A flexible semiparametric odds ratio model has been proposed to unify and to extend both the log-linear model and the joint normal model for data with a mix of discrete and continuous variables. The semiparametric odds ratio model is particularly useful for analyzing biased sampling designs. However, statistical inference of the model has not been systematically studied when more than one nonparametric component is involved in the model. In this article, we study the maximum semiparametric likelihood approach to estimation and inference of the semiparametric odds ratio model. We show that the maximum semiparametric likelihood estimator of the odds ratio parameter is consistent and asymptotically normally distributed. We also establish statistical inference under a misspecified semiparametric odds ratio model, which is important when handling weak identifiability in conditionally specified models under biased sampling designs. We use simulation studies to demonstrate that the proposed approaches have satisfactory finite sample performance. Finally, we illustrate the proposed approach by analyzing multiple traits in a genome-wide association study of high-density lipid protein. Supplementary materials for this article are available online.

KEY WORDS: Approximate odds ratio function; Extreme-value sampling design; Misspecified model; Semiparametric likelihood; Weak identifiability.

1. INTRODUCTION

The log-linear model (Bishop, Fienberg, and Holland 1975) and the joint normal model are the two most frequently used modeling approaches to discrete and continuous data, respectively. For data with a mix of discrete and continuous variables, a conditional normal model combined with a log-linear model is often used for the data analysis. Such approaches, however, are not flexible in model specification, nor robust against model misspecification. For example, it cannot accommodate possible interaction between two continuous variables. In addition, the normal modeling approach cannot address the problem of asymmetry and/or the long tail of the distribution of continuous variables (Azzalini and Valle 1996). Two versions of semiparametric odds ratio model have been proposed, respectively, by Chen (2004, 2007, 2010) and Osius (2004, 2009) to extend both the log-linear model and the joint normal model to address these issues in a unified way. Chen parameterized the model by odds ratio functions and conditional distributions at fixed conditions, while Osius parameterized the model by an odds ratio function and marginal distributions. The parameterization adopted by Chen (2004, 2007, 2010) has a closed-form expression for the joint density and is easier to study and generalize.

The semiparametric odds ratio model is particularly useful for handling biased sampling designs (Chen 2003, 2007, 2011; Osius 2009), such as case–control design, matched case–control design, and extreme-value sampling design widely used in epidemiological studies. Traditionally, the prospective logistic regression approach (Anderson 1972; Breslow, 1976, 1996; Prentice and Pyke 1979; Breslow and Day 1980) is employed for the case–control design. This approach is simple and efficient, and was also extended to various settings where multiplicative models were assumed (Rabinowitz 1997; Scott and Wild 1997; Chen 2003). When a nonmultiplicative model, such as the probit model, is fit to the case–control data, the issue of weak identifiability (Chen 2011) emerges. Such weak identifiability occurs more frequently in the analysis of case–control genetic association studies where complex structures, such as gene–environment independence and Hardy–Weinberg equilibrium in the general population, are incorporated into the model for efficiency gain (Chatterjee and Carroll 2005; Lin and Zeng 2006, 2009; Chen et al. 2008, 2009). As Chen (2011) demonstrated, weakly identifiable parameters can be very hard to estimate, especially when the true value of the parameter is close to the points where loss of identifiability occurs, because the observed data contain little information on the parameter. Testing hypotheses at those points, also known as testing under loss of identifiability (Lindsay 1995), was studied by Dacunha-Castelle and Gassiat (1999), Liu and Shao (2003), Zhu and Zhang (2006), and Song, Kosorok, and Fine (2009) among others. The distribution of such tests under the null hypothesis can be very difficult to characterize. To circumvent the problem, supplemental information in terms of the case–prevalence rate in the general population is sometimes assumed (Chatterjee and Carroll 2005; Lin and Zeng 2009). In practice, such information may not be available or may only be known loosely in a plausible range. Without supplemental information, existing retrospective likelihood approaches often perform poorly (Chen 2011; Chen and Chen 2011). Furthermore, when the outcome is continuous and the sampling design is the extreme-value design that selects only those subjects with very high and very low quantitative outcome measures (Feldt 1961;
Risch and Zhang 1995, 1996; Zhang and Risch 1996; Chen and Li 2011), it is unclear what supplemental information is needed to achieve strong identifiability and computation stability.

Osus (2009) studied the asymptotic properties of the likelihood estimator of his version of the semiparametric odds ratio model for the case where outcome-dependent samples are taken on a fixed number of outcome values. A systematic theoretical study of the likelihood approach to semiparametric odds ratio model has not been done, especially when more than one nonparametric component are involved in the model. In this article, we develop a maximum semiparametric likelihood approach to estimation and inference of the semiparametric odds ratio model (Chen 2004, 2007, 2010, 2011) aiming at addressing the issue of weakly identifiability in applications. Our study reveals that when all the parameters are strongly identifiable, such as in the generalization of log-linear model for discrete data and joint normal model for continuous data, the maximum likelihood estimator of parameters in the semiparametric odds ratio model is consistent and asymptotically normally distributed. When weakly identifiable parameters are involved in the odds ratio function, which is often induced by nonmultiplicative model or conditionally specified models under biased sampling designs, directly maximizing the likelihood is computationally challenging and the theoretical results may not apply. An approximate semiparametric odds ratio model is proposed to circumvent the inherent computational problem associated with estimating the weakly identifiable parameters. In this case, statistical inference on the model parameters requires us to study the asymptotic properties of the maximum likelihood estimator under the misspecified semiparametric odds ratio model that has not been done previously.

The remainder of this article is organized as follows. In Section 2, the semiparametric odds ratio modeling framework is briefly reviewed. In Section 3, theories for the maximum likelihood estimation and inference on both correctly specified and misspecified semiparametric odds ratio model are, respectively, established. Section 4 formulates applications of the developed theory through approximate likelihood for problems with weakly identifiable parameters frequently encountered in genetic association studies involving conditionally specified models under biased sampling designs. Section 5 deals with a specific application of the proposed approach to the genome-wide association study of lipoproteins, which involves multiple traits under extreme-value sampling design. Extensive simulation studies are also conducted in this section to assess the performance of the proposed approach. Section 6 concludes the article with a discussion of the proposed approach and further research. Proofs of the theoretical results are provided in the online supplementary appendix.

2. SEMIPARAMETRIC ODDS RATIO MODEL

Let $Y$ denote the data from a sampling unit. Suppose that variables in $Y$ are divided into $t$ groups as $Y_1, \ldots, Y_t$, where $Y_j$ taking values in $\mathcal{Y}_j$ is a vector of $d_j \geq 1$ dimension for $j = 1, \ldots, t$. Let $p(y_1, \ldots, y_t)$ be the density of $(Y_1, \ldots, Y_t)$ under an appropriate product of count measures and Lebesgue measures. Assume the positivity condition (Besag 1974, 1996) holds, that is, $p(y_1, \ldots, y_t) > 0$ if $p_j(y_j) > 0$ for $j = 1, \ldots, t$, where $p_j(y_j)$ denotes the marginal density of $Y_j$. Chen (2007, 2010) showed that, on one hand, any given joint density can be represented by odds ratio functions and conditional densities at a fixed condition. On the other hand, the semiparametric model that models the odds ratio functions parametrically and the conditional densities at a fixed condition nonparametrically is very flexible and unifies many existing parametric and semiparametric models. To keep this article self-contained, we give a brief review of these results here. Without loss of generality, we mainly focus on $t = 3$ in the remainder of this article.

