Statistics and Its Interface Volume 7 (2014) 43–50

Estimation of genetic effects incorporating prior information

Ao Yuan*, Qizhai Li†, Jing Qin, and Gang Zheng‡

We study estimations of the genetic effect of a marker by adjusting out covariates and incorporating the results of previous potentially heterogenous studies of the same genetic marker. Without prior information on the covariates, the procedures are based on both frequentist and Bayesian methods by simultaneously maximizing the likelihood function for the coefficients of the covariates and minimizing the loss function for the genetic effect, and hence are regarded as hybrid estimations. Although we focus on an application to case-control genetic association studies, we describe a general method for various types of traits. For the application, we show that the proposed hybrid inference based on the prospective sampling can be applied to retrospectively collected case-control data. Simulations and applications using hybrid inference are presented.

AMS 2000 subject classifications: 62F12; 62F15; 62P10.
Keywords and phrases: Asymptotics, Bayesian analysis, Case-control genetic study, Hybrid inference, Loss function, Power prior, Prospective and retrospective.

1. INTRODUCTION

In case-control genetic association studies, the main goal is to identify genetic markers, single nucleotide polymorphisms (SNPs), that are associated with a common disease. In this paper, we are interested in estimating the genetic effect of a SNP, often in terms of the odds ratio (OR), that is associated with a disease after adjusting out confounding covariates.

In the era of genome-wide association studies (GWASs), millions of SNPs have been genotyped and studied for many common diseases. Information of these genetic studies and GWAS data are often publicly available. For example, GWASs funded by National Institutes of Health (NIH) and the Wellcome Trust Case-Control Consortium (WTCCC) [15] are available through the database of Genotypes and Phenotypes (dbGaP) and the WTCCC websites, respectively. Incorporating past results into the analysis of a new genetic association study is expected to improve the power to detect true associated SNPs. Estimations of the ORs of the associated SNPs incorporating the past results would provide more accurate sample size calculations in designing genetic association studies in future.

The results from past studies can be used to elicit an informative prior for the analysis of a new study, which is referred to as the "current study" in the following. However, it is highly possible that the data from the past studies and the data of the current study are drawn from different populations due to different definitions of covariates and population characteristics. Hence, heterogeneity between the past and current studies often exists.

To our best knowledge, information from past genetic association studies has been incorporated into the analysis of the current study in two ways. The first method is to elicit a mixture distribution with positive probabilities for <1, 0, and >1 log-OR [4], where most mass of the prior probability is placed on log-OR = 0. Then a full Bayesian analysis is conducted using the samples drawn from the posterior distribution. The second one is for a multi-marker association study to incorporate the probabilities of the markers being associated with the disease [3]. The first approach is a full Bayesian analysis for the genetic effect, while the second one is a frequentist approach, but does not consider the same problem as we do here.

We consider a hybrid inference which involves both Bayesian and frequentist procedures. Our hybrid estimation is an extension of Yuan (2009) [16], who considered a hybrid estimate for the parameters (α, β) based on the likelihood $f(Y|\alpha, \beta)$, where Y is the response variable and β is the parameter of interest and α is a nuisance parameter. We consider a hybrid estimate for the conditional likelihood $f(Y|X, G, \alpha, \beta)$, where X is the vector of covariates with coefficients α, G is the genotype of the SNP (or the risk factor for a general regression model) with a coefficient β. Like [16], we assume information from a past study for β is available, but, information for the covariates is often not available due to different studies designs and samples. One can put a non-informative prior on α, so that a full Bayesian analysis can be performed, but we use the frequentist analysis on α. This makes the modeling and computation simpler. Also, an incorrect prior can cause misleading results when sample size

*Partially supported by the National Institute on Minority Health and Health Disparities of the National Institutes of Health under Award Number G12MD007597.
†Partially supported by National Nature Science Foundation of China, No. 61134013.
‡Corresponding author.

is not large. Our setting is general as Y can be either binary or quantitative. We also consider the 0–1 loss function that was not considered in [16]. We prove that the hybrid inference is asymptotically first-order equivalent to the classical frequentist inference. Hence we examine the small sample benefits through simulations. For applications, we consider a case-control association study. Hence, we study whether or not the hybrid inference based on prospective sampling can be applied to the retrospectively sampled case-control data. How to use a power prior [6, 2] in the hybrid inference to adjust heterogeneity between the past and current studies or to weight more on the current study is discussed. Application to real GWAS data illustrates the use of hybrid inference.

2. MODELS AND DATA

Let $G = (g_0, g_1, g_2) = (AA, AB, BB)$ be the genotypes of a SNP, $Y = 0$ (1) stand for a control (case), and X denote covariates. The retrospective and prospective likelihoods are given by $P(X, G|Y)$ and $P(Y|X, G)$, respectively, whose relationship can be described as a two-sample semi-parametric model with a biased sampling [12] or as a mixture model [11]. We study hybrid inference based on the prospective model, which is equivalent to using the retrospective model (Sect. 4).

Suppose r cases and s controls are obtained whose genotype counts for G are (r_0, r_1, r_2) in cases and (s_0, s_1, s_2) in controls with $n = r + s$ samples. Let $c(g_0) = 0$, $c(g_1) = c$, and $c(g_2) = 1$, where c is determined by the genetic model. For the recessive (REC), additive (ADD), and dominant (DOM) diseases [17], we use $c = 0$, $1/2$, and 1, respectively. Let Y_j be the outcome of the jth individual $(j = 1, \dots, n)$ with covariates $X_j = (X_{j1}, \dots, X_{jk})^T$, where $X_{j1} = 1$ and $k \geq 1$, and genotype G_j. Denote $X^n = (X_1, \dots, X_n)$, $G^n = (G_1, \dots, G_n)$ and $Y^n = (Y_1, \dots, Y_n)$. Then the likelihood function $L(\alpha^T, \beta) = f(Y^n|X^n, G^n, \alpha^T, \beta)$ can be written as

$$(1)\qquad L(\alpha^T, \beta) = \frac{\exp\{\sum_{j=1}^r \alpha^T X_j + \beta \sum_{i=0}^2 r_i c_i\}}{\prod_{j=1}^n \left[1 + \exp\{\alpha^T X_j + \beta c(G_j)\}\right]},$$

where $G_j = g_0$, g_1 or g_2, β is the log-OR associated with the SNP, and α contains the rest of the parameters. Under the null hypothesis of no association H_0, $\beta = 0$. Denote $\theta = (\alpha^T, \beta)$ and the dimension of θ is $\dim(\theta) = k + 1 \geq 2$.

