Journal of Data Science 0 (0), 1–16
??? 2024
DOI: 10.6339/24-JDS1155
Data Science Reviews

Bibliographical Connections for Semiparametric Analysis in Case-Control Studies on Gene-Environment Interactions

Tianying Wang[1], Jianxuan Liu[2,*], and Aijing Wu[2]

[1]*Department of Statistics, Colorado State University, Fort Collins, CO, USA*
[2]*Department of Mathematics, Syracuse University, Syracuse, NY, USA*

Abstract

Analyzing the gene-environment interaction (GEI) is crucial for understanding the etiology of many complex traits. Among various types of study designs, case-control studies are popular for analyzing gene-environment interactions due to their efficiency in collecting covariate information. Extensive literature explores efficient estimation under various assumptions made about the relationship between genetic and environmental variables. In this paper, we comprehensively review the methods based on or related to the retrospective likelihood, including the methods based on the hypothetical population concept, which has been largely overlooked in GEI research in the past decade. Furthermore, we establish the methodological connection between these two groups of methods by deriving a new estimator from both the retrospective likelihood and the hypothetical population perspectives. The validity of the derivation is demonstrated through numerical studies.

Keywords *gene-environment interaction; hypothetical population; retrospective likelihood; semiparametric analysis*

1 Introduction

With growing research studies on gene-trait associations, such as genome-wide association studies (GWAS), numerous discoveries show that the risk of complex diseases is determined by the combined effects of genetic and environmental or non-genetic exposures (Hunter, 2005; Hutter et al., 2013; Meisner et al., 2019; Emdin et al., 2021; Gauderman et al., 2017; Murcray et al., 2009; Gauderman et al., 2013). Gene-environment interactions have also attracted interest in various domains such as agricultural genetics (Crossa, 2012), cancer genetics (Hunter, 2005), and environmental genetics (Thomas, 2010). To study the effects of gene-environment interaction, developing statistical methods under efficient designs, such as case-control study designs, is essential. The goal of this paper is to review recently proposed statistical approaches for gene-environment interaction analysis in case-control studies and shed light on their statistical connections.

Common statistical inference methods for case-control studies are based on a standard prospective likelihood or a retrospective likelihood with a gene-environment independence assumption (Han and Chatterjee, 2018). Among them, semiparametric models are attractive as no assumptions are made on the distribution of the environmental variables. Logistic regression is one of the standard methods based on a prospective likelihood of case-control data (Prentice and

*Corresponding author. Email: jliu193@syr.edu.

Received June 30, 2024; Accepted September 23, 2024

Pyke, 1979; Breslow et al., 2000). Though the retrospective nature of the sampling scheme is ignored, logistic regression is consistent on non-intercept coefficients regardless of the retrospective sampling scheme, and it is the most efficient approach if no additional assumption has been made (Prentice and Pyke, 1979). However, logistic regression requires a large sample size to achieve substantial statistical power in case-control studies, especially for detecting gene-environment interactions. Thus, additional assumptions, such as parametric or semiparametric structures for the covariates, are exploited to improve the estimation efficiency. For example, the rare disease assumption has been imposed in Piegorsch et al. (1994a), yet it has been shown later that such an assumption could lead to estimation bias when the disease has moderate prevalence or with a small marginal probability in the source population (Chatterjee and Carroll, 2005).

To improve the estimation efficiency, the gene-environment independence assumption, for which the genetic variable is assumed to be independent of the environmental variables in the source population, is commonly exploited (Chatterjee and Carroll, 2005). Under this assumption, Chatterjee and Carroll (2005) shows that logistic regression is inefficient, whereas a more efficient estimate of the gene-environment interaction can be obtained by using the profile likelihood technique. In their work, the genetic variable is assumed to be discrete, and the environmental variable is fully nonparametric. Some follow-up studies on this topic can be found in Chatterjee et al. (2005); Chen et al. (2008, 2009); Han et al. (2012); Lobach et al. (2008); Luo et al. (2009); Mukherjee et al. (2012); Ma (2010); Spinka et al. (2005), whereas *parametric* modeling of the distribution genetic variable given environmental variables has been exploited. Such assumptions on the genetic variable make those methods less practical because the genetic variable can be continuous when the polygenic risk score is considered (Crouch and Bodmer, 2020; Khera et al., 2018; Curtis, 2018).

To relax the assumptions on the genetic variable, Stalder et al. (2017) showed that only an expectation of a simple function of the genetic variables given the environmental variables is essential, rather than the explicit distribution of the genetic variables. By showing that the expectation can be consistently estimated with fully unspecified genetic distribution, Stalder et al. (2017) proposed a new estimator, which does not require any distributional assumptions on both genetic and environmental variables but only assumes the gene-environment independence. Recently, Wang and Asher (2021) proposed to further improve the efficiency of estimating gene-environment interaction terms by utilizing the overlooked mathematical symmetry in the method of Stalder et al. (2017) with no additional assumptions required. Different from the aforementioned works based on retrospective likelihood framework, Ma (2010) proposed a framework by introducing a hypothetical population; Liang et al. (2019) further extended this idea to model both environmental and genetic variables in a nonparametric fashion.

In this paper, we aim to bridge the connections between the two predominant approaches, namely, hypothetical population and retrospective likelihood, used in case-control studies for gene-environment interactions. We first provide a comprehensive methodological review of the semiparametric analysis in this context, which allow fully unspecified distribution of both genetic and environmental variables under the gene-environment independence assumption. In particular, many of those methods, especially the ones based on the hypothetical population, are overlooked in gene-environment interaction analysis research. Second, we establish the connection between the methods based on retrospective likelihood and those based on hypothetical population. We illustrate their connections by deriving new algorithms from both the retrospective likelihood and the hypothetical population perspectives. This connection offers a unified framework that allows researchers to better understand the theoretical relationships between these methods. While computational gains or improvements in efficiency are not the primary

focus of this estimator, the contribution lies in shedding new light on the fundamental similarities and differences between the two predominant approaches, which is critical for advancing the theoretical development of semiparametric models in this context. Furthermore, the new algorithm we derive from the hypothetical population perspective does not require constructing the nuisance tangent space and its orthogonal complement in a Hilbert space or solving complex integral equations which are often required in other semiparametric approaches. The algorithms can be easily implemented without additional assumptions rather than gene-environment independence. We illustrate its validity through simulation studies.

2 Review of Estimation Methods

We first introduce the notations and background of the gene-environment interaction problem in case-control studies. Then we review two types of methods based on (1) the retrospective likelihood framework: Chatterjee and Carroll (2005); Stalder et al. (2017), and Wang and Asher (2021), and (2) the hypothetical population framework: Ma (2010) and Liang et al. (2019).