Let $(y_{10}, y_{20}, y_{30})$ be a fixed reference point in the sample space. The conditional odds ratio function of $y_1$ versus $y_2$ given $y_3$ with the reference point $(y_{10}, y_{20})$ is defined as

$$
\eta(y_1, y_{10}; y_2, y_{20} \mid y_3) = \frac{p(y_1, y_2 \mid y_3)p(y_{10}, y_{20} \mid y_3)}{p(y_1, y_2 \mid y_3)p(y_{10}, y_{20} \mid y_3)}.
$$

Note also that the conditional odds ratio function can be equivalently defined as

$$
\eta(y_1, y_{10}; y_2, y_{20} \mid y_3) = \frac{p(y_1 \mid y_2, y_{20})p(y_{10} \mid y_{20})}{p(y_1 \mid y_2, y_{20})p(y_{10} \mid y_{20})}.
$$

Other conditional odds ratio functions, such as $y_1$ versus $y_2$ given $y_3$ with the reference point $(y_{10}, y_{30})$, can be defined similarly. We may interpret the conditional odds ratio function similarly to the conditional odds ratio parameter in a logistic regression. That is, given $Y_3 = y_3$ and $\Delta \geq 0$, the odds of $Y_1$ falls in the interval $[y_1 - \Delta, y_1 + \Delta]$ versus in the interval $[y_{10} - \Delta, y_{10} + \Delta]$ when $Y_2 = y_2$ is approximately

$$
\frac{P(Y_1 \in [y_1 - \Delta, y_1 + \Delta] \mid y_2, y_3)}{P(Y_1 \in [y_{10} - \Delta, y_{10} + \Delta] \mid y_2, y_3)} \approx \frac{p(y_1 \mid y_2, y_3)}{p(y_{10} \mid y_2, y_3)}.
$$

The conditional odds ratio function

$$
\eta(y_1, y_{10}; y_2, y_{20} \mid y_3) \approx \frac{P(Y_1 \in [y_1 - \Delta, y_1 + \Delta] \mid y_2, y_3)}{P(Y_1 \in [y_{10} - \Delta, y_{10} + \Delta] \mid y_2, y_3)} \frac{P(Y_1 \in [y_{10} - \Delta, y_{10} + \Delta] \mid y_2, y_3)}{P(Y_1 \in [y_1 - \Delta, y_1 + \Delta] \mid y_2, y_3)},
$$

is thus approximately the ratio of the odds of $Y_1$ falls in the interval $[y_1 - \Delta, y_1 + \Delta]$ versus in the interval $[y_{10} - \Delta, y_{10} + \Delta]$ when $Y_2 = y_2$ to the odds of $Y_1$ falls in the interval $[y_1 - \Delta, y_1 + \Delta]$ versus in the interval $[y_{10} - \Delta, y_{10} + \Delta]$ when $Y_2 = y_{20}$ for a given $Y_3 = y_3$. Note that the qualification “approximately” can be removed when $Y_1$ takes finite values and $\Delta$ is sufficiently small. For binary $Y_1$, this interpretation reduces to the commonly used interpretation for the conditional odds ratio parameter in a logistic regression model when $y_2 = y_{20} + 1$.

The unconditional odds ratio function for $y_1$ versus $(y_2, y_3)$ with the reference point $(y_{10}, y_{20}, y_{30})$ is defined as

$$
\eta(y_1, y_{10}; y_2, y_{20}, y_{30}) = \frac{p(y_1, y_2, y_3)p(y_{10}, y_{20}, y_{30})}{p(y_1, y_2, y_3)p(y_{10}, y_{20}, y_{30})},
$$

or equivalently as

$$
\eta(y_1, y_{10}; y_2, y_{20}, y_{30}) = \frac{p(y_1 \mid y_2, y_3)p(y_{10} \mid y_{20}, y_{30})}{p(y_1 \mid y_2, y_3)p(y_{10} \mid y_{20}, y_{30})}.
$$

See Chen (2010) for more details. Other unconditional odds ratio functions can be defined similarly. For brevity, we suppress the reference point in the odds ratio expression from now on. For example, use $\eta(y_1; y_2 \mid y_3)$ for $\eta(y_1, y_{10}; y_2, y_{20} \mid y_3)$. The
following easily verified identity,
\[ \eta \{ y_1; (y_2, y_3) \} = \eta(y_1; y_2 | y_3) \eta(y_1; y_3 | y_2), \]  
(1)

connects an unconditional odds ratio function to conditional odds ratio functions.

Chen (2010) gave an odds ratio representation of the joint density with arbitrary \( t \) groups as
\[
p(y_1, \ldots, y_i) \frac{\prod_{j=1}^{t-1} \eta_j(y_j; (y_{j+1}, \ldots, y_i)) \{ (y_{10}, \ldots, y_{(i-1)0}) \}}{\prod_{j=1}^{t} p_j(y_j | y_{-j0})} = \int \cdots \int \frac{\prod_{j=1}^{t-1} \eta_j(y_j; (y_{j+1}, \ldots, y_i)) \{ (y_{10}, \ldots, y_{(i-1)0}) \}}{\prod_{j=1}^{t} p_j(y_j | y_{-j0}) d \mu_j(y_j)},
\]
where \( y_{-j0} = (y_{10}, \ldots, y_{(j-1)0}, y_{(j+1)0}, \ldots, y_{100}) \) and \( y_0 = (y_{10}, \ldots, y_{10}) \) is a reference point, and \( \mu_j \) denotes Lebesgue measure or the counting measure or a product of both of \( y_j \) for \( j = 1, \ldots, t \). For \( t = 3 \), this representation can be specified as
\[
p(y_1, y_2, y_3) = \frac{\eta_1(y_1; (y_2, y_3)) \eta_2(y_2; (y_3) | y_{10}) p_1^2(y_1) p_2^2(y_2) p_3^2(y_3)}{\int \int \eta_1(y_1; (y_2, y_3)) \eta_2(y_2; (y_3) | y_{10}) p_1^2(y_1) p_2^2(y_2) p_3^2(y_3) d \mu_1(y_1) d \mu_2(y_2) d \mu_3(y_3)}, \tag{2}
\]
where \( p_1^2(y_1) = p_1(y_1 | y_{20}, y_{30}), p_2^2(y_2) = p_2(y_2 | y_{10}, y_{30}), p_3^2(y_3) = p_3(y_3 | y_{10}, y_{20}). \)

As illustrative examples, we show how to represent log-linear model and joint normal model in odds ratio form. For an \( I \times J \times K \) table, let the cell probability be \( \pi_{ijk} = P(Y_1 = i, Y_2 = j, Y_3 = k) \), where \( i = 1, \ldots, I \), \( j = 1, \ldots, J \), and \( k = 1, \ldots, K \). Suppose that the log-linear model has the form
\[
\log \pi_{ijk} = \lambda + \lambda_{1i} + \lambda_{j} + \lambda_{k} + \lambda_{ij} + \lambda_{ik} + \lambda_{jk} + \lambda_{ijk},
\]
where \( \sum_{i=1}^{I} \lambda_{1i} = 0, \sum_{j=1}^{J} \lambda_j = 0, \sum_{k=1}^{K} \lambda_k = 0, \sum_{i=1}^{I} \lambda_{ij} = 0, \sum_{j=1}^{J} \lambda_{ij} = 0, \sum_{i=1}^{I} \lambda_{ik} = 0, \sum_{k=1}^{K} \lambda_{ik} = 0, \sum_{j=1}^{J} \lambda_{jk} = 0, \sum_{k=1}^{K} \lambda_{jk} = 0. \)

