.16 (0.02)	0.02 (−0.03)	0.06 (0.03)
	$\hat{\beta}_{\text{EB2}}$	0.10 (−0.05)	0.23 (0.03)	0.05 (−0.03)	0.12 (0.00)	0.02 (−0.05)	0.05 (0.01)
	$\hat{\beta}_{\text{PL1}}$	0.09 (−0.07)	0.20 (0.03)	0.05 (−0.05)	0.11 (0.01)	0.02 (−0.04)	0.05 (0.02)
	$\hat{\beta}_{\text{PL2}}$	0.10 (−0.03)	0.25 (0.04)	0.05 (−0.01)	0.14 (0.01)	0.02 (−0.03)	0.06 (0.02)
$f = 0$	$\hat{\beta}_{\text{free}}$	0.11 (0.01)	0.37 (0.03)	0.06 (0.00)	0.17 (0.03)	0.02 (0.00)	0.07 (0.02)
$\gamma_{1,3} = -0.4$	$\hat{\beta}_{\text{model}}$	0.11 (0.17)	0.35 (−0.46)	0.07 (0.16)	0.27 (−0.45)	0.05 (0.17)	0.26 (−0.48)
	$\hat{\beta}_{\text{HWD}}$	0.12 (0.17)	0.36 (−0.46)	0.07 (0.16)	0.28 (−0.46)	0.05 (0.16)	0.26 (−0.48)
	$\hat{\beta}_{\text{EB1}}$	0.11 (0.01)	0.35 (0.00)	0.05 (0.01)	0.16 (0.06)	0.02 (0.00)	0.07 (0.00)
	$\hat{\beta}_{\text{EB2}}$	0.10 (0.06)	0.32 (−0.11)	0.05 (0.04)	0.17 (−0.09)	0.02 (0.03)	0.08 (−0.05)
	$\hat{\beta}_{\text{PL1}}$	0.10 (0.09)	0.29 (−0.16)	0.05 (0.06)	0.16 (−0.09)	0.02 (0.03)	0.07 (−0.04)

		$N_1 = N_0 = 150$ MSE (bias)		$N_1 = N_0 = 300$ MSE (bias)		$N_1 = N_0 = 600$ MSE (bias)	
		H	**$H \times X$**	**H**	**$H \times X$**	**H**	**$H \times X$**
	$\hat{\beta}_{PL2}$	0.10 (0.04)	0.31 (−0.06)	0.05 (0.02)	0.16 (−0.02)	0.02 (0.00)	0.07 (0.01)
$f = 0.05$	$\hat{\beta}_{free}$	0.11 (−0.03)	0.33 (0.10)	0.05 (−0.00)	0.16 (−0.01)	0.03 (0.00)	0.07 (0.00)
$\gamma_{1,3} = -0.4$	$\hat{\beta}_{model}$	0.09 (0.04)	0.36 (−0.46)	0.04 (0.05)	0.28 (−0.46)	0.02 (0.05)	0.26 (−0.48)
	$\hat{\beta}_{HWD}$	0.12 (0.17)	0.38 (−0.48)	0.07 (0.16)	0.28 (−0.46)	0.05 (0.17)	0.26 (−0.48)
	$\hat{\beta}_{EB1}$	0.11 (−0.03)	0.31 (0.07)	0.05 (0.00)	0.16 (−0.02)	0.03 (0.00)	0.07 (−0.02)
	$\hat{\beta}_{EB2}$	0.10 (−0.01)	0.28 (−0.06)	0.04 (0.01)	0.16 (−0.11)	0.02 (0.02)	0.08 (−0.08)
	$\hat{\beta}_{PL1}$	0.09 (0.00)	0.25 (−0.10)	0.05 (0.01)	0.15 (−0.11)	0.02 (0.00)	0.07 (−0.05)
	$\hat{\beta}_{PL2}$	0.10 (−0.02)	0.27 (0.00)	0.05 (0.00)	0.15 (−0.04)	0.03 (0.00)	0.07 (−0.02)

NOTE: $\hat{\beta}_{model}$: model-based estimator; $\hat{\beta}_{free}$: model-free estimator; $\hat{\beta}_{HWD}$: estimator accounting for Hardy-Weinberg disequilibrium; $\hat{\beta}_{EB1}$ and $\hat{\beta}_{EB2}$: multivariate and component-wise empirical-Bayes estimators; $\hat{\beta}_{PL1}$ and $\hat{\beta}_{PL2}$: L_1 and L_2 penalized estimators. Here f is the fixation index quantifying departures from HWE, and γ_1, 3 quantifies departures from gene-environment independence. HWE is $f = 0$, and G-E independence is γ_1, 3 = 0. Monte Carlo standard errors of MSEs (biases) in the table range from 0.001–0.016 (0.004–0.019).