3. HYBRID ESTIMATES

The discussion is this section is based on a more general likelihood than the one in Sect. 2.

3.1 Definition

Let $\pi(\cdot)$ be the prior density for β, $w(\cdot, \cdot)$ be the loss function for inferring β, and $d = d(Y^n|X^n, G^n)$ be the decision for β based on Y^n given X^n and G^n. Denote the parameter spaces for α and β as Λ and Γ, respectively. For a fixed α, the Bayes estimate β^* for β is given by

$$\begin{aligned}\beta^* &= \beta^*(Y^n|X^n, G^n, \alpha^T)\\ &= \arg\inf_{d\in\Gamma} \int_\Gamma w(d, \beta) f(Y^n|X^n, G^n, \theta)\pi(\beta)d\beta,\end{aligned}$$

and for a fixed β, the maximum likelihood estimate (MLE) α^* for α is given by

$$\alpha^* = \alpha^*(Y^n|X^n, G^n, \beta) = \arg\sup_{\alpha\in\Lambda} f(Y^n|X^n, G^n, \theta).$$

Following Yuan (2009) [16], the hybrid estimate of $\theta = (\alpha^T, \beta)$, denoted by $\tilde{\theta} = (\hat{\alpha}^T, \check{\beta})$, satisfies

$$(2)\qquad (\hat{\alpha}^T, \check{\beta}) = \arg\sup\inf{}_{\alpha\in\Lambda, d\in\Gamma} \int_\Gamma w(d, \beta) f(Y^n|X^n, G^n, \theta)\pi(\beta)d\beta.$$

Denote $H(\alpha, d) = w(d, \beta) f(Y^n|X^n, G^n, \theta)\pi(\beta)$. Then $H(\hat{\alpha}, \check{\beta}) \leq H(\hat{\alpha}, d)$ for $\forall d \in \Gamma$ and $H(\hat{\alpha}, \check{\beta}) \geq H(\alpha, \check{\beta})$ for $\forall \alpha \in \Lambda$. The hybrid estimate $\tilde{\theta}$ generally exists and is locally unique because it can be formulated as a Bayesian estimator under the 0–1 loss with a constant prior for α. Note that the operations inf and sup in (2) are applied jointly and simultaneously. Generally, $\sup\inf_{\alpha\in\Lambda, d\in\Gamma}$ is not equivalent to $\sup_{\alpha\in\Lambda}(\inf_{d\in\Gamma})$ or $\inf_{d\in\Gamma}(\sup_{\alpha\in\Lambda})$ [16].

3.2 Finding hybrid estimate given a loss function

In general, there is no closed form for $(\hat{\alpha}^T, \check{\beta})$ satisfying (2). Some details of computations under three common loss functions are discussed here. Denote the conditional posterior density for β as $\pi(\beta|Y^n, \alpha^T, X^n, G^n) = f(Y^n|X^n, G^n, \theta)\pi(\beta)/m(Y^n|X^n, G^n, \alpha^T)$, where the conditional marginal density is given by $m(Y^n|X^n, G^n, \alpha^T)$.

Given the quadratic loss $w(d, \beta) = (d - \beta)^2$, for a fixed $\hat{\alpha}$, $\check{\beta}$ is the posterior mean. Thus,

$$(3)\qquad \begin{aligned}\hat{\alpha} &= \arg\sup_{\alpha\in\Lambda} l(\alpha^T, \check{\beta}),\\ \check{\beta} &= E(\beta|Y^n, \hat{\alpha}^T, X^n, G^n),\end{aligned}$$

where $l(\alpha^T, \beta) = \log L(\alpha^T, \beta)$ is the conditional log-likelihood function. Generally, $(\hat{\alpha}^T, \check{\beta})$ can not be evaluated in a closed form, and a numerical method is required for their computation.

Given the absolute error loss $w(d, \beta) = |d - \beta|$, for a fixed $\hat{\alpha}$, $\check{\beta}$ is the posterior median. Thus, $\hat{\alpha}$ satisfies (3) and $\check{\beta}$ is given by $\check{\beta} = \text{Med}(\beta|Y^n, \hat{\alpha}^T, X^n, G^n)$. There is no closed form for the hybrid estimate and a numerical method is required to find $(\hat{\alpha}^T, \check{\beta})$.

For the 0–1 loss, $w(d, \beta) = 0$ if $d = \beta$, and $w(d, \beta) = 1$ otherwise. Then, for a fixed $\hat{\alpha}$, $\check{\beta}$ is the posterior mode.

Hence $\hat{\alpha}$ still satisfies (3) but $\check{\beta}$ is given by

$$(4) \qquad \begin{aligned}\check{\beta} &= \arg\sup_{\beta\in\Gamma} \pi(\beta|Y^n, \hat{\alpha}^T, X^n, G^n) \\ &= \arg\sup_{\beta\in\Gamma} \{f(Y^n|X^n, G^n, \hat{\alpha}^T, \beta)\pi(\beta)\}.\end{aligned}$$

Consequently,

$$(\hat{\alpha}^T, \check{\beta}) = \arg\sup \sup_{\alpha\in\Lambda,\beta\in\Gamma} \big(l(\alpha^T, \beta) + \log \pi(\beta)\big).$$

Hence, $\tilde{\theta}$ is simpler to compute than using the previous two loss functions because $\tilde{\theta}$ can be regarded as the MLE from $l^*(\alpha^T, \beta) = l(\alpha^T, \beta) + \log \pi(\beta)$.

3.3 Asymptotic properties

Let (α_0, β_0) be the true parameters in the model. The proofs of the following results are given in the appendix.

