SEMIPARAMETRIC ANALYSIS OF COMPLEX POLYGENIC GENE-ENVIRONMENT INTERACTIONS IN CASE-CONTROL STUDIES

A Dissertation

by

ALEXANDER ALLEN ASHER

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Raymond J. Carroll
Committee Members,	Jeffrey D. Hart
	Edward R. Jones
	Jan S. Suchodolski
Head of Department,	Valen E. Johnson

August 2018

Major Subject: Statistics

Copyright 2018 Alexander Allen Asher

ABSTRACT

Gene-environment interactions can be efficiently estimated in case-control data by existing retrospective methods that assume gene-environment independence in the source population, but such techniques require parametric modeling of the genetic variables. Standard logistic regression analysis of case-control data has low power to detect gene-environment interactions, but it has been the only method capable of analyzing complex polygenic data for which parametric distributional models are not feasible.

This dissertation proposes a general, computationally simple, semiparametric method for analysis of case-control studies that allows exploitation of the assumption of gene-environment independence without any further parametric modeling assumptions about the marginal distributions of any of the two sets of factors. The method relies on the key observation that an underlying efficient profile likelihood depends on the distribution of genetic factors only through certain expectation terms that can be evaluated empirically.

This method is further improved by treating the genetic and environmental variables symmetrically to generate two sets of parameter estimates that are combined to generate a more efficient estimate. A semiparametric framework is employed to develop the asymptotic theory of the estimators, and their performance is evaluated via simulation studies. The methods are illustrated using data from a case-control study of breast cancer, and free software implementing both methods is demonstrated.

DEDICATION

To the memory of Dr. Sue Allen Warren.

ACKNOWLEDGMENTS

There are no words to express my gratitude to my wife, who put her career on hold to switch continents and raise our children as a single parent while I pursued my PhD. She is the axis around which our family revolves, and her constant encouragement and moral support made everything possible. There is, similarly, no way to sufficiently thank my children, who took everything in stride. They are my motivation, and I am proud to be their *doctor de números*.

I also thank my parents and sister, who were with me every step of the way, for the unwavering support they gave me, my wife, and our children. They inspired me to embark on this project, and they have done everything possible to help me succeed. Thanks also go to my mother-in-law, father-in-law, sister-in-law, and brother-in-law, who can always be counted on to lend a hand, and whose collaboration makes our family thrive.

I owe a debt of gratitude to my advisor, Raymond Carroll, for his guidance and encouragement. Ray is widely acclaimed for his fundamental contributions to the field of statistics, but he deserves equal recognition for his ability to inspire and mentor budding statisticians, and challenge them to accomplish more than they thought possible.

My committee members deserve extra thanks for all their efforts on my behalf. Jeff Hart has always been extremely generous with his time, and he has been my go-to expert on a wide range of topics since my days as a distance MS student. Ed Jones has been a trusted mentor since my first days on campus, and he has taught me skills invaluable to my career as a professional statistician. Jan Suchodolski has spent untold hours discussing statistical analysis with me, and he has taught me volumes about how to effectively communicate my message to a wider audience.

I would also like to thank my colleague Tianying Wang, whose tireless work ethic, meticulous attention to detail, sharp mind, and cheerful disposition make it a joy to work with her.

Special thanks to Sabrina Barahona, Nilanjan Chatterjee, Eva Garcia, Valen Johnson, Eli Kravitz, Liang Liang, Yanyuan Ma, Yabo Niu, Riddhi Pratim Ghosh, and Odile Stalder. Additionally, countless friends, colleagues, and faculty contributed to this research by sharing their time and wisdom. There is no room to name everyone, but I thank you all.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Distinguished Professor Raymond J. Carroll and Professors Jeffrey D. Hart and Edward R. Jones of the Department of Statistics and Professor Jan S. Suchodolski of the Department of Veterinary Medicine & Biomedical Sciences.

Section 2 of this dissertation was published in Biometrika as Stalder et al. (2017) and is reprinted with permission. Coauthors of this section are: Odile Stalder (Department of Clinical Research and Institute of Social and Preventive Medicine, University of Bern, Switzerland), Liang Liang (Biostatistics Department, Harvard University, Cambridge, Massachusetts), Raymond J. Carroll (Department of Statistics, Texas A&M University), Yanyuan Ma (Department of Statistics, Penn State University, University Park, Pennsylvania), and Nilanjan Chatterjee (Department of Biostatistics, Bloomberg School of Public Health, and Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, Maryland).

Section 3 of this dissertation was coauthored with Tianying Wang and Raymond J. Carroll, both of the Department of Statistics, Texas A&M University.

The data analyzed for Sections 2 and 3 were provided by the National Cancer Institute and are available via a data transfer agreement.