2.1 Notations

Denote the genetic information by G, environmental exposures by X, and the disease status by D, where $D = 1$ for cases and $D = 0$ for controls. For a case-control study, let n_1 be the number of cases and n_0 be the number of controls, while $\pi_1 = P(D = 1)$ is the disease prevalence rate in the source population and $\pi_0 = 1 - \pi_1$. Further, denote $f_{G,X}(\cdot, \cdot)$ as the joint density or mass of X and G in the source population, and $f_G(\cdot)$ and $f_X(\cdot)$ as the marginal density or mass functions of G and X, respectively. Then, the well-accepted gene-environment independence assumption can be written as $f_{X,G}(x, g) = f_G(g) \times f_X(x)$. Both Chatterjee and Carroll (2005) and Ma (2010) leave $f_X(x)$ unspecified and assume G to be discrete or continuous. Recent advancements allow both the distributions of X and G to be arbitrary and treated as infinite-dimensional nuisance parameters in the work of Stalder et al. (2017), Liang et al. (2019), and Wang and Asher (2021).

Assume we have case-control observations $(D_i = 0, X_i, G_i)$, $i = 1, \ldots, n_0$ and $(D_i = 1, X_i, G_i)$, $i = n_0 + 1, \ldots, n = n_0 + n_1$. The aforementioned works assume that the risk of disease, given genetic and environmental factors in the source population, follows the logistic distribution function:

$$P(D = 1|X_i, G_i, \alpha, \boldsymbol{\beta}) = \frac{\exp\{\alpha + m(X_i, G_i, \boldsymbol{\beta})\}}{1 + \exp\{\alpha + m(X_i, G_i, \boldsymbol{\beta})\}},$$

where $m(X_i, G_i, \boldsymbol{\beta})$ is a parametric function defining the joint effect of X and G; $\boldsymbol{\beta}$ is the parameter we are interested in. Further, we denote

$$h(D_i, X_i, G_i) \equiv \frac{\exp\{D_i(\alpha + m(X_i, G_i, \boldsymbol{\beta}))\}}{1 + \exp\{\alpha + m(X_i, G_i, \boldsymbol{\beta})\}}. \tag{1}$$

Prentice and Pyke (1979) showed that when the joint distribution of X and G is unspecified, ignoring the scheme of case-control studies and fitting the standard prospective logistic regression is equivalent to maximum likelihood estimation, leading to the consistent estimate of $\boldsymbol{\beta}$. Under the assumption of the independence between genetic factors G and the environmental exposures X, α and $\boldsymbol{\beta}$ are identifiable (Chatterjee and Carroll, 2005, Lemma 1).

2.2 Methods Based on Retrospective Likelihood

The methods of Chatterjee and Carroll (2005), Stalder et al. (2017), and Wang and Asher (2021) follow the retrospective likelihood framework. Given the disease status of each subject, the retrospective likelihood is

$$\begin{aligned} & P(G=g, X=x|D=d) \\ = & \frac{P(G=g, X=x, D=d)}{P(D=d)} \\ = & \frac{P(D=d|X=x, G=g)P(G=g, X=x)}{\sum_t \sum_s P(D=d|G=t, X=x)P(G=t, X=s)}, \end{aligned}$$

if X and G are discrete. Similarly, if X and G are continuous, the retrospective likelihood takes the form

$$\frac{f_G(g)f_X(x)\exp[d\{\alpha_0+m(g,x,\beta)\}]/[1+\exp\{\alpha_0+m(g,x,\beta)\}]}{\int f_G(u)f_X(v)\exp[d\{\alpha_0+m(u,v,\beta)\}]/[1+\exp\{\alpha_0+m(u,v,\beta)\}]dudv}.$$

As the logistic intercept α_0 converges to $\kappa=\alpha_0+\log(n_1/n_0)-\log(\pi_1-\pi_0)$ (Prentice and Pyke, 1979), some approaches reparameterized α_0 in terms of κ.

In the method of Chatterjee and Carroll (2005), X is treated as discrete. Then by profiling out $f_X(\cdot)$, a semiparametric profile likelihood can be obtained as

$$L_X(D,G,X,\kappa,\beta,f_G)=f_G\frac{S(D,G,X,\kappa,\beta)}{R_X(X,\kappa,\beta)},$$

where

$$\begin{aligned} S(d,g,x,\kappa,\beta) &= \frac{\exp[d\{\kappa+m(g,x,\beta)\}]}{1+\exp\{\kappa-\log(n_1/n_0)+\log(\pi_1/\pi_0)+m(g,x,\beta)\}}, \\ R_X(x,\kappa,\beta) &= \sum_{r=0}^{1}\int f_G(v)S(r,v,x,\kappa,\beta)dv. \end{aligned}$$

Based on the method of Chatterjee and Carroll (2005), Stalder et al. (2017) further developed an unbiased estimator of $R_X(x,\kappa,\beta)$ with $f_G(\cdot)$ being treated nonparametrically, denoted as $\hat{R}_X$. Define $\Omega=(\kappa,\beta^\top)^\top$. Then, the score function for the profile likelihood can be estimated consistently by

$$\hat{S}_X(\Omega)=n^{-1/2}\sum_{i=1}^{n}\left\{\frac{S_\Omega(D_i,G_i,X_i,\Omega)}{S(D_i,G_i,X_i,\Omega)}-\frac{\hat{R}_{X\Omega}(X_i,\Omega)}{\hat{R}_X(X_i,\Omega)}\right\},$$

where $S_\Omega(d,g,x,\Omega)=\partial S(d,g,x,\Omega)/\partial\Omega$ and $\hat{R}_{X\Omega}(x,\Omega)=\partial\hat{R}_X(x,\Omega)/\partial\Omega$. The consistent estimate of Ω, namely $\hat{\Omega}_X$, is obtained by solving equation $\hat{S}_X(\Omega)=0$.

Based on Stalder et al. (2017), Wang and Asher (2021) further improved the efficiency of the estimate by observing the mathematical symmetry of X and G. Wang and Asher (2021) proposed to swap X and G in the method of Stalder et al. (2017), and obtained an estimate by profiling G out, namely $\hat{\Omega}_G$. Then, Wang and Asher (2021) proposed an optimal combination of the symmetric estimators $\hat{\Omega}_X$ and $\hat{\Omega}_G$: $\hat{\Omega}_{\text{Combo}}=(\mathcal{X}^\top\Lambda^{-1}\mathcal{X})^{-1}\mathcal{X}^\top\Lambda^{-1}\mathcal{Y}$ where $\mathcal{X}=(I_p,I_p)^\top$, $\mathcal{Y}=(\hat{\Omega}_X^\top,\hat{\Omega}_G^\top)^\top$ and $\Lambda=cov(\mathcal{Y})$. p is the length of the vector Ω. Wang and Asher (2021) showed that $\hat{\Omega}_{\text{Combo}}$ is guaranteed to have an improved (or at least no worse) estimation efficiency than the method of Stalder et al. (2017) on the gene-environment interaction.