Let \( (Y_{10}, Y_{20}, Y_{30}) = (1, 1, 1) \) be the reference category. The odds ratio function
\[
\log \eta_1(Y_1 = i; (Y_2 = j, Y_3 = k)) = \lambda_{ij} + \lambda_{ik} - \lambda_{1i} - \lambda_{j} - \lambda_{k} - \lambda_{ij} - \lambda_{ik},
\]
\[
\log \eta_2(Y_2 = j; (Y_1 = 1, Y_3 = k)) = \lambda_{ij} + \lambda_{ik} - \lambda_{ij} - \lambda_{ik} - \lambda_{jk},
\]
\[
\log \eta_3(Y_3 = k; (Y_1 = 1, Y_2 = 1)) = \lambda_{ij} + \lambda_{ik} - \lambda_{1i} - \lambda_{j} - \lambda_{k} - \lambda_{ij} + \lambda_{jk}.
\]
The conditional probability mass functions at the fixed condition are
\[
p(Y_1 = i | Y_2 = 1, Y_3 = 1) = \frac{\exp(\lambda_{ij} + \lambda_{ik} + \lambda_{jk})}{\sum_{j=1}^{J} \sum_{k=1}^{K} \exp(\lambda_{ij} + \lambda_{ik} + \lambda_{jk})},
\]
\[
p(Y_2 = j | Y_1 = 1, Y_3 = 1) = \frac{\exp(\lambda_{ij} + \lambda_{ik} + \lambda_{jk})}{\sum_{j=1}^{J} \sum_{k=1}^{K} \exp(\lambda_{ij} + \lambda_{ik} + \lambda_{jk})},
\]
\[
p(Y_3 = k | Y_2 = 1, Y_1 = 1) = \frac{\exp(\lambda_{ij} + \lambda_{ik} + \lambda_{jk})}{\sum_{k=1}^{K} \sum_{j=1}^{J} \exp(\lambda_{ij} + \lambda_{ik} + \lambda_{jk})}.
\]
which are, respectively, denoted by \( p_1^1, p_2^1, p_3^1 \) in the following. The probability \( \pi_{ijk} \) can be represented as
\[
\pi_{ijk} = \frac{\eta_1(Y_1 = i; (Y_2 = j, Y_3 = k)) \eta_2(Y_2 = j; Y_3 = k | Y_1 = 1) p_1^1 p_2^1 p_3^1}{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \eta_1(Y_1 = i; (Y_2 = j, Y_3 = k)) \eta_2(Y_2 = j; Y_3 = k | Y_1 = 1) p_1^1 p_2^1 p_3^1} \tag{3}
\]
For a trivariate joint normal density \( (Y_1, Y_2, Y_3) \sim N(\mu, \Sigma) \), the joint density can be represented as
\[
p(y_1, y_2, y_3) = \frac{\eta_1(y_1; (y_2, y_3)) \eta_2(y_2; y_3 | y_{10}) \prod_{j=1}^{3} p_j(y_j)}{\int \int \eta_1(y_1; (y_2, y_3)) \eta_2(y_2; y_3 | y_{10}) \prod_{j=1}^{3} p_j(y_j) dy_j}, \tag{4}
\]
where
\[
\log \eta_1(y_1; (y_2, y_3)) = -\sigma_1^2(y_1 - y_{10}) (y_2 - y_{20}) - \sigma_{13} (y_1 - y_{100} (y_3 - y_{300}),
\]
\[
\log \eta_2(y_2; y_3 | y_{10}) = -\sigma_2^2(y_2 - y_{20}) (y_3 - y_{30}),
\]
\[
p_j(y_j) = \sqrt{\frac{\sigma j j}{2 \pi}} \exp \left[ -\frac{\sigma j j}{2} \left( y_j - \mu_j + \sum_{k \neq j} (y_{kj} - \mu_k) \frac{\sigma_{jk}}{\sigma j j} \right)^2 \right],
\]
and \( \mu = (\mu_1, \mu_2, \mu_3) \) and \( \Sigma^{-1} = (\sigma_{jk}) \). In general, the representation applies equally well to densities of a mix of discrete and continuous variables.

Two important features of the odds ratio representation make it useful in practice. The first feature is that the representation separates the association part, that is, odds ratio functions, from the marginal part, that is, conditional densities at a fixed condition, in a joint density. In an unconstrained joint density, the two parts are variation independent (Chen 2007). In many frequently used parametric models, parameters in the two parts can be naturally separated after some reparameterization. For example, both representations (3) and (4) achieve the parameter separation after reparameterization. Parameter separation in the odds ratio representation of the log-linear model is easy to see. In the representation of the joint normal model, the association parameters \( -\sigma_{jk}, j \neq k \) are separated from the marginal parameters \( \sigma j j, \mu j - \sum_{k \neq j} (y_{kj} - \mu_k) \frac{\sigma_{jk}}{\sigma j j} \). The second feature of the odds ratio representation is its flexibility in handling biased sampling designs. Let \( S = 1 \) denote a subject being included in the sample. Consider a general biased sampling design having sampling probability
\[
P(S = 1 | Y_1, Y_2, Y_3) \propto \pi_1(Y_1) \pi_2(Y_2) \pi_3(Y_3), \tag{5}
\]
with known or unknown \( \pi_j(Y_j), j = 1, 2, 3 \). This sampling design includes both the case–control design and the extreme-value sampling design as special cases. Under this sampling design,
\[
p(y_1, y_2, y_3 | S = 1) = \frac{\eta_1(y_1; (y_2, y_3)) \eta_2(y_2; y_3 | y_{10})}{\prod_{j=1}^{3} q_j(y_j)} \frac{\prod_{j=1}^{3} p_j(y_j)}{\prod_{j=1}^{3} q_j(y_j)} \tag{6}
\]
where
\[
q_j(y_j) = \frac{\pi_j(y_j)p_j(y_j \mid y_{-j0})}{\int \pi_j(y_j)p_j(y_j \mid y_{-j0})d\mu_j(y_j)}.
\]

It can be seen from expression (6) that the odds ratio functions are invariant under sampling design (5). When \(\pi_j(y_j)\) is unknown, such as in a case-control design or an extreme-value sampling design, \(q_j(y_j)\) has a nonparametric form even if \(p_j(y_j \mid y_{-j0})\) is of a parametric form. This motivates us to consider the semiparametric odds ratio model. Which models the odds ratio functions parametrically and \(q_j\) or the original \(p_j(y_j \mid y_{-j0})\) nonparametrically.

Let \(\eta_1(y_1; y_2, y_3) = \eta_1(y_1; y_2, y_3) = 1\) for all \(y_1, y_2, y_3\) in a bounded closed set in Euclidean spaces. Let \(P_{Y_k}\) be the collection of all distribution functions absolutely continuous with respect to \(\mu_k\) on \(\mathbb{Y}_k\) for \(k = 1, 2, 3\), Assume the following conditions hold.

1. The set \(B\) is convex compact in \(\mathbb{R}^q\) and \(\mathbb{Y}_0\) is an inner point of \(B\). Parameter \(\gamma \in B\) is identifiable from \(\eta(y_1; y_2, y_3; \gamma)\).

2. The set \(\mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{Y}_3\) is compact in Euclidean space. There exists \(c_1 > 0\) such that, for any \(y_k \in \mathbb{Y}_k, k = 1, 2, 3, \gamma \in B,\)

\[
0 < c_1 < \eta(y_1; y_2, y_3; |\gamma|) \leq c_1^{-1} < +\infty,
\]

and \(\eta(y_1; y_2, y_3; |\gamma|)\) has bounded continuous derivatives up to the third order with respect to \(\gamma\).

3. \(\eta(y_1; y_2, y_3; |\gamma|)\) and its first derivatives with respect to \(\gamma \in B\) are bounded variation functions with respect to \(y_k, k = 1, 2, 3, \gamma \in B\), and the total variations are uniformly bounded. That is,

\[
\sup \left\{|\eta|, |\eta|, |\eta|, |\eta|, |\eta|, |\eta|, |\eta|, |\eta|, \forall r, s, t \in \gamma \right\} < \infty,
\]

with the supremum being taken over \(y_k \in \mathbb{Y}_k, k = 1, 2, 3, \gamma \in B\).

4. The Kullback–Leibler information exists and is finite. That is,

\[
E_0 \left( \log \frac{p(y_1, y_2 \mid y_3, \gamma, Q_1, Q_2)}{p(y_1, y_2 \mid y_3, \gamma, Q_{10}, Q_{20})} \right) < +\infty,
\]

for all \(\gamma \in B, \ D_0 \in P_{Y_0}, k = 1, 2, 3, y_0, Q_{10}, Q_{20}\) is the true density function of \(Y_1, Y_2\) given \(Y_3\).

5. For almost all \(Y_3 \in \mathcal{Y}_3,\)

\[
\text{var}_{(y_0, Q_{10}, Q_{20})} \left( \frac{\partial \log \eta}{\partial \gamma} (Y_1; Y_2; Y_3 | y_0) \right) Y_3 > 0.
\]

The identifiability condition 1 is an important requirement, which may fail when weak identifiability occurs. Conditions 2 and 3 are technique regularity conditions for applying uniform law of large number and central limit theorem to obtain the asymptotic results. We did not explore the best possible
conditions, and as a result, these conditions may be relaxed. Condition 4 is required for consistency and condition 5 for $\sqrt{n}$-convergence rate of the parameter estimator.