Proposition 1. *Under the regularity conditions A1–A9 of [16], when the squared error loss function is used, we have (i) $(\hat{\alpha}^T, \check{\beta}) \to (\alpha_0^T, \beta_0)$ (a.s.), and (ii) $\sqrt{n}(\hat{\alpha}^T - \alpha_0^T, \check{\beta} - \beta_0) \xrightarrow{D} N(0, I^{-1}(\theta_0))$, where $I(\theta)$ is the Fisher information matrix for a single sample.*

Proposition 2. *Assume that 1) there is a convex set A such that $\inf_{\theta\in A} |I(\theta)| > 0$, $(\alpha_0^T, \beta_0) \in A$ and $(\hat{\alpha}^T, \check{\beta}) \in A$ for all large n, 2) on A, $0 < \pi(\cdot) < \infty$, 3) the first and second derivatives, $\pi^{(k)}(\cdot)$ $(k = 1, 2)$, are bounded and away from zero on A, 4) $I(\cdot)$ is continuous at θ_0, and 5) $\partial \sum_G \int f(y, x, G|\theta)dydx/\partial\theta = \sum_G \int \partial f(y, x, G|\theta)/\partial\theta dydx$. Then under the 0–1 loss, we have (i) $(\hat{\alpha}^T, \check{\beta}) \to (\alpha_0^T, \beta_0)$ (a.s.), and (ii) $\sqrt{n}(\hat{\alpha}^T - \alpha_0^T, \check{\beta} - \beta_0) \xrightarrow{D} N(0, I^{-1}(\theta_0))$.*

From both results and the Bernstein-von Mises theorem [8, 9], the Bayes estimator, the MLE and the hybrid estimate of θ for the conditional model $f(Y^n|X^n, G^n, \theta)$ are asymptotically first-order equivalent and efficient. The prior $\pi(\cdot)$ for β is elicited based on a past study (or past studies). Without loss of generality, we assume that $\pi(\cdot)$ is based on a single past study with sample size m. The above asymptotic results require $m/n \to 0$ as $n \to \infty$. Hence, the contribution of the past study asymptotically vanishes. However, finite-sample properties of the hybrid and frequentist inferences are generally different. In practice, when we calculate the asymptotic variance or construct a confidence interval for β using hybrid inference, $\theta_0 = (\alpha_0^T, \beta_0)$ is replaced by the hybrid estimate $(\hat{\alpha}^T, \check{\beta})$.

4. EQUIVALENCE TO USING THE RETROSPECTIVE LIKELIHOOD

Seaman and Richardson (2004) [13] proved the equivalence of using the prospective and retrospective likelihoods for a full Bayesian inference with some mild conditions on the prior distributions. Here we show an equivalence for the hybrid inference.

The results presented before are based on a prospective likelihood. Denote $\alpha^T = (\alpha_1, \tilde{\alpha}^T)$ where $\tilde{\alpha}^T = (\alpha_2, \ldots, \alpha_k)$ ($\tilde{\alpha}^T$ vanishes if $k = 1$), $X_j = (1, \tilde{X}_j)$, $z_j = c(G_j)$, $\delta = 0$ for controls and 1 for cases, and $Y_{\delta j} = Y_j$ where $Y_{0j} = 0$ and $Y_{1j} = 1$. Then $f_P(\alpha_1, \tilde{\alpha}^T, \beta) = f(Y^n|X^n, G^n, \theta)$ is given by

$$(5) \quad f_P(\alpha_1, \tilde{\alpha}^T, \beta) = \prod_{j=1}^{n} \prod_{\delta=0}^{1} \left\{ \frac{\exp\left(\delta\alpha_1 + \delta\tilde{\alpha}^T \tilde{X}_j + \delta\beta z_j\right)}{\sum_{l=0}^{1} \exp\left(l\alpha_1 + l\tilde{\alpha}^T \tilde{X}_j + l\beta z_j\right)} \right\}^{Y_{\delta j}}.$$

The retrospective likelihood is $f_R(\theta^*) = f(X^n, G^n|Y^n, \theta^*)$, where $\theta^* = (\tilde{\alpha}_1, \tilde{\alpha}^T, \beta)$ and $\tilde{\alpha}_1 \neq \alpha_1$. However, based on the parameterization of Prentice and Pyke (1979) [10], subject to a normalization constant, the retrospective likelihood can be written as

$$(6) \qquad f_R(\tilde{\alpha}_1, \tilde{\alpha}^T, \beta) = f_P(\tilde{\alpha}_1, \tilde{\alpha}^T, \beta)q(X^n, G^n),$$

where $q(X^n, G^n)$ is a non-specified density (mass) function and does not involve the parameter $\tilde{\theta}$. It follows that the posterior density of β based on (6), $\pi_R(\beta|\cdot)$, has the same form as that based on (5), $\pi(\beta|\cdot)$, that is, $\pi_R(\beta|Y^n, \tilde{\alpha}_1, \tilde{\alpha}^T, G^n) = \pi(\beta|Y^n, \tilde{\alpha}_1, \tilde{\alpha}^T, G^n)$. The nonparametric MLE of $q(X^n, G^n)$, denoted as $\hat{q}(X^n, G^n)$, is the empirical distribution, which assigns mass m/n to any observed value of (X, G) with multiplicity m and 0 otherwise [10].

The following approach to show the equivalence of the hybrid inference using f_R and f_P is based on the profile retrospective likelihood function. One advantage of this approach is that there is no condition specified for the priors. Let Λ_1 and Λ_2 be subspaces for $\tilde{\alpha}_1$ and $\tilde{\alpha}$, respectively. The profile retrospective likelihood of (6) after maximization of q is given by $\hat{f}_R(\tilde{\alpha}_1, \tilde{\alpha}^T, \beta) = f_P(\tilde{\alpha}_1, \tilde{\alpha}^T, \beta)\hat{q}(X^n, G^n)$. Then the hybrid estimate is given by

$$\begin{aligned}(\hat{\tilde{\alpha}}_1, \hat{\tilde{\alpha}}^T, \check{\beta}) = {} & \arg\sup \sup \inf_{\tilde{\alpha}_1\in\Lambda_1, \tilde{\alpha}^T\in\Lambda_2, d\in\Gamma} \\ & \int w(d, \beta)\hat{f}_R(\tilde{\alpha}_1, \tilde{\alpha}^T, \beta)\pi(\beta)d\beta \\ = {} & \arg\sup \sup \inf_{\tilde{\alpha}_1\in\Lambda_1, \tilde{\alpha}^T\in\Lambda_2, d\in\Gamma} \\ & \int w(d, \beta)f_P(\tilde{\alpha}_1, \tilde{\alpha}^T, \beta)\hat{q}(X^n, G^n)\pi(\beta)d\beta.\end{aligned}$$

Given either of the three loss functions (quadratic, absolute error, and 0–1) and $\check{\beta}$, $(\hat{\tilde{\alpha}}_1, \hat{\tilde{\alpha}}^T)$ satisfies $(\hat{\tilde{\alpha}}_1, \hat{\tilde{\alpha}}^T) = \arg\sup \sup_{\tilde{\alpha}_1\in\Lambda_1, \tilde{\alpha}^T\in\Lambda_2} f_P(\tilde{\alpha}_1, \tilde{\alpha}^T, \check{\beta})\hat{q}(X^n, G^n) = \arg\sup \sup_{\tilde{\alpha}_1\in\Lambda_1, \tilde{\alpha}^T\in\Lambda_2} f_P(\tilde{\alpha}_1, \tilde{\alpha}^T, \check{\beta})$, as if they were obtained from a prospective likelihood $f_P(\tilde{\alpha}_1, \tilde{\alpha}^T, \check{\beta})$. Given $(\hat{\tilde{\alpha}}_1, \hat{\tilde{\alpha}}^T, \hat{q})$, $\check{\beta}$ is the posterior mean, median and mode with the posterior density $\pi_R(\beta|Y^n, \hat{\tilde{\alpha}}_1, \hat{\tilde{\alpha}}^T, G^n)$, which is the same as the posterior density of β based on the prospective likelihood. Thus, the hybrid estimate can be obtained from the prospective model.