2.3 Methods Based on Hypothetical Population

Given the same statistical task in case-control studies, Ma (2010) proposed the concept of *hypothetical population* and derived the efficient estimator under a semiparametric model. The notation and the model remain the same as described before. It is well known that the samples from a case-control study are sampled conditioning on the disease status, which violates the random sampling scheme. Different from the method of Chatterjee and Carroll (2005), where the Lagrange multiplier argument is necessary, Ma (2010) considered a hypothetical population with infinite population size, and the disease to non-disease ratio is fixed at n_1/n_0. The hypothetical population probability density/mass function of (D_i, X_i, G_i) is

$$\begin{aligned} & f^s(X_i, G_i, D_i) \\ = \; & f^s_D(D_i) f^s_{X,G|D}(X_i, G_i \mid D_i) \\ = \; & \frac{n_{D_i}}{n} f^t_{X,G|D}(X_i, G_i \mid D_i) \\ = \; & \frac{n_{D_i}}{n} \frac{f^t_X(X_i) f^t_G(G_i) h(D_i, X_i, G_i)}{\int f^t_X(x) f^t_G(g) h(D_i, x, g) d\mu(x) d\mu(g)}, \end{aligned} \tag{2}$$

where $h(D_i, X_i, G_i)$ is defined in (1). Quantities under the true model have a superscript ^{t}, while under the hypothetical population, they have a superscript ^{s}. The case-control data form an i.i.d. sample of the hypothetical population with the above joint distribution. Ma (2010) showed that the case-control sample could be viewed as an i.i.d. random sample from the hypothetical population of interest, and the usual semiparametric analysis as exemplified by Bickel et al. (1993) and Tsiatis (2006) can be applied naturally. Through a geometric approach, Ma (2010) constructed an estimator by projecting the score vector of the parameter onto the orthogonal complement of the nuisance tangent space, $\Lambda^{\perp} = [h(D, X, G) : E\{h(D, X, G) \mid X\} = E\{E[h(D, X, G) \mid D] \mid X\}]$. Such a procedure can bypass estimating the unspecified distribution of X, and the resulting estimator still achieves optimal efficiency.

Liang et al. (2019) also solves this problem from the hypothetical population perspective. Liang et al. (2019) treated both of the unspecified distributions of X and G as infinite-dimensional parameters. The estimation is made by constructing a Hilbert space and decomposing it into nuisance tangent space and its orthogonal complement. The resulting efficient score is $\mathbf{S}_{\text{eff}}(D, X, G) = \mathbf{S} - \mathbf{a}(G) - \mathbf{b}(X) - E(\mathbf{S} \mid D) + E\{\mathbf{a}(G) + \mathbf{b}(X) \mid D\}$, where $\mathbf{S} = \mathbf{S}(D, X, G) = \{D - h(1, X, G)\}[\{\partial m(D, X, G)/\partial \boldsymbol{\beta}\}^{\mathrm{T}}, 1]^{\mathrm{T}}$, and $\mathbf{a}(G)$ and $\mathbf{b}(X)$ satisfy

$$E\{\mathbf{a}(G) \mid X\} - \mathbf{b}(X) - E\{E(\mathbf{a} + \mathbf{b} \mid D) \mid X\} = E(\mathbf{S} \mid X) - E\{E(\mathbf{S} \mid D) \mid X\}, \tag{3}$$

and

$$\mathbf{a}(G) + E\{\mathbf{b}(X) \mid G\} - E\{E(\mathbf{a} + \mathbf{b} \mid D) \mid G\} = E(\mathbf{S} \mid G) - E\{E(\mathbf{S} \mid D) \mid G\}. \tag{4}$$

While the distribution function of X drops in the procedure in Ma (2010), Liang et al. (2019) treated both unspecified distributions as nuisance parameters and proposed to estimate them through kernel methods by conditioning on the disease status. When X or G is continuous, numerical approximation, such as a discretizing technique (Tsiatis and Ma, 2004; Liu and Ma, 2019), was adopted. Due to the hypothetical population mechanisms, the case-control sample can be viewed as a simple random sample. Thus, the classic semiparametric methods explained in Bickel et al. (1993) and Tsiatis (2006) are applicable in Liang et al. (2019).

2.4 Rare Diseases when π_1 Is Unknown

When the disease rate in the source population, i.e., π_1, is unknown and hard to estimate due to its rareness, a rare disease assumption is often assumed in case-control studies (Piegorsch et al., 1994b; Modan et al., 2001; Lin and Zeng, 2006). When $\pi_1 \approx 0$ is assumed, the estimator of $\Omega = (\kappa, \beta^\top)^\top$ on the three retrospective likelihood methods could be slightly biased. That is, $\hat{\Omega}$ converges to Ω^* instead of Ω, where Ω^* is the solution to the estimating equation with $\pi_1 = 0$. In Stalder et al. (2017) and Wang and Asher (2021), several simulation studies in these three papers showed that small bias would be introduced due to the rare disease assumption or misspecified π_1, but it has little effect on the coverage probabilities of confidence intervals. We generate the environmental exposure X from a standard normal distribution while the genetic susceptibility G is binary with probability 0.6. The true π_1 is 4.5%. We reproduced their results and evaluated the performance of the methods in Stalder et al. (2017), and Wang and Asher (2021) under settings with various misspecified π_1 (see Table 1). Though the bias and coverage rate of 95% confidence intervals are relatively stable against the misspecified π_1 as reported in the original papers, we observed a decrease in the efficiency for each coefficient. Meanwhile, though Ma (2010) estimates π_1 through a sample mean approximation, the methods of Ma (2010) and Liang et al. (2019) do not require a pre-specified π_1, making the hypothetical population thread of methods appealing when π_1 is unknown.

3 Connecting the Two Threads of Work

We here illustrate a new simple estimator, which can be derived from Chatterjee and Carroll (2005) and Ma (2010) without any additional assumptions other than gene-environment independence. Of note, Ma (2010) shows that the results of discrete G are very similar to that in Chatterjee and Carroll (2005) based on numerical experiments, yet we connect these two methods by deriving the estimator from both the retrospective likelihood perspective and the hypothetical population perspective. Further, the derivation is illustrated by unspecified X and G distributions, assuming they are discrete. When X or G is continuous, the derivation is similar. Nonparametric kernel density estimation is adopted to estimate the unknown probability density/mass functions.