**Theorem 1.** Under Conditions 1–5, the maximum likelihood estimator of $\gamma$ from maximizing (8) exists and is almost surely consistent. Furthermore, the maximum likelihood estimator of $\gamma$, denoted by $\hat{\gamma}_n$, is asymptotically normally distributed. That is,

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) \rightarrow N(0, V_0),$$

where the asymptotic variance satisfies $h_0^T V h_0 = (h_0, 0, 0) h_0^{-1} (h_0, 0, 0)$ with $\sigma$ defined in the Appendix and $h_0$ being a vector having the same dimension as $\gamma$. The variance function $\sigma$ can be consistently estimated by the sampling version of the inverse of the information matrix or by the inverse of the information matrix of the profile likelihood for $\gamma$ from (8).

### 3.2 Misspecified Semiparametric Odds Ratio Model

To account for possible misspecification involved in the odds ratio functions, we prove a generalized version of Theorem 1 under modified conditions. Let

$$(\gamma_0, Q_{10}, Q_{20}) = \arg \min_{\gamma, Q_{0}, Q_{2}} E_0 \left\{ -\log \frac{p(y_1, y_2 | y_3, \gamma, Q_1, Q_2)}{p_0(y_1, y_2 | y_3)} \right\},$$

(9)

where $p_0(y_1, y_2)$ is the true density generating the observed data. The modification keeps Conditions 1–3 and replaces Conditions 4 and 5 by the following conditions:

4a. The Kullback–Leibler information exists and is finite, that is,

$$E_0 \left\{ \left|\frac{\partial^2}{\partial \gamma^2} \log p(y_1, y_2 | y_3, \gamma, Q_1, Q_2) \right| \right\} < +\infty,$$

where $E_0$ is the expectation computed under $f_0$.

5a. The minimizer in (9) exists and is unique. Furthermore, let $H_0$ be a bounded set in $R^p$. Let $H_k$ be the collection of functions of $y_k$ with bounded variations on the support of $Q_{10}$ and $h_k(y_k)dQ_{00}(y_k) = 0$ for all $h_k \in H_k$, $k = 1, 2$. For any nonzero $(h_0, h_1, h_2) \in H_0 \times H_1 \times H_2$,

$$-h_0^T E_0 \left[ \frac{\partial^2}{\partial \gamma^2} \log \eta(Y_1; Y_2; Y_3 | \gamma_0) - E_{(\gamma_0, Q_{10}, Q_{20})} \right] + E_0 \left[ \frac{\partial^2}{\partial \gamma^2} \log \eta(Y_1; Y_2; Y_3 | \gamma_0) + h_1(Y_1) + h_2(Y_2) \right] > 0.$$

Condition 4a is the same as Condition 4 except that the true density $f_0(y_1, y_2, y_3)$ generated the data is different from the best possible model $f(y_1, y_2, y_3 | \gamma_0, Q_{10}, Q_{20})$ in the semiparametric odds ratio family. The second part of Condition 5a is a strengthened version of an inequality that can be derived from the first part of Condition 5a. The modified conditions ensure that the density minimizing the Kullback–Leibler divergence within the semiparametric odds ratio model exists, the minimizer is asymptotically unique, and the matrix of the second derivatives is invertible. Theorem 2 establishes that the estimator from maximizing the likelihood under misspecified model converges and is asymptotically normally distributed. Proof of Theorem 2 is given in the online supplementary appendix.

**Theorem 2.** Under Conditions 1–3, 4a, and 5a, the maximum likelihood estimator of $\gamma$ from maximizing (8) exists and is almost surely converges to $\gamma_0$, which is the value of $\gamma$ minimizing the Kullback–Leibler divergence. Furthermore, the maximum likelihood estimator of $\gamma$ is asymptotically normally distributed. That is,

$$\sqrt{n}(\hat{\gamma}_n - \gamma_0) \rightarrow N(0, V),$$

where the asymptotic variance satisfies $h_0^T V h_0 = P_0(y_0, Q_{10}, Q_{20})(Y_1, Y_2, Y_3)^{-1}(h_0, 0, 0)^2$ with $\sigma$ defined in the Appendix. $V$ is consistently estimated by the corresponding sampling version of the sandwich estimator.

Theorem 1 is a special case of Theorem 2. To see this, note that Conditions 4 and 5 imply Conditions 4a and 5a. It is easy to see that Condition 4 implies Condition 4a. Under Conditions 4 and 5,

$$E_0 \left\{ \log \frac{f_1(y_1, y_2 | y_3, \gamma, Q_1, Q_2)}{f_0(y_1, y_2 | y_3)} \right\} \leq 0$$

by Jensen’s inequality. The equality holds only when $f(y_1, y_2 | y_3, \gamma, Q_1, Q_2) = f_0(y_1, y_2 | y_3)$, which implies the minimizer in (9) exists, unique, and is the same as the true parameter values by the identifiability Condition 1. Note that the identifiability here refers to the misspecified semiparametric odds ratio model, not the odds ratio model corresponding to the family of distributions containing the truth, which may be weakly identifiable only. See the next section for more details. Furthermore, $f_0(y_1, y_2, y_3) = f(y_1, y_2, y_3 | \gamma_0, Q_{10}, Q_{20})$ implies that

$$E_0 \left[ \frac{\partial}{\partial \gamma} \log \eta(Y_1; Y_2; Y_3 | \gamma_0) \right] = E_{(\gamma_0, Q_{10}, Q_{20})} \left\{ \frac{\partial}{\partial \gamma} \log \eta(Y_1; Y_2; Y_3 | \gamma_0) \right\}.$$
Define the odds ratio function for $t_1(y_2, y_3)$ as
\[ \eta_1(y_2; y_3) = \frac{t_1(y_2, y_3)p_1(y_2, y_3)}{t_1(y_2, y_3)p_1(y_2, y_3)}. \]

By comparison to (2) and the uniqueness of the odds ratio representation (Chen 2010, 2011), it can be seen that $\eta_2(y_2; y_3) = \eta_1(y_2; y_3)\eta_3(y_2; y_3)$, and
\[ q_2(y_2) = \frac{t_1(y_2, y_3)p_2(y_2 | y_3)}{t_1(y_2, y_3)p_2(y_2 | y_3)}, \]
\[ q_3(y_3) = \frac{t_1(y_2, y_3)p_3(y_3) | y_3 = y_3)}{t_1(y_2, y_3)p_3(y_3) | y_3 = y_3}. \]

The density $q_3(y_3)$ is usually unspecified. In this case, the conditional likelihood corresponding to (8) becomes
\[ \prod_{i=1}^n \int \eta_{1_i}(y_{1_i}; y_{1_i} | y_1)q_{1_i}(y_1) | y_1)p_1(y_1 | y_20, y_30)d\gamma_1. \]

The identifiability result in Chen (2011) guarantees that $\eta_1$, $\eta_2$, and $q_3(y_j)$, $j = 1, 2, 3$, are identifiable from the biased sampling design (5). However, some parameters in $\eta_1$ and $\eta_2$ may be unidentifiable or only weakly identifiable from the biased sample. This can lead to poor performance of the estimator for the model parameters with practically attainable sample sizes when the parameters take values close to an unidentifiable point. We approach this problem using the following approximation.

4.2 Approximation to Odds Ratio Functions

For simplicity of discussion, we assume in the following that both $q_2^*$ and $q_3^*$ are unconstrained density functions so that both $q_2$ and $q_3$ are unconstrained. In this case, the primary problem lies in the estimation of parameters in $\eta_1$. The parameters in $\eta_1$ is determined by $\eta_1$ that is identifiable from the biased sampling design, and by $p_1(y_1 | y_20, y_30)$, which may or may not be identifiable from the biased sampling design. We propose an approximation to $\eta_1(y_2, y_3)$ to circumvent the problem.