5. CHOOSING PRIORS

For our application to case-control genetic association studies, β is the log-OR. If the data of the past study are available, let z_m and τ_m^2 be the MLE of β and the consistent estimate of its asymptotic variance, respectively, where m is the sample size of the past study. Suppose the genotype counts in cases and controls of the past study are reported as (r_{00}, r_{01}, r_{02}) and (s_{00}, s_{01}, s_{02}). Then the MLE for the log-OR and its asymptotic variance have closed forms for a given genetic model. Under the REC model, $\exp(z_m) = r_{02}(s_{00} + s_{01})/\{s_{02}(r_{00} + r_{01})\}$ and $\tau_m^2 = 1/(r_{00} + r_{01}) + 1/r_{02} + 1/s_{02} + 1/(s_{00} + s_{01})$, under the DOM model, $\exp(z_m) = (r_{01} + r_{02})s_{00}/\{(s_{01} + s_{02})r_{00}\}$ and $\tau_m^2 = 1/r_{00} + 1/(r_{01} + r_{02}) + 1/s_{00} + 1/(s_{01} + s_{02})$, and under the ADD model, $\exp(z_m) = [(2r_{02}+r_{01})(2s_{00}+s_{01})/\{(2s_{02}+s_{01})(2r_{00} + r_{01})\}]^2$ and $\tau_m^2 = 4\{1/(2r_{02} + r_{01}) + 1/(2s_{00} + s_{01}) + 1/(2s_{02} + s_{01}) + 1/(2r_{00} + r_{01})\}$. We do not need individual genetic data when calculating (z_m, τ_m^2).

If the data of the past study are not available, z_m is often reported but τ_m^2 may not be reported although it can be obtained from the p-value. However, in the following example, τ_m^2 cannot be even converted from the p-value and z_m. Swaroop et al. (2007) [14] reported meta-analysis results for five candidate-genes for age-related macular degeneration (AMD). In their Table 1, they reported the number of studies (H), the total sample size (M) of H studies, allele frequencies in cases (p_1) and controls (p_0), the estimates of OR ($\exp(z_m)$), and p-values. For example, one SNP rs1061170 located in CFH gene had $\exp(z_m) = 2.00$ with $H = 14$, $M = 10{,}930$, $p_1 = 0.435$, $p_0 = 0.639$, and the p-value of the meta-analysis was reported as $<10^{-100}$. The asymptotic variance was not given and the genetic model was not specified either. Since no explicit p-value was reported, τ_m^2 cannot be estimated. Even if an explicit p-value were reported, one would be doubtful about the accuracy of such a small p-value. In this case, we can approximate τ_m^2 with Hardy-Weinberg equilibrium (HWE). The derivation for the ADD model is given in the appendix. Let $\widetilde{m} = M/H$. For the ADD, REC, and DOM models, τ_m^2 can be approximated by $\tau_m^2 = 4\widetilde{m}^{-1}[\{p_1(1 - p_1)\}^{-1} + \{p_0(1 - p_0)\}^{-1}]$, $\tau_m^2 = (\widetilde{m}/2)^{-1}[\{p_0^2(1 - p_0^2)\}^{-1} + \{p_1^2(1 - p_1^2)\}^{-1}]$, and $\tau_m^2 = (\widetilde{m}/2)^{-1}[\{(1-p_0)^2p_0(2-p_0)\}^{-1} + \{(1-p_1)^2p_1(2-p_1)\}^{-1}]$. If only population allele frequency is given, p_1 and p_0 can be both replaced by the known or estimated population allele frequency.

Let $f(\beta|z_m, \tau_m^2)$ be the normal density with mean z_m and variance τ_m^2. Let $\pi_0(\beta)$ be the prior without using the past data. Then the posterior prior is proportional to $\pi(\beta) = f(\beta|z_m, \tau_m^2)\pi_0(\beta)$. We choose $\pi_0(\beta) = 1$ here. If both the past study and the current study are based on the same study population and there is no concern of the quality of the data of either study, when $m \to \infty$, z_m converges to the true value of β but $\tau_m^2 \to 0$. The latter causes the problem applying the two asymptotic results. In practice, m may not be small relative to the sample size of the current study, especially many genetic studies with large sample sizes (e.g., $m \geq 5{,}000$) have been reported. Besides, it is likely that the data of the past study are not comparable to the data of the current study. To handle this situation, the power prior of Ibrahim and Chen (2000) [6] can be considered.

The power prior introduces a power parameter $\gamma \in [0, 1]$ in the prior as [6]

$$\pi(\beta|\gamma) = \pi^\gamma(\beta)\pi_0(\beta) = \pi^\gamma(\beta),$$

with $\pi_0(\beta) = 1$, where $\pi(\beta)$ is given before. The parameter γ controls the contribution of the past study or adjusts the heterogeneity of the past and current studies.

When $\pi(\beta)$ is $N(z_m, \tau_m^2)$,

$$\pi^\gamma(\beta) \sim \exp\left\{-\frac{(\beta - z_m)^2}{2\tau_m^2/\gamma}\right\}$$

indicates that the power prior has a normal density with the same mean but an inflated variance.