3.1 Derivation Based on Hypothetical Population

We adopt the hypothetical population notion with a disease to non-disease ratio is n_1/n_0, the case-control data form an i.i.d. sample from the hypothetical population with probability distribution function (2). Then, the log-likelihood is

$$\begin{aligned}\mathcal{L} \;=\; & \sum_{i=1}^{n} \log\left(\frac{n_{D_i}}{n}\right) + \log\{f_X^t(X_i)\} + \log\{f_G^t(G_i)\} + \log\{h(D_i, X_i, G_i)\} \\ & - \log\left\{\int f_X^t(x) f_G^t(g) h(D_i, x, g) dx dg\right\}. \end{aligned} \tag{5}$$

Maximizing $\mathcal{L}$ with respect to $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ and the supporting points of $f_X^t(x)$, $f_G^t(g)$. Recall that as in Chatterjee and Carroll (2005), we treat the density of X as discrete on the set of distinct observed values $(x_1, \ldots, x_K)$ with probability $\gamma_k = P(X = x_k), k = 1, \ldots, K$. Similarly, we assume

Table 1: Misspecified π_1 in Stalder et al. (2017) (Spmle), and Wang and Asher (2021) (Combo). Results of 100 simulations with $n_0 = n_1 = 100$. $X \sim N(0,1)$, $G \sim$ Bernoulli(0.6). ESE is the empirical standard error. ASE is the asymptotic standard error. CI is the 95% confidence interval. EFF is the efficiency.

		β_X	β_G	β_{XG}	β_X	β_G	β_{XG}
		use true π_1			set $\pi_1 = 0.06$		
	Bias	−0.0445	−0.0896	0.0388	−0.0517	−0.0906	0.0522
	ESE	0.2357	0.3381	0.2553	0.2366	0.3385	0.2576
Spmle	ASE	0.2195	0.3122	0.2411	0.2204	0.3118	0.2449
	MSE	0.0570	0.1212	0.0660	0.0581	0.1216	0.0684
	CI	0.9500	0.9300	0.9200	0.9500	0.9300	0.9300
	EFF	1.1723	1.0423	1.4122	1.1499	1.0386	1.3625
	Bias	0.0691	−0.0536	−0.0421	0.0671	−0.0584	−0.0330
	ESE	0.2056	0.3430	0.2294	0.2099	0.3390	0.2332
Combo	ASE	0.2087	0.3173	0.2350	0.2116	0.3170	0.2407
	MSE	0.0466	0.1194	0.0539	0.0481	0.1172	0.0549
	CI	0.9200	0.9400	0.9400	0.9200	0.9400	0.9400
	EFF	1.4318	1.0585	1.7306	1.3880	1.0781	1.6972
		set $\pi_1 = 0.1$			set $\pi_1 = 0.2$		
	Bias	−0.0676	−0.0936	0.0812	−0.0968	−0.1026	0.1366
	ESE	0.2397	0.3397	0.2655	0.2509	0.3429	0.2891
Spmle	ASE	0.2235	0.3114	0.2546	0.2309	0.3114	0.2786
	MSE	0.0615	0.1230	0.0764	0.0717	0.1270	0.1014
	CI	0.9700	0.9300	0.9300	0.9600	0.9300	0.9300
	EFF	1.0864	1.0272	1.2209	0.9317	0.9951	0.9194
	Bias	0.0596	−0.0645	−0.0153	0.0548	−0.0688	0.0124
	ESE	0.2182	0.3335	0.2451	0.2441	0.3266	0.2800
Combo	ASE	0.2180	0.3164	0.2545	0.2314	0.3176	0.2858
	MSE	0.0507	0.1143	0.0597	0.0620	0.1103	0.0778
	CI	0.9400	0.9400	0.9400	0.9200	0.9500	0.9600
	EFF	1.3171	1.1055	1.5617	1.0777	1.1451	1.1990

G also follows a discrete distribution on the set $(g_1, \ldots, g_S)$ with probability $\xi_s = P(G = g_s)$, $s = 1, \ldots, S$. Let $f_X^t(x_i) = \gamma_i$ and $f_G^t(g_i) = \xi_i$, then, we obtain the following gradient equations

$$\begin{aligned}
& \sum_{i=1}^{n} \left\{ \frac{\partial h(D_i, X_i, G_i)/\partial \boldsymbol{\theta}}{h(D_i, X_i, G_i)} - \frac{\int f_X^t(x) f_G^t(g) \partial h(D_i, x, g)/\partial \boldsymbol{\theta} dx dg}{\int f_X^t(x) f_G^t(g) h(D_i, x, g) dx dg} \right\} \\
= & \sum_{i=1}^{n} \left\{ \frac{\partial h(D_i, X_i, G_i)/\partial \boldsymbol{\theta}}{h(D, X_i, G_i)} - \frac{\sum_{s=1}^{S} \sum_{k=1}^{K} \gamma_k \xi_s \partial h(D_i, x_k, g_s)/\partial \boldsymbol{\theta}}{\sum_{s=1}^{S} \sum_{k=1}^{K} \gamma_k \xi_s h(D_i, x_k, g_s)} \right\} \\
= & \mathbf{0}.
\end{aligned}$$

Maximizing (5) with respect to γ and ξ yields

$$\frac{1}{\gamma_k} - \frac{n_0 \sum_{s=1}^{S} \xi_s h(0, x_k, g_s)}{\sum_{s=1}^{S}\sum_{k=1}^{K} \gamma_k \xi_s h(0, x_k, g_s)} - \frac{n_1 - n_1 \sum_{s=1}^{S} \xi_r h(0, x_k, g_s)}{1 - \sum_{s=1}^{S}\sum_{k=1}^{K} \gamma_k \xi_s h(0, x_k, g_s)} = 0,$$
$$\frac{1}{\xi_s} - \frac{n_0 \sum_{k=1}^{K} \gamma_k h(0, x_k, g_s)}{\sum_{s=1}^{S}\sum_{k=1}^{K} \gamma_k \xi_s h(0, x_k, g_s)} - \frac{n_1 - n_1 \sum_{k=1}^{K} \gamma_k h(0, x_k, g_s)}{1 - \sum_{s=1}^{S}\sum_{k=1}^{K} \gamma_k \xi_s h(0, x_k, g_s)} = 0,$$

for $k = 1, \ldots, K$ and $s = 1, \ldots, S$. Also, $\sum_{k=1}^{K} \gamma_k = 1$, $\sum_{s=1}^{S} \xi_s = 1$. Equivalently,

$$\sum_{i=1}^{n} \left\{ \frac{\partial h(D_i, X_i, G_i)/\partial \boldsymbol{\theta}}{h(D_i, X_i, G_i)} - \frac{\sum_{s=1}^{S}\sum_{k=1}^{K} \gamma_k \xi_s \partial h(D_i, x_k, g_s)/\partial \boldsymbol{\theta}}{\sum_{s=1}^{S}\sum_{k=1}^{K} \gamma_k \xi_s h(D_i, x_k, g_s)} \right\} = \mathbf{0}, \tag{6}$$