Note that
\[ \log t_1(y_2, y_3) = -\log \int \eta_{1_i}(y_1; y_2, y_3)p_1(y_1 | y_20, y_30)d\gamma_1. \]

In many applications, $\eta_1$ can be expressed as
\[ \log \eta_{1_i}(y_1; y_2, y_3) = (y_1 - y_{10})m(y_2, y_3), \]
where $m(y_2, y_3)$ is a function satisfying $m(20, 30) = 0$. For example, if $p(y_1 | y_2, y_3)$ is a logistic regression with interaction, $m(y_2, y_3) = \theta_1(y_2 - y_20) + \theta_2(y_3 - y_30) + \theta_12(y_2 y_3 - y_20 y_30)$. This function can be rewritten as $m(y_2, y_3) = \theta_1^*(y_2 - y_20) + \theta_2^*(y_3 - y_30) + \theta_12(y_2 - y_20)(y_3 - y_30)$, where $\theta_1^* = \theta_1 + \theta_1 y_30$ and $\theta_2^* = \theta_2 + \theta_2 y_20$. Let
\[ \varphi(m) = \log \int e^{(y_1 - y_{10})m} p_1(y_1 | y_20, y_30)d\gamma_1, \]
which is the moment generating function of $Y_1^* - y_{10}$ with $Y_1^*$ follows $p_1(y_1 | y_20, y_30)$. By Taylor expansion,
\[ \varphi(m) = \varphi(0) + \varphi'(0)m + \varphi''(0)\frac{m^2}{2} + \varphi'''(0)\frac{m^3}{6} + \cdots, \]
where \( \varphi(0) = 0, \varphi'(0) = E(Y_i^* - y_{10}), \) and \( \varphi''(0) = \text{var}(Y_i^*) \), the third- and higher-order derivatives are respectively, the third- and higher-order cumulants of the distribution of \( Y_i^* \). It follows from applying the expansion to \( \eta_i(y_2, y_3) \) that
\[
- \log \eta_i(y_2; y_3) = \varphi(0)(m(y_2, y_3) - m(y_{20}, y_3) - m(y_2, y_{30})) + \frac{1}{2} \varphi'(0)m^2(y_2, y_3) - m^2(y_{20}, y_3)
+ \frac{1}{6} \varphi''(0)(m^3(y_2, y_3) - m_{20}(y_2, y_3))
- m^3(y_{20}, y_3) + \cdots. \tag{14}
\]

When \( Y_1 \) is a binary outcome and the case-prevalence rate is close to zero, then \( p_0 \approx 0 \) and thus \( \varphi(0) = \varphi(1)(0) \approx 0 \), which imply \( \log \eta_i(y_2; y_3) \approx 0 \). The approximation reduces to the log-linear model approximation under the sparse case assumption (Chen and Chen 2011). When \( Y_1 \) is continuous and \( \rho(1) \) is a normal density, \( \varphi(0)(0) \) is equal for all \( k \geq 3 \). The approximation up to the second term is thus exact. In general, the third-order (in \( y_2 \) and \( y_3 \)) approximation yields
\[
\log \eta_i(y_2; y_3) \approx -(\varphi(0)\theta_{12} + \varphi''(0)\theta_{12}^2)(y_2 - y_{20})(y_3 - y_{30})
- \left\{ \varphi''(0)\theta_{12}^2 + \frac{\varphi''(0)}{2}\theta_{12}^2 \right\}y_2(y_3 - y_{30})
- \left\{ \varphi''(0)\theta_{12}^2 + \frac{\varphi''(0)}{2}\theta_{12}^2 \right\}y_2(y_3 - y_{30}).
\]

When one of the approximate semiparametric odds ratio models is used in likelihood (12), the approximate likelihood for the odds ratio parameters can be relatively easily maximized because those parameters in the odds ratio functions are strongly identifiable. Theorem 2 can be applied to inference on the model parameters under a misspecified semiparametric odds ratio model.

5. APPLICATION TO MULTIPLE TRAITS ANALYSIS
5.1 Models and Approximation for Multiple Traits Analysis

The odds ratio approximation framework can be applied to many problems with biased sampling designs using consecutively conditionally specified models, for example, in efficiency gain by exploiting the gene–environment independence assumption (Chatterjee and Carroll 2005; Lin and Zeng 2006; Chen, Chatterjee, and Carroll 2008, 2009; Chen and Chen 2011), and in secondary trait analysis (Monsees et al. 2009; Lin and Zeng 2009; Chen, Reilly, and Li 2013) of a case–control sample. Details of these applications are discussed elsewhere (Chen, Kittles, and Zhang 2013; Chen, Reilly, and Li 2013). Here, we focus on an application to identifying genes that are associated with multiple phenotypes with data from a biased sampling design. One biological mechanism that favors the joint analysis of multiple phenotypes is the pleiotropy, a single gene that affects multiple phenotypes.

To avoid lengthy discussion in a general treatment, we concentrate on the case where the primary outcome \( Y_1 \) is a quantitative trait and the extreme-value sampling design is employed. Let \( Y_2 \) denote other phenotypes of interest and \( Y_3 = G \) denote the genotype. The extreme-value sampling design has sampling probability \( p(S = 1 | Y_1, Y_2, G) = \pi(Y_i) \), where \( S \) is the sampling indicator and the exact functional form of \( \pi \) is assumed unknown. Under this sampling design, a simple retrospective analysis based on \( p(G | Y_1, Y_2) \) yields estimates of the effects of genotype \( G \) on phenotypes \( Y_1 \) and \( Y_2 \). However, since the sampling design depends on \( Y_1 \) only, the simple retrospective likelihood approach is inefficient. More efficient analysis uses the retrospective likelihood based on
\[
p(Y_2, G | Y_1) = \frac{p(Y_1, Y_2 | G)p(G)}{\sum_G p(Y_1, Y_2 | G)p(G)dY_2}.
\]

One parametric approach to this problem models
\[
Y_1 = \beta_0 + \beta_1 G + \beta_2 Y_2 + \beta_12 Y_2 G + \epsilon,
\]
\[
Y_2 = \alpha_0 + \alpha_1 G + \epsilon,
\]
where \( \epsilon \sim N(0, \sigma^2) \) and \( \epsilon \sim N(0, \omega^2) \). Note that we include the interaction term in the first model to make it more flexible.

Under the normal model, it can be shown that both \( \beta_0 \) and \( \sigma^2 \) are identifiable if \( \beta_1 \neq 0 \) or \( \beta_2 \neq 0 \). However, when \( \beta_1 = \beta_2 = 0 \),
\[
p(Y_2, G | Y_1) = \frac{\exp(\beta_1/\sigma^2)Y_1 G)^p(G)^1}{\sum_G \exp(\beta_1/\sigma^2)Y_1 G)^p(G)^{\sqrt{2\pi} \omega}}
\]
\[
\times \exp \left\{ \frac{1}{2\omega^2}Y_2 - \alpha_0 - \alpha_1 G \right\}^2, \tag{15}
\]
which suggests that neither \( \beta_0 \) nor \( \sigma^2 \) is identifiable, where \( \omega^* \) is an unrestricted distribution for \( G \). This means that \( \beta_0 \) and \( \sigma^2 \) are weakly identifiable parameters.

The foregoing discussion indicates that the more restrictive normal model assumption may not help to remove the weakly identifiable parameters in the extreme-value sampling design. For robustness consideration and ease of presentation, we switch to the semiparametric odds ratio models in the following. Suppose that the parametric odds ratio function (13) holds. Under the normal model for \( Y_1, (\theta_1, \theta_2, \theta_12) = (\beta_1, \beta_2, \beta_12)/\sigma^2 \). Suppose that we model \( Y_2 \) by a semiparametric odds ratio model with
\[
\log \eta_2^*(G; Y_2; G) = \alpha(G - G_0)(Y_2 - Y_{20}). \tag{16}
\]

Under the semiparametric odds ratio models (10) and (11), the retrospective likelihood based on \( p(Y_2, G | Y_1) \) is equivalent to the prospective likelihood based on \( p(Y_1, Y_2 | G, S = 1) \), which has the form
\[
\frac{\eta(Y_1, G; Y_2)q_1(Y_1)q_2(G)}{\sum_G \eta(Y_1, G; Y_2)q_1(Y_1)q_2(G)dY_1},
\]
with \( q_1(Y_1) \) and \( q_2(G) \) unspecified, and
\[
\log \eta(Y_1, G; Y_2) = \theta_2^*(Y_1 - y_{10})(G - G_0) + \theta_2^*(Y_1 - y_{10})
\times (Y_2 - Y_{20})
+ \psi_2(Y_1 - y_{10})(G - G_0)(Y_2 - Y_{20}) + \psi_1
\times (G - G_0) \times (Y_2 - Y_{20})
+ \psi_2(G - G_0)^2(Y_2 - Y_{20}) + \psi_4(G - G_0)
\times (Y_2 - Y_{20})^2 + \psi_4(G - G_0)^2(Y_2 - Y_{20})^2.
\]

The odds ratio functions are approximation when the semiparametric odds ratio models for \( Y_1 \) and \( Y_2 \) are correctly specified.
If the normal model for $Y_1$ is correctly specified, the odds ratio functions are exact.