To illustrate the effect of γ, we consider τ_m^2 under the ADD model. Let $\phi \in (0, 1)$ be a constant ratio of the number of cases over that of controls in the past study for any sample size m and denote, in the past study, the genotype probabilities in cases as (p_{00}, p_{01}, p_{02}) and in controls as (q_{00}, q_{01}, q_{02}). Then

$$(7) \qquad \frac{\tau_m^2}{\gamma} \approx \frac{4h(\mathbf{p}, \mathbf{q})}{m\gamma} = \frac{4h(\mathbf{p}, \mathbf{q})}{m'} \approx \tau_{m'}^2,$$

where $\tau_{m'}^2 = \tau_m^2/\gamma$, $m' = m\gamma \leq m$, and

$$h(\mathbf{p}, \mathbf{q}) = \frac{1 + \phi^{-1}}{2p_{02} + p_{01}} + \frac{1 + \phi}{2q_{00} + q_{01}} + \frac{1 + \phi}{2q_{02} + q_{01}} + \frac{1 + \phi^{-1}}{2p_{00} + p_{01}}$$

does not involve m or m'. Therefore, the power parameter γ controls the input of the past study by reducing its sample size m to m'. Optimal choice of γ has been studied by Bhattachrya (2009) [2] using the criteria based on Kullback-Leibler divergence. The optimal γ described below belongs to [2] (remark 3), which can be written as $\gamma = \lambda/(1 + \lambda)$, where λ solves

$$\inf_{\lambda \in R^+}\left[\lambda(\delta - e) + (1 + \lambda)\log\int L(Y^n|\theta)\{f(\beta|z_m, \tau_m^2)\}^{\frac{\lambda}{1+\lambda}}d\theta\right],$$

where $L(Y^n|\theta) = f(Y^n|X^n, G, \theta)$ is the likelihood of the current study, $\theta = (\alpha, \beta)$, $e = \log\int L(Y^n|\theta)f(\beta|z_m, \tau_m^2)d\theta$, and δ is user-specified. If the data of the past study are judged to have poor quality, one may choose $\delta = 0.01 \times \int L(Y^n|\theta)\log\{L(Y^n|\theta)/f(\beta|z_m, \tau_m^2)\}d\theta$ [2]. We can apply this rule because we want to place more weight on the current study than the past one.

Both δ and γ depend on sample sizes m and n as well as G. Thus, different δ and γ have to be found and used for different SNPs. In our simulation and application presented later, however, we choose a single power prior given by $m' = \min(200, m, n/3)$ if $m \leq 1{,}000$ and $m' = \min(300, n/3)$ otherwise. Thus, our choice of m' is about no more than 1/3 of n. With this fixed rule, the power prior that we apply to case-control genetic association studies is $\pi(\beta|\gamma) = N(z_m, \tau^2_{m'})$.

6. NUMERICAL RESULTS

6.1 Simulations

The data were simulated using the procedures given in Zheng et al. (2012) [17]. HWE was assumed to hold in the population. Three minor allele frequencies (MAFs) were 0.15, 0.30 and 0.45, which were also the frequencies for the risk allele under the alternative hypothesis H_1. For the past study, genotype counts (r_{00}, r_{01}, r_{02}) and (s_{00}, s_{01}, s_{02}) with sample size $m = 300$ were first simulated under H_0 (no genetic model) or H_1 for a given genetic model. The statistics $(z_m, \tau^2_{m'})$ were calculated for the same genetic model under H_0 and under H_1 with $m' = 200$ based on our rule for the power prior. Then, for the current study, genotype counts (r_0, r_1, r_2) and (s_0, s_1, s_2) were simulated under the same genetic model with sample size $n = 2{,}000$. Equal numbers of cases and controls were used in the simulations. The results were estimated based on 10,000 replicates, in each of which a past study was simulated and $(z_m, \tau^2_{m'})$ was calculated. The nominal level was 5% in the simulations. The log-ORs were 0, $\log(1.5) = 0.4055$, $\log(1.25) = 0.2231$ and $\log(1.5)$ under H_0, REC, ADD and DOM models, respectively. For each setting, we assumed a covariate taking 0 or 1 with true $\alpha_1 = \log(1.1)$. We calculate α_0 with the disease prevalence $\Pr(\text{case}) = 0.1$ by the standard logistic regression model. We considered the classical MLE and the hybrid estimate (HE) for β. The mean, mean squared error (MSE) and coverage probability (CP) are reported in Table 1. The results show that the means and CPs of both estimates are very close to the true values. However, the MSE of the HE is smaller than that of the MLE. In Table 2, we report the results of the HE for the covariate, which show that the HE for the covariate also has very good coverage and very small biases as expected.

6.2 Real applications

First, we used real GWAS data of the WTCCC (2007) [15] to calculate $(z_m, \tau^2_{m'})$ for a specific phenotype. Then we applied the hybrid inference to estimate the genetic effect for a candidate marker for AMD.

The WTCCC studied seven common diseases (bipolar disorder, coronary disease, Crohn's disease, hypertension, rheumatoid arthritic, type I diabetes, and type II diabetes) using Affymetrix 500K SNP platform. About 2,000 cases for each disease were collected with 3,000 shared controls. Genotype counts can be obtained from the WTCCC website, from which $(z_m, \tau^2_{m'})$ and useful power prior can be used for future genetic studies with $n > 3m'$. In this application, we focused on bipolar disorder. After standard quality control steps [15], we obtained 391,573 SNPs. Because the sample size $m = 5{,}000$ is quite large, we chose $m' = 300$

Table 1. Means, MSEs and CPs of the MLE and hybrid estimate (HE) for the log-OR of the genetic effect. The true values for the log-OR are 0, $\log(1.5) = 0.4055$, $\log(1.25) = 0.2231$ and $\log(1.5) = 0.4055$ under the null, REC, ADD and DOM models, respectively

	MLE			HE		
MAF	Mean	MSE	CP	Mean	MSE	CP
NULL						
0.15	0.0003	0.0079	0.951	−0.0001	0.0062	0.958
0.30	0.0012	0.0048	0.950	0.0008	0.0038	0.955
0.45	−0.0007	0.0040	0.952	−0.0009	0.0031	0.959
REC						
0.15	0.4116	0.0844	0.954	0.4056	0.0644	0.961
0.30	0.4069	0.0222	0.951	0.4059	0.0174	0.958
0.45	0.4063	0.0116	0.949	0.4058	0.0091	0.958
ADD						
0.15	0.2241	0.0076	0.949	0.2235	0.0059	0.956
0.30	0.2234	0.0046	0.950	0.2234	0.0036	0.960
0.45	0.2226	0.0042	0.952	0.2225	0.0032	0.959
DOM						
0.15	0.4070	0.0095	0.951	0.4065	0.0074	0.956
0.30	0.4050	0.0082	0.951	0.4053	0.0064	0.958
0.45	0.4025	0.0104	0.951	0.4030	0.0082	0.956

Table 2. Means, MSEs and CPs of the HE for the log-OR of the covariate. The true value is $\log(1.1) = 0.0953$ under the null, REC, ADD and DOM models, respectively