$$\begin{aligned} \frac{n_1}{\pi_1} + \left(\frac{n_0}{\pi_0} - \frac{n_1}{\pi_1}\right) \sum_{s=1}^{S} \xi_s h(0, x_k, g_s) &= \frac{1}{\gamma_k}, \\ \frac{n_1}{\pi_1} + \left(\frac{n_0}{\pi_0} - \frac{n_1}{\pi_1}\right) \sum_{k=1}^{K} \gamma_k h(0, x_k, g_s) &= \frac{1}{\xi_s}, \end{aligned} \tag{7}$$

$$\sum_{k=1}^{K} \gamma_k = 1, \quad \sum_{s=1}^{n} \xi_s = 1. \tag{8}$$

Here the disease rate in the source population, $\pi_1 = \sum_{s=1}^{S} \xi_s \{\sum_{k=1}^{K} \gamma_k h(1, x_k, g_s)\}$, and $\pi_0 = \sum_{s=1}^{S} \xi_s \{\sum_{k=1}^{K} \gamma_k h(0, x_k, g_s)\}$. Using the nonparametric maximum likelihood estimation approach, we solve equations (6)-(8) simultaneously for α, $\boldsymbol{\beta}^{\mathrm{T}}$ and the nuisance parameters γ_k's, ξ_s's. However, for example, when $m(X_i, G_i, \boldsymbol{\beta}) = \boldsymbol{\beta}_1^{\mathrm{T}} X_i + \boldsymbol{\beta}_2^{\mathrm{T}} G_i + \boldsymbol{\beta}_3^{\mathrm{T}} X_i G_i$, solving at least $2n+4$ equations can be computationally challenging. We consider a profile likelihood approach. According to the above setting, the loglikelihood is

$$\mathcal{L}_1 = \sum_{i=1}^{n} \left[\log\left(\frac{n_{D_i}}{n}\right) + \log(\gamma_i) + \log(\xi_i) + \log\{h(D_i, X_i, G_i)\} \right] - n_1 \log(\pi_1) - n_0 \log(\pi_0).$$

From (7), we have

$$\begin{aligned} \gamma_k &= \left\{ \frac{n_1}{\pi_1} + \left(\frac{n_0}{\pi_0} - \frac{n_1}{\pi_1}\right) \sum_{s=1}^{S} \xi_s h(0, x_k, g_s) \right\}^{-1}, \\ \xi_s &= \left\{ \frac{n_1}{\pi_1} + \left(\frac{n_0}{\pi_0} - \frac{n_1}{\pi_1}\right) \sum_{k=1}^{K} \gamma_k h(0, x_k, g_s) \right\}^{-1}. \end{aligned} \tag{9}$$

Although we share the same spirit of Ma (2010) and Liang et al. (2019) regarding solving estimating equations for $\boldsymbol{\beta}$, our approach is more straightforward and simple to implement. Specifically, we derive the estimating equations directly from the log-likelihood without constructing a Hilbert space or searching for the nuisance tangent space and its orthogonal complement. It is known that solving (3) and (4) is practically difficult. Our approach does not need to solve such complex equations. Furthermore, the log-likelihood we derive from a hypothetical population perspective can be achieved from a retrospective likelihood perspective, which unifies the two major threads of methods in case-control studies. We summarize the algorithm in Section 3.3.

3.2 Derivation Based on Retrospective Likelihood

We can also derive the same conclusion based on the retrospective likelihood methods. Following the profile likelihood framework by Chatterjee and Carroll (2005), Stalder et al. (2017) profiled out X and estimated the part related to the distribution of G unbiasedly and nonparametrically. Rather than using the two-step approach as in Stalder et al. (2017), we simultaneously profile out X and G and maximize the corresponding complete retrospective likelihood. The main idea is to discretize the densities of X and G, i.e., $f_X(x)$ and $f_G(g)$, respectively. Then, one can maximize the retrospective likelihood iteratively as an optimization problem.

Under the assumption that X and G are independent, we profile X out then maximize retrospective likelihood over $(\gamma_1, \ldots, \gamma_K)$, leading to

$$\gamma_k = \frac{\sum_{i=1}^{n} I(X_i = x_k)}{n \sum_{d,s} P(D = d | X = x_k, G = g_s)\xi_s \mu_d}, \ k = 1, \ldots, K, \tag{10}$$

where $\mu_d = n_d/(n\pi_d)$, $\pi_d = P(D = d)$. Due to the independence of G and X, we can also profile G out as follows:

$$\xi_s = \frac{\sum_{i=1}^{n} I(G_i = g_s)}{n \sum_{d,k} P(D = d | X = x_k, G = g_s)\gamma_k \mu_d}, \ s = 1, \ldots, S. \tag{11}$$

Note that $\sum_{i=1}^{n} I(X_i = x_k)$ and $\sum_{i=1}^{n} I(G_i = g_s)$ are not guaranteed to be equal to 1. These above equations can be simplified in reality once more information is provided. Hence, γ_i, ξ_i, π_1 and π_0 can be calculated as

$$\begin{aligned}
\gamma_k &= \{\sum_{i=1}^{n} I(X_i = x_k)\} \left\{ \frac{n_1}{\pi_1} \sum_{s=1}^{S} \xi_s + \left(\frac{n_0}{\pi_0} - \frac{n_1}{\pi_1} \right) \sum_{s=1}^{S} h(0, x_k, g_s)\xi_s \right\}^{-1}, \\
\xi_s &= \{\sum_{i=1}^{n} I(G_i = g_s)\} \left\{ \frac{n_1}{\pi_1} \sum_{k=1}^{K} \gamma_k + \left(\frac{n_0}{\pi_0} - \frac{n_1}{\pi_1} \right) \sum_{k=1}^{K} h(0, x_k, g_s)\gamma_k \right\}^{-1}, \\
\pi_1 &= \text{pr}(D = 1) = \sum_{k=1}^{K} \sum_{s=1}^{S} h(1, X_k, G_s)\gamma_k \xi_s, \\
\pi_0 &= 1 - \pi_1.
\end{aligned}$$

The retrospective likelihood is

$$\begin{aligned}
& \prod_{i=1}^{n} \text{pr}(G = G_i, X = X_i | D = D_i) \\
= & \prod_{i=1}^{n} \frac{\text{pr}(D = D_i | X = X_i, G = G_i)\text{pr}(G = G_i)\text{pr}(X = X_i)}{\text{pr}(D = D_i)} \\
= & \prod_{i=1}^{n} \frac{h(D_i, X_i, G_i)\xi_i \gamma_i}{\pi_{D_i}},
\end{aligned}$$

and the log-likelihood is

$$\mathcal{L}_2 = \sum_{i=1}^{n} \left[\log\{h(D_i, X_i, G_i)\} + \log(\xi_i) + \log(\gamma_i) \right] - n_1 \log(\pi_1) - n_0 \log(\pi_0).$$

Note that $\mathcal{L}_1$ derived by hypothetical population is proportional to $\mathcal{L}_2$, as its component $\log(n_{D_i}/n)$ does not include parameters. Hence, we conclude that the new estimator unifying the two threads of work can be obtained through optimizing $\mathcal{L}_1$ (or $\mathcal{L}_2$) or solving estimating equations (6)-(8).