When genotype is individually evaluated, $Y_2$ is usually of higher dimension than $G$. To simplify the computation, we choose to work with the prospective likelihood in the following. Parameter $\gamma$ in Theorem 2 becomes $(\theta^*, \psi)$ where $\theta = (\theta_1^*, \theta_2^*, \theta_1)$ and $\psi = (\psi_1, \psi_2, \psi_3, \psi_4)$. From the expansion (14), the parameters are related through

$$
\psi_1 = \alpha - \mu \theta_{12} - \tau^2 \theta_1^* \theta_2^*,
$$

$$
\psi_2 = -\tau^2 \theta_1^* \theta_{12},
$$

$$
\psi_3 = -\tau^2 \theta_2^* \theta_{12},
$$

$$
\psi_4 = -\frac{\tau^2}{2} \theta_1^2,
$$

where $\mu = \beta_0 + \beta_1 G_0 + \beta_2 Y_{20} + \beta_12 G_0 Y_{20} - Y_{10}$. Applying Theorem 2 yields

$$
\sqrt{n}[(\hat{\theta}^*, \hat{\psi}) - (\theta^*, \psi_0)] \sim N(0, V),
$$

where $(\theta^*, \psi_0)$ is the convergent point.

When $\theta_{12} \neq 0$, $\theta_3 = (\theta^*, \psi_1, \tau^2)$ are identifiable where $\psi_1 = \alpha - \theta_{12} \mu$. The identifiable parameters can be estimated using the weighted least-square approach

$$
\min_{\theta^*, \psi_1, \tau^2} \{Z - g(\theta^*, \psi_1, \tau^2)\}^T V^{-1} \{Z - g(\theta^*, \psi_1, \tau^2)\},
$$

where $Z = (\hat{\theta}^*, \hat{\psi})$ and $g(\theta^*, \psi_1, \tau^2) = (\theta^*, \psi_1)$ and $V$ is the variance of the maximum likelihood estimator for $(\theta^*, \psi_1)$. The least-square estimator has asymptotic variance $(X' V^{-1} X)^{-1}$, where $X = (\partial \theta^*, \psi_1) / \partial \theta_2$ evaluated at the convergent point. The $\theta$ estimator from the least-square combination is theoretically more efficient than the simple prospective analysis based on $p(Y_1 | G, Y_2, S = 1)$ alone because $p(Y_1 | G, Y_2, S = 1)$ contains information for estimating $\theta$. When $\theta_{12} = 0$, parameters $\psi_k = 0$ for $k = 2, 3, 4$. In this case, $\psi_1 = \alpha - \tau^2 \theta_1^* \theta_2$ is identifiable. But neither $\alpha$ nor $\tau^2$ is identifiable if $\partial \theta_{12} \neq 0$. Parameter $\psi_1$ is the association of $Y_2$ and $G$ conditional on $Y_1$ rather than the marginal association of $Y_2$ and $G$, which is $\alpha$.

5.2 Simulation Study

In the following, we conduct a simulation study to assess the performance of the proposed approach in an extreme-value sampling design. We simulated a minor allele with frequency 0.3 in the population. The simulated genotype $G$ follows the Hardy–Weinberg equilibrium. That is, $P(G = 2) = p^2$, $P(G = 1) = 2p(1 - p)$, and $P(G = 0) = (1 - p)^2$. A quantitative trait $Y_2$ is simulated to follow a normal distribution with mean $\alpha_0 + \alpha_1 G$ and variance $\omega^2 = 1$. Another quantitative trait $Y_1$ is simulated conditional on $Y_2$ and $G$ to follow a normal model with mean $\beta_0 + \beta_1 G + \beta_2 Y_2 + \beta_12 G Y_2$ and unit variance. The extreme-value sampling design is employed to select subjects whose quantitative trait $Y_2$ is outside mean plus-minus 2 standard deviations. A range of parameter values for $\alpha_1$ and $\beta = (\beta_1, \beta_2, \beta_12)$ are used in the simulation. We set $\alpha_0$ and $\beta_0$ to zero in generating the data. In the analysis of the data, a linear odds ratio model is assumed for the $G$–$Y_2$ association. However, neither the sampling probability is assumed known, nor the Hardy–Weinberg equilibrium is assumed to hold in the analysis. Three methods of analysis are employed. The first fits a simple linear regression model with $G$–$Y_2$ interaction in the model.

Two estimators are obtained by this method: one is the regression parameter estimators (LRM) and the other is the estimator for the odds ratio parameter (LMO) defined as the regression coefficient over the residual variance. The second method fits a semiparametric odds ratio model (SOR) for $P(Y_1 | G, Y_2, S = 1)$ with the $G$–$Y_2$ interaction included. The third method fits a joint odds ratio model (JOR) for $P(Y_1, G | Y_2, S = 1)$ with the odds ratio function specified in the previous subsection. When the interaction is strong, two additional estimators are obtained: one is the weighted least-square combination (WLS) of estimated odds ratio parameters, and the other is the unweighted least-square combination (LS).

The simulation results based on 1000 repetitions of a sample size 400 are given in Table 1 for the case with no $G$–$Y_2$ interaction and in Tables 2 and 3 for the case with strong $G$–$Y_2$ interaction. When there is no $G$–$Y_2$ interaction, the least-square combination is infeasible even if we did not assume we know there is no $G$–$Y_2$ interaction. When more extreme-valued samples are taken, for example, those beyond 3 standard deviations of the mean, our simulation results not shown here also indicate more efficiency is gained. It can be seen from Table 2 that both LMR and LMO may be subject to very large bias and their variance estimates may also be very different from the empirical variances. These observations suggest that inference based on the linear model under the extreme-value sampling design can be misleading. In comparison, the odds ratio parameter estimator based on SOR or JOR has small bias and the variance estimates are close to the empirical variances. The efficiency of the odds ratio parameter estimator based on JOR is higher than

| $(\theta_1, \theta_2)$ | Method | Bias | Var | Mev | Bias | Var | Mev |
|---------------------|--------|------|-----|-----|------|-----|-----|
| (1, -1)             | LMR    | 0.78 | 0.0122 | 0.0096 | -0.63 | 0.0057 | 0.0032 |
|                     | LMO    | 0.26 | 0.0170 | 0.0128 | -0.15 | 0.0119 | 0.0083 |
|                     | SOR    | 0.07 | 0.0447 | 0.0475 | -0.06 | 0.0348 | 0.0326 |
|                     | JOR    | 0.04 | 0.0189 | 0.0177 | -0.05 | 0.0200 | 0.0195 |
| (0.0)               | LMR    | 0.00 | 0.0010 | 0.0010 | 0.00 | 0.0028 | 0.0021 |
|                     | LMO    | 0.00 | 0.0000 | 0.0000 | 0.00 | 0.0009 | 0.0008 |
|                     | SOR    | 0.00 | 0.0000 | 0.0000 | 0.00 | 0.0009 | 0.0009 |
|                     | JOR    | 0.00 | 0.0000 | 0.0000 | 0.00 | 0.0009 | 0.0009 |

NOTE: LMR: least-square linear model estimator of the regression parameters, LMO: least-square linear model estimator of the odds ratio parameters, SOR: maximum likelihood estimator based on the semiparametric odds ratio model for $p(Y_1 | G, E, S = 1)$, JOR: maximum likelihood estimator based on the semiparametric odds ratio model for $p(Y_1, G | E, S = 1)$. LS: unweighted least-square estimator based on parameters estimated from JOR. Bias: estimated-truth. Var: empirical variance estimated from the 1000 replicates. Mev: mean of the estimated variances over the 1000 replicates.
that based on SOR. When the $G-Y_2$ interaction is very strong, the least-square combination approach can be applied. Even though theory predicts that the weighted least square is more efficient, our simulation results cannot confirm it. For both WLS and LS, it appears that the estimated variances are substantially different from the empirical variance. This is due to the intrinsic difficulty in estimating the weakly identifiable parameters. Table 3 shows the estimates of the weakly identifiable parameters when $G-Y_2$ interaction is strong. The estimates appear to be reasonably well. When the interaction effect is not strong, we found that the feasibility of the least-square combination becomes an issue in our simulation. This most likely ties to the weak identifiability of the parameters. Overall, the simulation results suggest that the JOR estimator without the least-square combination for the odds ratio parameters is favorable. It is more efficient and stable, and relatively easy to obtain. The least-square combination is of value only when the actual $G-Y_2$ interaction is very strong.