MAF	Mean	MSE	CP
NULL			
0.15	0.0954	0.0080	0.950
0.30	0.0958	0.0081	0.945
0.45	0.0960	0.0082	0.945
REC			
0.15	0.0951	0.0080	0.948
0.30	0.0950	0.0080	0.949
0.45	0.0952	0.0080	0.949
ADD			
0.15	0.0951	0.0081	0.950
0.30	0.0940	0.0081	0.949
0.45	0.0956	0.0082	0.948
DOM			
0.15	0.0952	0.0080	0.951
0.30	0.0953	0.0082	0.947
0.45	0.0958	0.0083	0.947

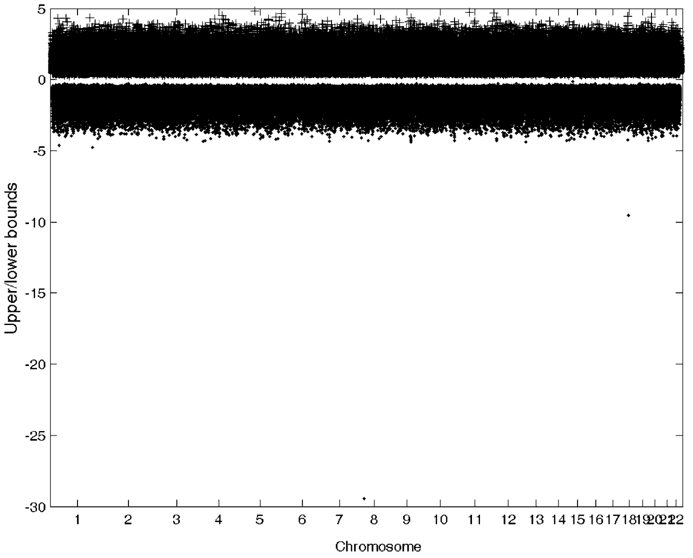

Figure 1. Plots of the upper $z_m + 2\tau_{m'}$ (+) and lower $z_m - 2\widehat{\tau}_{m'}$ (*) bounds for 391,573 SNPs for bipolar disorder [15]. Among all the intervals, there is only one on chromosome 8 that does not contain 0.

based on our rule for the power prior. Allele frequencies were also estimated using the original samples. Only the ADD model was considered. Then we calculated $z_m + 2\tau_{m'}$ (indicated by +) and $z_m - 2\tau_{m'}$ (indicated by *) and plotted the dots and *'s in Figure 1 along the physical locations and chromosomes of these SNPs. Among 391,573 intervals $(z_m - 2\tau_{m'}, z_m + 2\tau_{m'})$, there was only one interval that did not cover 0 (see Figure 1).

Our next application is AMD. Among the five candidate SNPs reported in [14], only one SNP rs10490924 was in a GWAS of AMD of Klein et al. (2005) [7]. This SNP is located in LOC387715 gene. We focused on this SNP for illustration. Its genotype counts are $(r_0, r_1, r_2) = (2, 17, 31)$ and $(s_0, s_1, s_2) = (12, 44, 40)$ with the total sample size $n = 146$. Without using any prior information, the MLEs of the OR and their 95% confidence intervals (CIs) under the REC, ADD and DOM models are 2.28 and (1.13, 4.60), 4.30 and (1.36, 13.67), and 3.43 and (0.74, 15.96), respectively (Table 3).

Based on the meta-analysis of 8 studies of this SNP, we have $H = 8$, $M = 8{,}473$, $p_0 = 0.207$, $p_1 = 0.420$, the estimate of the OR is 2.62, and the p-value in the meta-analysis $<10^{-100}$ [14]. There was no genetic model reported. So we took $z_m = \log(2.62) = 0.963$ for all the three genetic models. We calculated $\tau^2_{m'}$ using the formulas given before with $m' = 50 \approx n/3$ based on our rule for the power prior. The results are summarized in Table 3, which show $(z_m - 2\tau_{m'}, z_m + 2\tau_{m'})$ contains 0 for each genetic model at the 0.05 significance level. We also obtained the hybrid estimates of the OR under each genetic model and the corresponding 95% confidence intervals (CIs). Comparing the MLEs and the HEs and their confidence intervals (Table 3), the HEs of the OR yield similar but smaller point estimates, but have smaller variances so that the hybrid CIs are all smaller.

Table 3. Estimates of genetic effect of SNP rs10490924 with AMD

Model	$\tau^2_{m'}$	$(z_m - 2\tau_{m'}, z + 2\tau_{m'})$
REC	1.2506	$(-1.27, 3.20)$
ADD	0.8158	$(-0.84, 2.77)$
DOM	0.3506	$(-0.22, 2.15)$

Estimate	MLE		HE	
Model	OR	95% CI	OR	95% CI
REC	2.28	(1.13, 4.60)	2.37	(1.30, 4.32)
ADD	4.30	(1.36, 13.67)	3.72	(1.43, 9.69)
DOM	3.43	(0.74, 15.96)	3.14	(0.91, 10.81)

7. DISCUSSION

Incorporating the results from past genetic association studies into the analysis of a current genetic association study in a hybrid fashion has not been seen before, although general inference with both frequentist and Bayesian components has been discussed in the literature [1, 5]. Yuan (2009) [16] considered hybrid estimates with both frequentist and Bayesian components, but we extended his concept to also adjust out covariates and proposed to use the hybrid likelihood with the 0–1 loss function. In this paper, we focused on applications to genetic association studies, but analytic results and properties can be used for more general applications.

One key part of applying the hybrid inference in genetic association studies is the elicitation of an informative prior based on the results of past studies. For genetic studies, the past and current studies are likely heterogenous. Moreover, more weight should be placed on the current study than the past one. An existing approach of using the power prior [6, 2] was applied here, which is equivalent to inflate the variance or reduce the actual sample size of the past study to a smaller one. The optimal reduction rate was obtained [2] given a user-specified number reflecting the quality of the past data or the heterogeneity between the past and current studies. This rate needs to be determined by solving a non-linear optimality problem for each SNP. Hence, for the simplicity, we used a fixed rule to reduce the sample size of the past study by taking into account the sample size of the current study. Our rule, though simple, is not necessarily optimal, and its comparison to the optimal rule is not studied yet. But it worked well in our simulations reported here and in extra simulation studies for hybrid hypothesis testing (results are not reported here). It is worthwhile to further study how to obtain the optimal reduction rate for case-control genetic association studies. Even with the optimal reduction rate, sensitivity analysis with different rates is helpful to understand the robustness of using the power prior.