3.3 Algorithm for Solving Estimating Equation

Based on our experience, optimizing the loglikelihood could be computationally unstable, owing to the high dimension of nuisance parameters (i.e., ξ's and γ's). Thus, we recommend directly solving the estimating equations, which are the score functions of the loglikelihood. We summarize the strategy in Algorithm 1.

Algorithm 1

Step 1: Set initial values. $\widetilde{\boldsymbol{\theta}}$ from logistic regression with $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$.
$\widetilde{\gamma}_k$'s and $\widetilde{\xi}_s$'s are set based on the observed frequency of X and G.
Step 2: Calculate $h(0, X_s, G_k, \widetilde{\boldsymbol{\theta}})$, and then update
$\widetilde{\pi}_0 = \sum_{s=1}^{S} \widetilde{\xi}_s \{\sum_{k=1}^{K} \widetilde{\gamma}_k h(0, X_k, G_s, \widetilde{\boldsymbol{\theta}})\}$ and
$\widetilde{\pi}_1 = \sum_{s=1}^{S} \widetilde{\xi}_s \{\sum_{k=1}^{K} \widetilde{\gamma}_k h(1, X_k, G_s, \widetilde{\boldsymbol{\theta}})\}$.
Step 3: Use equations (7) to update γ and ξ, denoted by $\widehat{\gamma}$ and $\widehat{\xi}$.
Step 4: Use equations (8) to check density estimation from the previous step.
Step 5: Update π_1 with $\widehat{\pi}_1 = \sum_{s=1}^{S} \widehat{\xi}_s \{\sum_{k=1}^{K} \widehat{\gamma}_k h(1, X_k, G_s, \widetilde{\boldsymbol{\theta}})\}$ and $\widehat{\pi}_0 = 1 - \widehat{\pi}_1$.
Step 6: With $\widehat{\pi}_1, \widehat{\pi}_0$, use equations (6) to update $\boldsymbol{\theta}$, denoted by $\widehat{\boldsymbol{\theta}}$.
Step 7: Repeat Step 2–6 until $\widehat{\boldsymbol{\theta}}$ converged.
Step 8: Calculate π_1 with updated $\widehat{\boldsymbol{\theta}}$, $\widehat{\gamma}$ and $\widehat{\xi}$. i.e., $\widehat{\pi}_1 = \sum_{s=1}^{S} \widehat{\xi}_s \{\sum_{k=1}^{K} \widehat{\gamma}_k h(1, X_k, G_s, \widehat{\boldsymbol{\theta}})\}$.

3.4 Asymptotic Results

From the derivations in Sec 3.1–3.2, $Q(\boldsymbol{\beta})$ is the score function of $\mathcal{L}_1$ or $\mathcal{L}_2$, which is the derivative of $\mathcal{L}_1$ or $\mathcal{L}_2$ with respect to $\boldsymbol{\beta}$. To establish the asymptotic properties of $\widehat{\boldsymbol{\beta}}$, we first state a list of regularity conditions:

1. The function $Q(\boldsymbol{\beta})$ is twice differentiable, and its second derivative is Lipschitz continuous.
2. The density functions of X and G, denoted by f_X and f_G, respectively, have compact support and are positive on the support.
3. The matrix $A = E\{\partial Q_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^{\top}\}$ and $B = cov\{Q_i(\boldsymbol{\beta})\}$ are non-singular and their elements are bounded away from infinity.
4. The function $h(D, X, G)$ defined in (1) is differentiable with respect to α and $\boldsymbol{\beta}$.

Under mild regularity conditions listed above, the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ can be derived based on standard estimating equation theory. Denote the estimating equation of $\boldsymbol{\beta}$ as $Q(\boldsymbol{\beta})$ such that $E\{Q(\boldsymbol{\beta})\} = 0$. $\widehat{\boldsymbol{\beta}}$ solves

$$n^{-1} \sum_{i=1}^{n} Q_i(\boldsymbol{\beta}) = 0.$$

By Taylor series,

$$n^{-1/2} \sum_{i=1}^{n} Q_i(\widehat{\boldsymbol{\beta}}) - Q_i(\boldsymbol{\beta})$$

$$
\begin{aligned}
&= n^{-1/2}\sum_{i=1}^{n}\left\{\frac{\partial Q_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})+o_p(\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|)\right\}\\
&= n^{-1}\sum_{i=1}^{n}\frac{\partial Q_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})+o_p(1).
\end{aligned}
$$

Thus,

$$
n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = -\left\{n^{-1}\sum_{i=1}^{n}\frac{\partial Q_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}^{-1} n^{-1/2}\sum_{i=1}^{n}Q_i(\boldsymbol{\beta})+o_p(1).
$$

Let $A = E\{\partial Q_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^{\top}\}$ and $B = cov\{Q_i(\boldsymbol{\beta})\}$, we have $n^{-1}\sum_{i=1}^{n}\partial Q_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^{\top} = A + o_p(1)$, and $n^{-1/2}\sum_{i=1}^{n}Q_i(\boldsymbol{\beta}) \to N(0, B)$ as $n \to \infty$. After solving the estimating equation for $\widehat{\boldsymbol{\beta}}$, both A and B can be estimated by their sample counterparts. When the sample size is limited, bootstrap is often recommended for estimating B.

3.5 Numerical Study

To show the validity of the new estimator, we adopt the simulation setting in Liang et al. (2019). With 100 replications, we generate $X \sim N(0, 1)$ and $G \sim$ Bernnoulli (0.6) independently, and then generate the disease status D as follow $\Pr(D = 1 \mid X, G) = 1/[1 + \exp\{-(\alpha + \beta_X X + \beta_G G + \beta_{XG} XG)\}]$. The case-control data is collected as follows. We simulate a random sample (X, G, D) from a sufficiently large source population. We do not stop until both the number of cases and the number of controls reach $n_1 = 100$ and $n_0 = 100$, respectively. By setting $\alpha = -3.61$ and $\boldsymbol{\beta} = (\beta_X, \beta_G, \beta_{XG})^{\mathrm{T}} = (0.76, 0.36, -0.63)^{\mathrm{T}}$, the resulting disease rate is 4.5%. We estimate the parameters using logistic regression ("Logistic"), the method proposed in Stalder et al. (2017) ("Spmle"), the method proposed in Wang and Asher (2021) ("Combo"), the method proposed in Liang et al. (2019) ("Semi"), and the new estimator we derived based on hypothetical population and retrospective likelihood ("Unified"). We compare the mean, coverage rate, and efficiency of the five methods in Table 2. We observe that the new estimator achieved a nominal coverage rate as the prospective, retrospective, and hypothetical population methods. A more detailed summary of the bias ("Bias"), sample standard error ("ESE"), asymptotic standard error ("ASE"), mean squared errors ("MSE"), coverage rate ("95%"), and the efficiency ("EFF") of the Unified Estimator is provided in Table 3, with a direct comparison to the prospective method (logistic regression). Specifically, efficiency (EFF) is the ratio of two MSE where a large value indicates higher efficiency. From Table 3, we observe substantial improvements in empirical efficiency, which is due to the gene-environment assumption as other retrospective methods.