5.3 Analysis of the High-Density Lipoprotein Study

High-density lipoprotein (HDL) is one of the five groups of lipoproteins that enable cholesterol to be transported in the water-based blood stream. It is measured in blood test by the HDL-C level, which is the amount of cholesterol contained in the HDL particles. HDL particles can remove cholesterol from within the artery atheroma and transport it back to liver for excretion or reutilization. As a result, high HDL-C level is associated with lower risk of cardiovascular diseases. The Upenn HDL cholesterol study is a cross-sectional study of the genetic factors associated with elevated HDL-C levels. Subjects of European ancestry with HDL-C level less than 30 percentile or greater than 90 percentile in their age and sex group are sampled as cases and controls, respectively. The study genotyped 625 cases and 606 controls using the IBC 50 K SNP array. This genome-wide association study had been analyzed previously as a case–control study in Edmondson et al. (2011) and as an illustrative example in Chen and Li (2011) for extreme-value sampling design.
In addition to the HDL, other relevant phenotypes such as the low-density lipoprotein (LDL), the total cholesterol (TC), and the triglycerides (TR) levels are measured. The correlation between HDL and LDL, TR, and TC estimated from the biased sample are, respectively, 0.22, −0.46, and 0.61. In this analysis, we investigate the effects of single-nucleotide polymorphism (SNP) genotypes on these phenotypes. As suggested by the theoretical discussion and the simulation studies, the marginal effects of G on individually phenotype are unidentifiable or requires unknown weakly identifiable parameter for identification. We analyze the conditional associations here. Two sets of analysis are performed for comparison. The first fits a semiparametric odds ratio model using the inefficient retrospective likelihood approach based on \( p(G \mid HDL, LDL, TC, TR) \), which models the odds ratio function as

\[
\log \eta(G; (HDL, LDL, TR, TC)) = \theta_1 G \ast LDL + \psi_{11}G \ast TR + \psi_{13}G \ast TC.
\]

The second fits a semiparametric odds ratio model using the efficient prospective likelihood approach based on \( p(HDL, G \mid LDL, TC, TG, S = 1) \) which has odds ratio functions

\[
\log \eta(HDL; G; (LDL, TR, TC)) = \theta_1G \ast HDL + \psi_{11}G \ast TR + \psi_{13}G \ast TC,
\]

where \( \psi_{11} = (\psi_{11}, \psi_{12}, \psi_{13}) \) and \( \psi_{13} = (\psi_{11}, \psi_{12}, \psi_{13}) \). The second analysis is theoretically more efficient but less robust. Parameters in such models can be interpreted using the following general approach. For example, to interpret \( \theta_1 \), we compute the log-conditional odds ratio function log \( \eta(HDL; LDL \mid (G, TR, TC)) \) as

\[
\log \eta(HDL; G; (LDL, TR, TC)) - \log \eta(HDL; G; (LDL = 0, TR, TC)) - \log \eta(HDL = 0; G; (LDL, TR, TC)) + \log \eta(HDL = 0; G; (LDL = 0, TR, TC)) = \theta_1 G \ast HDL \ast LDL.
\]

This means \( \theta_1 \) can be interpreted as one unit increase in LDL leads to an increase of the odds of observing HDL + 1 versus HDL by exp(\( \theta_1 \)) times when TR, TC, G are fixed.

The study genotyped more than 44,000 SNPs. There are 28,202 left after SNPs with minor allele frequency less than 5% are excluded. The two methods are applied to each SNP. The estimates are displayed in Table 4 if either test for nonzero association between a genotype and HDL adjusted for the measurements on LDL, TR, and TC is significant at the family error level 0.05 after Bonferroni adjustment for the multiple tests performed with either method. We choose this criterion because other traits are less likely to be detected as significant. All the detected SNPs are within CETP, a gene located on chromosome 16 that is well known to be associated with HDL. The result also shows that, overall, the complex OR model tends to be more efficient than the simple OR model. But it is not uniform across all SNP genotypes. For some genotypes, the simple OR model yields stronger or equal evidence against the hypothesis of no association. The evidence of genetic association for traits other than HDL are weak. However, the direction of the association is fairly consistent in that those identified SNPs have either positive associations with HDL, LDL and TC, and negative association with TR, or negative associations with HDL, LDL, and TC, and positive association with TR. The implication of the result needs to be examined further.

### 6. DISCUSSION

We have established the asymptotic properties for the likelihood inference on the odds ratio parameters in the

| SNP               | Phenotype | Simple OR model | Complex OR model |
|-------------------|-----------|-----------------|-----------------|
| rs3764261         | HDL       | −1.30 0.241 6.9e−8 | −1.30 0.241 6.9e−8 |
| rs17231056        | LDL       | −0.13 0.210 5.4e−1 | −0.12 0.217 5.8e−1 |
| rs11076175        | LDL       | −0.19 0.082 2.0e−2 | −0.19 0.082 2.0e−2 |
| rs711752          | LDL       | 1.20 0.237 4.1e−7 | 1.20 0.237 4.1e−7 |
| rs708272          | LDL       | −0.15 0.176 3.9e−1 | −0.16 0.176 3.6e−1 |
| rs7203984         | LDL       | 0.24 0.207 2.5e−1 | 0.24 0.207 2.5e−1 |
| rs11508026        | TR        | 0.20 0.105 5.7e−2 | 0.19 0.100 5.7e−2 |
| rs12720922        | TR        | 0.28 0.176 1.2e−1 | 0.28 0.176 1.1e−1 |
| rs9932240         | TC        | 0.30 0.131 3.2e−1 | 0.28 0.137 3.1e−1 |
| rs1532624         | TC        | −0.42 0.281 1.4e−1 | −0.37 0.255 1.2e−1 |
| rs11076175        | TC        | −0.10 0.155 1.5e−1 | −0.10 0.159 1.5e−1 |
| rs7499892         | HDL       | 1.70 0.346 9.0e−7 | 1.60 0.316 9.1e−7 |
| rs11076176        | HDL       | 0.47 0.315 1.4e−1 | 0.42 0.277 1.3e−1 |
| rs289714          | HDL       | −0.47 0.332 1.6e−1 | −0.45 0.292 1.2e−1 |
| rs1076176         | HDL       | −0.10 0.071 1.1e−1 | −0.10 0.071 1.1e−1 |
| rs7499892         | HDL       | 0.30 0.136 3.0e−1 | 0.29 0.128 3.0e−1 |
| rs289714          | HDL       | 0.10 0.071 1.1e−1 | 0.10 0.071 1.1e−1 |
| rs11076176        | HDL       | 0.25 0.110 2.3e−2 | 0.24 0.105 1.7e−2 |

NOTE: Est. = odds ratio estimate times 100, Sterr = standard error estimate of the odds ratio estimator times 100, SNP in the parentheses has the same estimates as the one not in parentheses.
semiparametric odds ratio model with possible misspecification of the odds ratio functions. The results can be applied to handling the weakly identifiable parameter in the analysis of biased sampling designs. We emphasize that the issue of weakly identifiability is common in nonmultiplicative models, such as probit model for case–control design or conditionally specified models under biased sampling designs, and is not a consequence of using semiparametric odds ratio model. The applications of the proposed approximation approach include among others, efficiency gain by exploiting gene–environment independence, secondary trait analysis, and the analysis of family-based designs, in addition to the extreme-value sampling designs discussed in this article.

7. SUPPLEMENTARY MATERIALS

A supplemental appendix for the proof of Theorem 2 is available online.