Table 3

MSE (bias in parenthesis) over 1,000 simulations: recessive model.

		$N_1 = N_0 = 300$ MSE (bias)		$N_1 = N_0 = 600$ MSE (bias)		$N_1 = N_0 = 1{,}000$ MSE (bias)	
		H	**H × X**	**H**	**H × X**	**H**	**H × X**
$f = 0$	$\hat{\beta}_{\text{free}}$	0.15 (0.03)	0.59 (0.04)	0.08 (−0.01)	0.26 (0.02)	0.05 (0.01)	0.14 (0.00)
$\gamma_{1,3} = 0$	$\hat{\beta}_{\text{model}}$	0.09 (−0.01)	0.20 (−0.07)	0.04 (−0.03)	0.10 (−0.05)	0.03 (−0.01)	0.05 (−0.04)
	$\hat{\beta}_{\text{HWD}}$	0.10 (−0.04)	0.20 (−0.03)	0.04 (−0.03)	0.10 (−0.06)	0.03 (−0.01)	0.05 (−0.04)
	$\hat{\beta}_{\text{EB1}}$	0.12 (0.01)	0.43 (0.01)	0.06 (−0.01)	0.20 (0.00)	0.04 (0.01)	0.11 (−0.02)
	$\hat{\beta}_{\text{EB2}}$	0.12 (0.01)	0.39 (−0.01)	0.06 (−0.02)	0.17 (−0.01)	0.04 (0.01)	0.10 (−0.02)
	$\hat{\beta}_{\text{PL1}}$	0.11 (0.00)	0.37 (0.00)	0.06 (−0.01)	0.18 (−0.01)	0.04 (0.01)	0.11 (−0.02)
	$\hat{\beta}_{\text{PL2}}$	0.12 (0.02)	0.43 (0.00)	0.07 (−0.01)	0.22 (0.00)	0.05 (0.01)	0.13 (−0.01)
$f = 0.05$	$\hat{\beta}_{\text{free}}$	0.15 (−0.02)	0.51 (0.03)	0.07 (0.00)	0.21 (0.02)	0.04 (0.01)	0.11 (−0.01)
$\gamma_{1,3} = 0$	$\hat{\beta}_{\text{model}}$	0.11 (0.14)	0.19 (−0.05)	0.06 (0.16)	0.08 (−0.05)	0.05 (0.18)	0.04 (−0.04)
	$\hat{\beta}_{\text{HWD}}$	0.10 (−0.04)	0.19 (−0.05)	0.04 (−0.01)	0.08 (−0.05)	0.02 (0.00)	0.04 (−0.04)
	$\hat{\beta}_{\text{EB1}}$	0.13 (0.02)	0.40 (0.01)	0.06 (0.03)	0.17 (0.01)	0.04 (0.04)	0.09 (−0.02)
	$\hat{\beta}_{\text{EB2}}$	0.12 (0.03)	0.35 (0.00)	0.06 (0.04)	0.14 (−0.01)	0.04 (0.05)	0.07 (−0.03)
	$\hat{\beta}_{\text{PL1}}$	0.12 (0.03)	0.34 (−0.03)	0.06 (0.04)	0.16 (−0.02)	0.04 (0.03)	0.09 (−0.03)
	$\hat{\beta}_{\text{PL2}}$	0.14 (0.00)	0.40 (−0.01)	0.06 (0.01)	0.18 (0.00)	0.04 (0.01)	0.10 (−0.02)
$f = 0$	$\hat{\beta}_{\text{free}}$	0.19 (−0.02)	0.85 (0.03)	0.09 (0.01)	0.52 (0.01)	0.06 (0.00)	0.25 (0.01)
$\gamma_{1,3} = -0.4$	$\hat{\beta}_{\text{model}}$	0.10 (0.01)	0.89 (−0.72)	0.05 (0.04)	0.73 (−0.72)	0.03 (0.05)	0.65 (−0.73)
	$\hat{\beta}_{\text{HWD}}$	0.10 (0.02)	0.87 (−0.71)	0.05 (0.04)	0.72 (−0.72)	0.04 (0.06)	0.65 (−0.73)
	$\hat{\beta}_{\text{EB1}}$	0.15 (−0.02)	0.74 (−0.15)	0.08 (0.00)	0.50 (−0.12)	0.05 (0.00)	0.26 (−0.08)
	$\hat{\beta}_{\text{EB2}}$	0.14 (−0.02)	0.71 (−0.25)	0.07 (0.00)	0.48 (−0.20)	0.05 (0.01)	0.26 (−0.15)
	$\hat{\beta}_{\text{PL1}}$	0.14 (0.01)	0.76 (−0.34)	0.07 (0.02)	0.51 (−0.19)	0.05 (0.02)	0.26 (−0.10)