The new estimator does not require prior knowledge of π_1, and π_1 is updated iteratively in the algorithm. We summarize the π_1 estimates in the initial step, intermediate step, and the final estimate (see Figure 1). We observe that the estimation of π_1 is fairly stable during iterations.

4 Discussion

In this paper, we summarize the recently developed semiparametric analysis methods for gene-environment interaction. Specifically, we focus on the methods based on retrospective likelihood and hypothetical population perspectives due to their efficiency improvement under only the common gene-environment independent assumption. As Han and Chatterjee (2018) pointed out, this assumption is plausible due to the fact that the genetic variation one inherited from

Table 2: Simulation results from 100 simulated case-control samples of size $n_0 = n_1 = 100$ taken from a population with a disease rate of 4.5% with $X \sim N(0, 1)$, $G \sim \text{Bern}(0.6)$.

Method	True	β_X 0.76	β_G 0.36	β_{XG} −0.63
Logistic	Mean	0.8087	0.4512	−0.6747
	95%	0.95	0.93	0.97
Spmle	Mean	0.8045	0.4496	−0.6688
	95%	0.95	0.93	0.92
	EFF	1.1723	1.0423	1.4122
Combo	Mean	0.6909	0.4136	−0.5879
	95%	0.92	0.94	0.94
	EFF	1.4318	1.0585	1.7306
Semi	Mean	0.7610	0.36	−0.6300
	95%	0.95	0.94	0.94
	EFF	1.0030	1.325	1.5660
Unified	Mean	0.6742	0.3178	−0.5957
	95%	0.96	1.00	0.93
	EFF	8.5196	2.1541	4.7292

Table 3: A more detailed comparison for the new estimator and logistic regression. Results of 100 simulations with $n_0 = n_1 = 100$. $X \sim N(0, 1)$, $G \sim \text{Bern}(0.6)$.

		β_X	β_G	β_{XG}
Logistic	Bias	−0.0511	−0.1119	0.0957
	ESE	0.3825	0.2618	0.3069
	ASE	0.7044	1.7684	0.6866
	MSE	0.1474	0.0804	0.1024
	95%	0.90	0.95	0.93
Unified	Bias	0.0858	0.0422	−0.0343
	ESE	0.1002	0.1894	0.1438
	ASE	0.1689	1.115	0.1777
	MSE	0.0173	0.0373	0.0217
	95%	0.96	1.00	0.93
	EFF	8.5196	2.1541	4.7292

parents is determined during the meiosis stage. Thus, it is not affected by subsequent environmental exposures after birth. We connect the methods based on retrospective likelihood and the notion of a hypothetical population by developing a new estimator that unifies the two important approaches. The development of the new estimator serves as a bridge to help researchers understand the methodological connection between the two threads of work. Further, as pointed

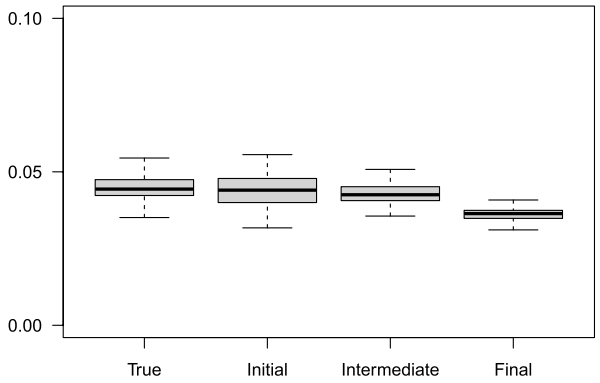

Figure 1: Results for $\hat{\pi}_1$ during the iterations of solving estimating equations in the "unified estimator". "True" is the true π_1 that generates the data. "Initial" is the initial π_1 in step 2 of the algorithm. "Intermediate" is the π_1 in step 5 of the algorithm. "Final" is the final estimate of π_1.

out by Wang and Asher (2021), the apparent efficiency improvement observed in the simulation studies does not necessarily make a substantial difference in real data discoveries, in addition to smaller p values. On the other hand, Han and Chatterjee (2018) also explained that sample size is an important factor for identifying gene-environment interactions, yet many studies do not have adequate measurements to identify the interaction effect of modest magnitude. Popular tools such as `CGEN` (Bhattacharjee et al., 2023) R packages provide various functions to test gene-environment interactions, requiring both gene and environmental exposure to be coded into three categories that have limited applications. Thus, we focus on showing the validity of the unified estimator through a more standard simulation study from Liang et al. (2019).

Possible future work includes justifying the efficiency gain theoretically and optimizing the algorithm to improve computational efficiency. Methodology-wise speaking, the new estimator is general and flexible, which could be obtained for unspecified multivariate genetic and environmental factors. However, the large number of nuisance parameters makes it computationally unstable. One possible solution to overcome the unstable computation problem is to consider the profile likelihood of α obtained as $L(\alpha, \hat{\beta}(\alpha), \hat{\xi}(\alpha), \hat{\gamma}(\alpha))$, and perform a grid search for α.

Supplementary Material

The code that implements Algorithm 1 in Section 3.3 is provided in the Supplementary Materials

Acknowledgement

The authors are grateful to the guest editor, and two anonymous referees for their constructive comments that helped improve the paper.

References

Bhattacharjee S, Chatterjee N, Han S, Song M, Wheeler W, de Rochemonteix M, et al. (2023). CGEN: An R package for analysis of case-control studies in genetic epidemiology. R package version 3.36.1.