One may think of an alternative approach of meta-analysis to incorporate prior data into the current analysis. Meta-analysis is commonly used in genetic association studies. Our approach, however, has advantages over the meta-analysis. One major concern of the meta-analysis is heterogeneity of the studies it combines. This includes the studies with substantially different sample sizes and populations which are not comparable across the studies. As in the real application that we have shown, if the actual small p-value of the previous study were reported, it would dominate the outcome of a meta-analysis combining the previous and current studies. Unlike meta-analysis, by treating the prior study as historical data, our approach can adjust the heterogeneity of the two studies [2, 6]. Since our analysis is based on the log-OR, whose HE asymptotically follows a normal distribution, and we use a normal prior for the log-OR, it is not surprising that our results are very close to those from the meta-analysis when there is no heterogeneity across the studies.

APPENDIX

Proof of Proposition 1

Let $q(X,G)$ be the density-mass function for (X,G). The full likelihood for (Y,X,G) is $f(Y,X,G|\theta) = f(Y|X, G,\theta)q(X,G)$, and the Fisher information $I(\theta)$ under $f(Y,X,G|\theta)$ is the same as that under $f(Y|X,G,\theta)$. Denote $l(Y^n|X^n,G^n,\theta) = \log f(Y^n|X^n,G^n,\theta)$ and $l(Y^n,X^n, G^n|\theta) = \log f(Y^n,X^n,G^n|\theta)$. Let $l^{(k)}(Y^n|X^n,G^n,\theta)$ and $l^{(k)}(Y^n,X^n,G^n|\theta)$ be the corresponding partial derivatives with respect to θ ($k = 1,2$). Then $l^{(k)}(Y^n|X^n,G^n,\theta) \equiv l^{(k)}(Y^n,X^n,G^n|\theta)$, and

$$\begin{aligned}\tilde{\theta} = {} & \arg\sup\inf\nolimits_{\alpha\in\Lambda,d\in\Gamma} \\ & \int_\Gamma w(d,\beta)f(Y^n|X^n,G^n,\alpha,\beta)\pi(\beta)d\beta \\ = {} & \arg\sup\inf\nolimits_{\alpha\in\Lambda,d\in\Gamma} \\ & \int_\Gamma w(d,\beta)f(Y^n,X^n,G^n|\alpha,\beta)\pi(\beta)d\beta,\end{aligned}$$

i.e., the estimator $\tilde{\theta}$ obtained under the conditional likelihood is the same as that obtained under the full model. Thus conclusion (i) follows from Theorem 2.1 in [16]. Also, by Theorem 2.4 in [16], $\sqrt{n}(\hat{\alpha}^T - \alpha_0^T, \check{\beta} - \beta_0) = \Delta_1(\theta_0) + o_P(1)$, where $\Delta_1(\theta_0) = n^{-1/2}l^{(1)}(Y^n,X^n,G^n|\theta_0)$ is the scaled score function for the full model, thus (ii) follows.

Proof of Proposition 2

Let $l^{*(k)}(\theta)$ and $l^{(k)}(\theta)$ ($k = 1,2$), and $f^{(1)}(Y^n|X^n,G^n,\theta)$ be partial derivatives with respect to θ. Denote $h(\beta) = \log\pi(\beta)$ and define $h^{(k)}(\beta)$ ($k = 1,2$) similarly.

(i) By the definition of $\tilde{\theta}$, $l^{*(1)}(\tilde{\theta}) = 0$, so we have $-l^{*(1)}(\theta_0) = l^{*(1)}(\tilde{\theta}) - l^{*(1)}(\theta_0) = l^{*(2)}(\dot{\theta})(\tilde{\theta} - \theta_0)$, where $\dot{\theta} = (\dot{\alpha}^T, \dot{\beta})$ is between $\tilde{\theta}$ and θ_0. By the given condition, $\dot{\theta} \in A$, and so for large n, $-\frac{1}{n}l^{(2)}(\dot{\theta}) \approx I(\dot{\theta})$ (a.s.), which is non-singular by the assumption. The conditions $0 < \pi(\cdot) < \infty$ and $\pi^{(2)}(\cdot)$ being bounded and away from 0 on A imply $h^{(2)}(\cdot)$ is bounded on A. Thus

$$\begin{aligned}-\frac{1}{n}l^{*(2)}(\dot{\theta}) &= -\frac{1}{n}l^{(2)}(\dot{\theta}) - \frac{1}{n}h^{(2)}(\dot{\beta}) \\ &= -\frac{1}{n}l^{(2)}(\dot{\theta}) - O(n^{-1}) \\ &= I(\dot{\theta}) + o(1), \quad \text{a.s.}\end{aligned}$$

i.e., $-\frac{1}{n}l^{*(2)}(\dot{\theta}) \approx I(\dot{\theta})$ is non-singular (a.s.) for all large n. Thus, for large n, a.s.

$$\tilde{\theta} - \theta_0 = \left\{-\frac{1}{n}l^{*(2)}(\dot{\theta})\right\}^{-1}\left\{\frac{1}{n}l^{*(1)}(\theta_0)\right\}.$$

Similarly, the given conditions imply that $h^{(1)}(\beta_0)$ is finite, so

$$\frac{1}{n}l^{*(1)}(\theta_0) = \frac{1}{n}l^{(1)}(\theta_0) + \frac{1}{n}h^{(1)}(\beta_0) = \frac{1}{n}l^{(1)}(\theta_0) + O(n^{-1}) \quad \text{a.s.}$$

Since $f^{(1)}(Y^n|X^n,G^n,\theta) = f^{(1)}(Y^n,X^n,G^n|\theta)$ and

$$\begin{aligned}\frac{1}{n}l^{(1)}(\theta_0) &\overset{a.s.}{\to} E\left(\frac{f^{(1)}(Y^n,X^n,G^n|\theta_0)}{f(Y^n,X^n,G^n|\theta_0)}\right) \\ &= \frac{\partial}{\partial\theta_0}\sum_G \int f(y,x,G|\theta_0)dydx = 0,\end{aligned}$$

we have

$$\tilde{\theta} - \theta_0 = \left\{-\frac{1}{n}l^{*(2)}(\tilde{\theta})\right\}^{-1}\left\{\frac{1}{n}l^{(1)}(\theta_0) + O(n^{-1})\right\} \to 0, \quad \text{a.s.}$$

or $\tilde{\theta} \to \theta_0$ (a.s.).