REFERENCES

Anderson, J. A. (1972), “Separate Sample Logistic Discrimination,” Biometrika, 59, 19–35. [1125]
Azzalini, A., and Valle, D. A. (1996), “The Multivariate Skew-Normal Distribution,” Biometrika, 83, 715–726. [1125]
Besag, J. (1974), “Spatial Interaction and the Statistical Analysis of Lattice Systems” (with discussion), Journal of the Royal Statistical Society, Series B, 36, 192–236. [1126]
——— (1996), “On Statistical Analysis of Dirty Pictures” (with discussion), Journal of the Royal Statistical Society, Series B, 48, 259–302. [1126]
Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), Discrete Multivariate Analysis: Theory and Practice, Cambridge, MA: MIT Press. [1125]
Breslow, N. E. (1976), “Regression Analysis of the Log Odds Ratio: A Method for Retrospective Studies,” Biometrics, 32, 409–416. [1125]
——— (1996), “Statistics in Epidemiology: The Case-Control Study,” Journal of the American Statistical Association, 91, 14–28. [1125]
Breslow, N. E., and Day, N. (1980), Statistical Methods in Cancer Research. Volume I: The Analysis of Case-Control Studies, Lyon: IARC Scientific Publications. [1125]
Chatterjee, N., and Carroll, R. J. (2005), “Semiparametric Maximum Likelihood Estimation Exploiting Gene-Environment Independence in Case-Control Studies,” Biometrics, 92, 399–418. [1125,1131]
Chen, H. Y. (2003), “A Note on Prospective Analysis of Outcome-Dependent Samples,” Journal of the Royal Statistical Society, Series B, 65, 575–584. [1125]
——— (2004), “Nonparametric and Semiparametric Models for Missing Covariates in Parametric Regressions,” Journal of the American Statistical Association, 99, 1176–1189. [1125,1126]
——— (2007), “A Semiparametric Odds Ratio Model for Measuring Association: Biometrics, 63, 413–421. [1125,1126,1127,1130]
——— (2010), “Compatibility of Conditionally Specified Models,” Statistics and Probability Letters, 80, 670–677. [1125,1126,1127,1130]
——— (2011), “A Unified Framework for Parameter Identifiability and Estimation in Biased Sampling Designs,” Biometrics, 98, 163–175. [1125,1126,1128,1300]
Chen, H. Y., and Chen, J. (2011), “On Information Coded in Gene-Environment Independence in Case-Control Design,” American Journal of Epidemiology, 174, 736–743. [1125,1131]
Chen, H. Y., Kittles, R., and Zhang, W. (2013), “Bias Correction in Secondary Traits Analysis With Case-Control Design,” Statistics in Medicine, 32, 1494–1508. [1131]
Chen, H. Y., and Li, M. (2011), “Improving Power and Robustness for Detecting Genetic Association With Extreme-Value Sampling Design,” Genetic Epidemiology, 35, 823–830. [1126,1133]
Chen, Y. H., Reilly, M. P., and Li, M. (2013), “Semiparametric Odds Ratio Model for Case-Control and Matched Case-Control Designs,” Statistics in Medicine, 32, 3126–3142. [1131]
Chen, Y. H., Chatterjee, N., and Carroll, R. J. (2008), “Retrospective Analysis of Haplotype-Based Case-control Studies Under a Flexible Model for Gene-Environment Association,” Biostatistics, 9, 81–99. [1125,1131]
——— (2009), “Shrinkage Estimators for Robust and Efficient Inference in Haplotype-Based Case-Control Studies,” Journal of the American Statistical Association, 104, 220–233. [1125,1131]
Dacunha-Castelle, D., and Gasqiat, E. (1999), “Testing the Order of a Model Using Locally Conic Parameterization: Population Mixture and Stationary ARMA Processes,” The Annals of Statistics, 27, 1178–1209. [1125]
Edmondson, A. C., Braund, P. S., Stylianou, I. M., Khera, A. V., Nelson, C. P., Wolfe, M. L., Derrohanessian, S. L., Keating, B. J., Qu, L., He, J., Tobin, M. D., Tomaszewski, M., Baumert, J., Kloppe, N., Frodina, A., Thorand, B., Li, M., Reilly, M. P., Koenig, W., Samani, N. J., and Rader, D. J. (2011), “Dense Genotyping of Candidate Gene Loci Identifies Variants Associated With High-Density Lipoprotein Cholesterol,” Circulation Cardiovascular Genetics, 4, 145–155. [1133]
Feldt, L. S. (1961), “The Use of Extreme Groups to Test for the Presence of a Relationship,” Psychometrika, 26, 307–316. [1125]
Gilbert, P. B., Lele, S. R., and Vardi, Y. (1999), “Maximum Likelihood Estimation in Semiparametric Selection Bias Models With Application to AIDS Vaccine Trials,” Biometrika, 86, 27–43. [1128]
Huang, A., and Rathouz, P. J. (2012), “Proportional Likelihood Ratio Models for Mean Regression,” Biometrika, 99, 223–229. [1128]
Li, D. Y., and Zeng, D. (2006), “Likelihood-Based Inference on Haplotype Effects in Genetic Association Studies” (with discussion), Journal of the American Statistical Association, 101, 89–104. [1125,1131]
——— (2009), “Proper Analysis of Secondary Phenotype Data in Case-Control Association Studies,” Genetic Epidemiology, 33, 256–265. [1125,1131]
Lindsay, B. G. (1995), Mixture Models: Theory, Geometry and Applications, Hayward, CA: IMS. [1125]
Liu, X., and Shao, Y. (2003), “Asymptotics for Likelihood Ratio Test Under Loss of Identifiability,” The Annals of Statistics, 31, 807–823. [1125]
Luo, X., and Tsai, W. Y. (2012), “A Proportional Likelihood Ratio Model,” Biometrika, 99, 211–222. [1126]
Monesse, G. M., Tamimi, R. M., and Kraft, P. (2009), “Genome-Wide Association Scans for Secondary Traits Using Case-Control Samples,” Genetic Epidemiology, 33, 718–728. [1131]
Osius, G. (2004), “The Association Between Two Random Elements: A Complete Characterization and Odds Ratio Models,” Metrika, 60, 261–277. [1125]
——— (2009), “Asymptotic Inference for Semiparametric Association Models,” The Annals of Statistics, 37, 459–489. [1125,1126]
Prentice, R. L., and Pyke, R. (1979), “Logistic Disease Incidence Models and Case-Control Studies,” Biometrika, 66, 403–411. [1125]
Qin, J. (1998), “Inferences for Case-Control Data and Semiparametric Two-Sample Density Ratio Model,” Biometrika, 85, 619–630. [1128]
Rubinowicz, D. (1997), “A Note on Efficient Estimation From Case-Control Data,” Biometrika, 84, 486–488. [1125]
Rathouz, P. J., and Gao, L. P. (2009), “Generalized Linear Models With Unspecified Reference Distribution,” Biostatistics, 10, 205–218. [1128]
Risch, N. J., and Zhang, H. (1995), “Extreme Discordant Sib Pairs for Mapping Quantitative Trait Loci in Humans,” Science, 268, 1584–1589. [1126]
——— (1996), “Mapping Quantitative Trait Loci With Extreme Discordant Sib Pairs: Sampling Considerations,” The American Journal of Human Genetics, 58, 836–843. [1126]
Scott, A. J., and Wild, C. J. (1997), “Fitting Regression Models to Case-control Data by Maximum Likelihood,” Biometrika, 84, 57–71. [1125]
Song, R., Kosorok, M. R., and Fine, J. P. (2009), “On Asymptotically Optimal Tests Under Loss of Identifiability in Semiparametric Models,” The Annals of Statistics, 37, 2409–2444. [1125]
Zhang, H., and Risch, N. (1996), “Mapping Quantitative-trait Loci in Human by Use of Extreme Concordant Sib Pairs: Selected Sampling by Parental Phenotypes,” American Journal of Human Genetics, 59, 951–957. [1125]
Zhu, H., and Zhang, H. (2006), “Asymptotics for Estimation and Testing Procedures Under Loss of Identifiability,” Journal of Multivariate Analysis, 97, 19–45. [1125]