Bickel PJ, Klassen CAJ, Ritov Y, Wellner JA (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Breslow NE, Robins JM, Wellner JA (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli*, 6: 447–55. https://doi.org/10.2307/3318670

Chatterjee N, Carroll RJ (2005). Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. *Biometrika*, 92: 399–418. https://doi.org/10.1093/biomet/92.2.399

Chatterjee N, Kalaylioglu Z, Carroll RJ (2005). A new paradigm of conditional-likelihoods for exploiting gene-environment independence in family based case-control studies. *Genetic Epidemiology*, 28: 138–156. https://doi.org/10.1002/gepi.20049

Chen YH, Chatterjee N, Carroll RJ (2008). Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association. *Biostatistics*, 9: 81–99. https://doi.org/10.1093/biostatistics/kxm011

Chen YH, Chatterjee N, Carroll RJ (2009). Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association*, 104: 220–233. https://doi.org/10.1198/jasa.2009.0104

Crossa J (2012). From genotype x environment interaction to gene x environment interaction. *Current Genomics*, 13(3): 225–244. https://doi.org/10.2174/138920212800543066

Crouch DJ, Bodmer WF (2020). Polygenic inheritance, gwas, polygenic risk scores, and the search for functional variants. *Proceedings of the National Academy of Sciences*, 117(32): 18924–18933. https://doi.org/10.1073/pnas.2005634117

Curtis D (2018). Polygenic risk score for schizophrenia is not strongly associated with the expression of specific genes or gene sets. *Psychiatric Genetics*, 28(4): 59–65. https://doi.org/10.1097/YPG.0000000000000197

Emdin CA, Haas M, Ajmera V, Simon TG, Homburger J, Neben C, et al. (2021). Association of genetic variation with cirrhosis: A multi-trait genome-wide association and gene–environment interaction study. *Gastroenterology*, 160(5): 1620–1633. https://doi.org/10.1053/j.gastro.2020.12.011

Gauderman WJ, Mukherjee B, Aschard H, Hsu L, Lewinger JP, Patel CJ, et al. (2017). Update on the state of the science for analytical methods for gene-environment interactions. *American Journal of Epidemiology*, 186(7): 762–770. https://doi.org/10.1093/aje/kwx228

Gauderman WJ, Zhang P, Morrison JL, Lewinger JP (2013). Finding novel genes by testing g× e interactions in a genome-wide association study. *Genetic Epidemiology*, 37(6): 603–613. https://doi.org/10.1002/gepi.21748

Han SS, Chatterjee N (2018). Review of statistical methods for gene-environment interaction analysis. *Current Epidemiology Reports*, 5: 39–45. https://doi.org/10.1007/s40471-018-0135-2

Han SS, Rosenberg PS, Garcia-Closas M, Figueroa JD, Silverman D, Chanock SJ, et al. (2012). Likelihood ratio test for detecting gene (g)-environment (e) interactions under an additive risk model exploiting ge independence for case-control data. *American Journal of Epidemiology*, 176: 1060–1067. https://doi.org/10.1093/aje/kws166

Hunter DJ (2005). Gene–environment interactions in human diseases. *Nature Reviews. Genetics*, 6(4): 287–298. https://doi.org/10.1038/nrg1578

Hutter CM, Mechanic LE, Chatterjee N, Kraft P, Gillanders EM, Tank NGET (2013). Gene-environment interactions in cancer epidemiology: A national cancer institute think tank report. *Genetic Epidemiology*, 37(7): 643–657. https://doi.org/10.1002/gepi.21756

Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. (2018). Genome-wide

polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9): 1219–1224. https://doi.org/10.1038/s41588-018-0183-z

Liang L, Ma Y, Carroll RJ (2019). A semiparametric efficient estimator in case-control studies for gene–environment independent models. *Journal of Multivariate Analysis*, 173: 38–50. https://doi.org/10.1016/j.jmva.2019.01.006

Lin DY, Zeng D (2006). Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association*, 101(473): 89–104. https://doi.org/10.1198/016214505000000808

Liu J, Ma Y (2019). Locally efficient semiparametric estimators for a class of Poisson models with measurement error. *Canadian Journal of Statistics*, 47(2): 157–181. https://doi.org/10.1002/cjs.11483

Lobach I, Carroll RJ, Spinka C, Gail MH, Chatterjee N (2008). Haplotype-based regression analysis of case-control studies with unphased genotypes and measurement errors in environmental exposures. *Biometrics*, 64: 673–684. https://doi.org/10.1111/j.1541-0420.2007.00930.x

Luo S, Mukherjee B, Chen J, Chatterjee N (2009). Shrinkage estimation for robust and efficient screening of single-SNP association from case-control genome-wide association studies. *Genetic Epidemiology*, 33: 740–750. https://doi.org/10.1002/gepi.20428

Ma Y (2010). A semiparametric efficient estimator in case-control studies. *Bernoulli*, 16: 585–603.

Meisner A, Kundu P, Chatterjee N (2019). Case-only analysis of gene-environment interactions using polygenic risk scores. *American Journal of Epidemiology*, 188(11): 2013–2020. https://doi.org/10.1093/aje/kwz175

Modan B, Hartge P, Hirsh-Yechezkel G, Chetrit A, Lubin F, Beller U, et al. (2001). Parity, oral contraceptives, and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *The New England Journal of Medicine*, 345: 235–240. https://doi.org/10.1056/NEJM200107263450401

Mukherjee B, Ahn J, Gruber SB, Chatterjee N (2012). Testing gene-environment interaction in large-scale case-control association studies: Possible choices and comparisons. *American Journal of Epidemiology*, 175: 177–190. https://doi.org/10.1093/aje/kwr367

Murcray CE, Lewinger JP, Gauderman WJ (2009). Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology*, 169(2): 219–226. https://doi.org/10.1093/aje/kwn353

Piegorsch WW, Weinberg CR, Taylor JA (1994a). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*, 13(2): 153–162. https://doi.org/10.1002/sim.4780130206

Piegorsch WW, Weinberg CR, Taylor JA (1994b). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Statistics in Medicine*, 13: 153–162. https://doi.org/10.1002/sim.4780130206

Prentice RL, Pyke R (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66: 403–411. https://doi.org/10.1093/biomet/66.3.403

Spinka C, Carroll RJ, Chatterjee N (2005). Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology*, 29: 108–127. https://doi.org/10.1002/gepi.20085

Stalder O, Asher A, Liang L, Carroll RJ, Ma Y, Chatterjee N (2017). Semiparametric analysis of complex polygenic gene-environment interactions in case-control studies. *Biometrika*, 104(4): 801–812. https://doi.org/10.1093/biomet/asx045

Thomas D (2010). Methods for investigating gene-environment interactions in candidate

pathway and genome-wide association studies. *Annual Review of Public Health*, 31: 21. https://doi.org/10.1146/annurev.publhealth.012809.103619

Tsiatis AA (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

Tsiatis AA, Ma Y (2004). Locally efficient semiparametric estimators for functional measurement error models. *Biometrika*, 91: 835–848. https://doi.org/10.1093/biomet/91.4.835

Wang T, Asher A (2021). Improved semiparametric analysis of polygenic gene–environment interactions in case–control studies. *Statistics in Biosciences*, 13(3): 386–401. https://doi.org/10.1007/s12561-020-09298-9