(ii) From (i), we have $\dot{\theta} \to \theta_0$ (a.s.). Since $I(\cdot)$ is continuous at θ_0, $-\frac{1}{n}l^{*(2)}(\dot{\theta}) = -\frac{1}{n}l^{(2)}(\dot{\theta}) - \frac{1}{n}h^{(2)}(\dot{\beta}) \to I(\theta_0)$ (a.s.). Since $n^{-1/2}l^{(1)}(\theta_0) \overset{D}{\to} N(0, I(\theta_0))$, we have

$$\begin{aligned}&\sqrt{n}(\hat{\alpha}^T - \alpha_0^T, \check{\beta} - \beta_0) \\ &\quad = \left\{-n^{-1}l^{*(2)}(\dot{\theta})\right\}^{-1}\left\{n^{-1/2}l^{*(1)}(\theta_0)\right\} \\ &\quad = \left\{-n^{-1}l^{*(2)}(\dot{\theta})\right\}^{-1}\left\{n^{-1/2}l^{(1)}(\theta_0) + O_P(n^{-1/2})\right\} \\ &\quad \overset{D}{\to} N(0, I^{-1}(\theta_0))\end{aligned}$$

by Slutzky's theorem.

Estimation of τ_m^2 without genotype counts

Let $m_1 = r_{00} + r_{01} + r_{02}$ and $m_0 = s_{00} + s_{01} + s_{02}$ be the numbers of cases and controls in the past study, respectively. So $m = m_0 + m_1$. Note that $(r_{00}, r_{01}, r_{02}) \sim Mul(m_1; p_{00}, p_{01}, p_{02})$ and $(s_{00}, s_{01}, s_{02}) \sim$

$Mul(m_0; q_{00}, q_{01}, q_{02})$. Let p_1 and p_0 be the frequency of the minor (or risk) allele in the cases and controls, respectively. Assume HWE holds in the data.

Under the ADD model, we approximate the log-OR and its asymptotic variance by an allelic inference. Then an estimate of the variance for the log-OR is

$$\begin{aligned}
\tau_m^2 &\approx 4\{(2r_{02}+r_{01})^{-1}+(2r_{00}+r_{01})^{-1} \\
&\quad +(2s_{02}+s_{01})^{-1}+(2s_{00}+s_{01})^{-1}\} \\
&\approx \frac{4}{m_0}\left\{(2p_{02}+p_{01})^{-1}+(2p_{00}+p_{01})^{-1}\right\} \\
&\quad +\frac{4}{m_1}\left\{(2q_{02}+q_{01})^{-1}+(2q_{00}+q_{01})^{-1}\right\} \\
&\approx \frac{8}{m}[(2p_1)^{-1}+\{2(1-p_1)\}^{-1} \\
&\quad +(2p_0)^{-1}+\{2(1-p_0)\}^{-1}] \\
&= \frac{4}{m}\left[\{p_1(1-p1)\}^{-1}+\{p_0(1-p_0)\}^{-1}\right],
\end{aligned}$$

where m_0 and m_1 are replaced by $m/2$, where $m = M/H$ in the meta-analysis example.

The REC and DOM models are symmetric. We only show for the REC model, under which

$$\begin{aligned}
\tau_m^2 &\approx (r_{00}+r_{01})^{-1}+r_{02}^{-1}+(s_{00}+s_{01})^{-1}+s_{02}^{-1} \\
&\approx \frac{1}{m_1}\left\{(p_{00}+p_{01})^{-1}+p_{02}^{-1}\right\} \\
&\quad +\frac{1}{m_0}\left\{(q_{00}+q_{01})^{-1}+q_{02}^{-1}\right\} \\
&\approx \frac{2}{m}\left[\{p_0^2(1-p_0^2)\}^{-1}+\{p_1^2(1-p_1^2)\}^{-1}\right].
\end{aligned}$$

Received 27 November 2012

REFERENCES

[1] Bayarri, M. J. and Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Stat. Sci.* **19** 58–80. MR2082147

[2] Bhattacharya, B. (2009). Optimal use of historical information. *J. Stat. Plann. Inf.* **139** 4051–4063. MR2558349

[3] Darnell, G., Duong, D., Han, B., and Eskin, E. (2012). Incorporating prior information into association studies. *Bioinformatics* **28** 147–153.

[4] Fridley, B. L. (2010). Bayesian mixture models for the incorporation of prior knowledge to inform genetic association studies. *Genet. Epdemiol.* **34** 418–426.

[5] Good, I. J. (1992). The Bayes/non-Bayes compromise: A brief review. *J. Am. Stat. Assoc.* **87** 597–606. MR1185188

[6] Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Stat. Sci.* **15** 46–60. MR1842236

[7] Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., Sangiovanni, J. P., Mane, S. M., Mayne, S. T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308** 385–389.

[8] LeCam, L. M. and Yang, G. (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer, New York. MR1066869

[9] Prakasa Rao, B. L. S. (1987). *Asymptotic Theory of Statistical Inference*. Wiley, New York. MR0874342

[10] Prentice, R. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–411. MR0556730

[11] Qin, J. and Liang, K.-Y. (2011). Hypothesis testing in a mixture case-control model. *Biometrics* **67** 182–193. MR2898830

[12] Qin, J. and Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **84** 609–618. MR1603924

[13] Seaman, S. and Richardson, S. (2004). Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika* **91** 15–25. MR2050457

[14] Swaroop, A., Branham, K. E., Chen, W., and Abecasis, G. (2007). Genetic susceptibility to age-related macular degeneration: A paradigm for dissecting complex disease traits. *Hum. Molecular Genet.* **16** 174–182.

[15] The Wellcome Trust Case-Control Consortium (WTCCC) (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447** 661–678.

[16] Yuan, A. (2009). Bayesian frequentist hybrid inference. *Ann. Stat.* **37** 2458–2501. MR2543699

[17] Zheng, G., Yang, Y., Zhu, X., and Elston, R. C. (2012). *Analysis of Genetic Association Studies*. Springer, New York. MR2895171

Ao Yuan
Statistical Genetics and Bioinformatics Unit
National Human Genome Center, Howard University
Washington DC 20059
USA
E-mail address: ayuan@howard.edu

Qizhai Li
Academy of Mathematics and Systems Science
Academy of Sciences, Beijing
China
E-mail address: liqz@amss.ac.cn

Jing Qin
Biostatistics Research Branch
National Institute of Allergy and Infectious Diseases
Bethesda, MD 20892
USA
E-mail address: jingqin@niaid.nih.gov

Gang Zheng
Office of Biostatistics Research
National Heart, Lung and Blood Institute
Bethesda, MD 20892
USA
E-mail address: zhengg@nhlbi.nih.gov