| | | $N_1 = N_0 = 300$ | | $N_1 = N_0 = 600$ | | $N_1 = N_0 = 1{,}000$ | |
| | | MSE (bias) | | MSE (bias) | | MSE (bias) | |
		H	$H \times X$	H	$H \times X$	H	$H \times X$
	$\widehat{\beta}_{\text{PL2}}$	0.16 (0.01)	0.71 (−0.25)	0.08 (0.01)	0.47 (−0.12)	0.06 (0.01)	0.24 (−0.04)
$f = 0.05$	$\widehat{\beta}_{\text{free}}$	0.17 (0.01)	0.74 (0.00)	0.08 (−0.04)	0.45 (0.09)	0.05 (0.00)	0.26 (0.02)
$\gamma_{1,3} = -0.4$	$\widehat{\beta}_{\text{model}}$	0.16 (0.23)	0.84 (−0.72)	0.11 (0.25)	0.69 (−0.73)	0.10 (0.27)	0.64 (−0.74)
	$\widehat{\beta}_{\text{HWD}}$	0.12 (0.02)	0.87 (−0.73)	0.05 (0.04)	0.70 (−0.73)	0.04 (0.07)	0.64 (−0.74)
	$\widehat{\beta}_{\text{EB1}}$	0.15 (0.05)	0.65 (−0.17)	0.07 (0.00)	0.43 (−0.03)	0.05 (0.04)	0.27 (−0.07)
	$\widehat{\beta}_{\text{EB2}}$	0.15 (0.06)	0.62 (−0.26)	0.08 (0.02)	0.41 (−0.11)	0.05 (0.08)	0.27 (−0.13)
	$\widehat{\beta}_{\text{PL1}}$	0.15 (0.09)	0.67 (−0.36)	0.08 (0.02)	0.43 (−0.11)	0.05 (0.04)	0.27 (−0.09)
	$\widehat{\beta}_{\text{PL2}}$	0.16 (0.05)	0.61 (−0.25)	0.08 (−0.01)	0.41 (−0.02)	0.05 (0.01)	0.25 (−0.03)

NOTE: $\widehat{\beta}_{\text{model}}$: model-based estimator; $\widehat{\beta}_{\text{free}}$: model-free estimator; $\widehat{\beta}_{\text{HWD}}$: estimator accounting for Hardy-Weinberg disequilibrium; $\widehat{\beta}_{\text{EB1}}$ and $\widehat{\beta}_{\text{EB2}}$: multivariate and component-wise empirical-Bayes estimators; $\widehat{\beta}_{\text{PL1}}$ and $\widehat{\beta}_{\text{PL2}}$: L_1 and L_2 penalized estimators. Here f is the fixation index quantifying departures from HWE, and $\gamma_{1,3}$ quantifies departures from gene-environment independence. HWE is $f = 0$, and G-E independence is $\gamma_{1,3} = 0$. Monte Carlo standard errors of MSEs (biases) in the table range from 0.002–0.033 (0.006–0.027).

Table 4

Results for haplotype-environment interaction analyses of the colorectal adenoma study (Example 1) and the prostate cancer study (Example 2). The full models fitted are described in Section 6.1

Method	Interaction	$\hat{\beta}$	SE	p value	Interaction	$\hat{\beta}$	SE	p-value
	Example 1				**Example 2**			
Model-free	101010*Smk1	0.149	0.217	0.494	000*VD	−0.206	0.123	0.093
	101010*Smk2	−0.558	0.279	0.045	001*VD	0.127	0.085	0.135
Model-based	101010*Smk1	0.076	0.153	0.618	000*VD	−0.188	0.080	0.019
	101010*Smk2	−0.291	0.181	0.108	001*VD	0.046	0.061	0.453
EB	101010*Smk1	0.090	0.173	0.602	000*VD	−0.189	0.082	0.021
	101010*Smk2	−0.461	0.295	0.118	001*VD	0.095	0.089	0.285
PL	101010*Smk1	0.125	0.182	0.490	000*VD	−0.188	0.080	0.019
	101010*Smk2	−0.492	0.249	0.048	001*VD	0.121	0.082	0.143

NOTE: *model-free*: assuming HWE and haplotype-environment independence conditional on genotype; *model-based*: assuming unconditional HWE and haplotype-environment independence; *EB*: the component-wise empirical Bayes method; *PL*: the penalized likelihood method based on Lasso; *Smk1*: former smoker; *Smk2*: current smoker; *VD*: [25(OH)D)] level.