All other work conducted for this dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by the Department of Statistics and a grant from the National Cancer Institute (U01-CA057030).

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Recent genome-wide association studies indicate that the genetic predisposition for diseases such as cancer and diabetes involves hundreds, if not thousands, of genetic variants (Chatterjee et al., 2016; Fuchsberger et al., 2016). In order to understand disease mechanisms and develop strategies for disease prevention, it is critical to know how these genetic factors interact with environmental risk factors.

Case-control studies are retrospective observational studies in which the sample consists of a group of healthy subjects and a group of diseased subjects. A crucial aspect of the case-control design is that the outcome, disease status, is known *before* sampling. The ability to deliberately oversample diseased subjects makes the case-control design cost effective, which is why it is widely popular in studies of gene-environment interactions.

Existing methods of estimating gene-environment interactions in case-control data are either inefficient or not flexible enough to handle more than a few genetic variants. The focus of this dissertation is to provide flexible methodology, theory, and software for the efficient estimation of polygenic gene-environment interactions in case-control data.

2. SEMIPARAMETRIC ANALYSIS OF COMPLEX POLYGENIC GENE–ENVIRONMENT INTERACTIONS IN CASE–CONTROL STUDIES*

Many methods have been recently proposed for efficient analysis of case–control studies of gene-environment interactions using a retrospective likelihood framework that exploits the natural assumption of gene-environment independence in the underlying population. However, for polygenic modeling of gene-environment interactions, a topic of increasing scientific interest, applications of retrospective methods have been limited due to a requirement in the literature for parametric modeling of the distribution of the genetic factors. We propose a general, computationally simple, semiparametric method for analysis of case–control studies that allows exploitation of the assumption of gene–environment independence without any further parametric modeling assumptions about the marginal distributions of any of the two sets of factors. The method relies on the key observation that an underlying efficient profile likelihood depends on the distribution of genetic factors only through certain expectation terms that can be evaluated empirically. We develop asymptotic inferential theory for the estimator and evaluate numerical performance using simulation studies. An application of the method is presented.

2.1 Introduction

Recent genome-wide association studies indicate that complex diseases, such as cancers, diabetes and heart diseases, are in general extremely polygenic (Chatterjee et al., 2016; Fuchsberger et al., 2016). Genetic predisposition to a single disease may involve thousands of genetic variants, each of which may have a very small effect individually, but in combination they can explain substantial variation in risk in the underlying population. As discoveries from genome-wide association studies continue to enhance understanding of complex diseases, in the future, it will be critical to understand how these genetic factors interact with environmental risk factors for both

*Reprinted with permission from "Semiparametric analysis of complex polygenic gene-environment interactions in case-control studies" by Stalder, O., Asher, A., Liang, L., Carroll, R. J., Ma, Y., and Chatterjee, N., 2017. *Biometrika*, 104, 801-812, Copyright 2017 by Oxford University Press.

understanding disease mechanisms and developing public health strategies for disease prevention.

Because of its sampling efficiency, the case–control design is widely popular for conducting studies of genetic associations and gene–environment interactions. A variety of analytic methods have been proposed to increase the efficiency of analysis of case–control data for studies of gene–environment interactions by exploiting an assumption of gene–environment independence in the underlying population. It has been shown that under the assumptions of gene–environment independence and rare disease, the interaction odds-ratio parameters of a logistic regression model can efficiently be estimated based on cases alone (Piegorsch et al., 1994). A general logistic regression model can be fit to case–control data under the gene–environment independence assumption using a log-linear modeling framework (Umbach and Weinberg, 1997) or a semiparametric retrospective profile likelihood framework (Chatterjee and Carroll, 2005). More recently, the assumption of gene–environment independence has been exploited to propose a variety of powerful hypothesis testing methods for conducting genome-wide scans of gene–environment interactions (Gauderman et al., 2013; Han et al., 2015; Hsu et al., 2012; Mukherjee et al., 2012; Mukherjee and Chatterjee, 2008; Murcray et al., 2009).

We consider developing methods for efficient analysis of case–control studies for modeling gene–environment interactions involving multiple genetic variants simultaneously. To develop parsimonious models for joint effects, many studies have recently focused on developing models for gene–environment interactions using underlying polygenic risk scores that could be defined by all known genetic variants associated with the diseases (Chatterjee et al., 2016, 2013; Dudbridge, 2013; Meigs et al., 2008; Wacholder et al., 2010). Further, for obtaining improved biological insights and for enhancing statistical power for detection, it may often be desired to model gene–environment interactions using multiple variants within genomic regions or/and biologic pathways (Chatterjee et al., 2006; Jiao et al., 2013; Lin et al., 2013, 2015). In standard prospective logistic regression analysis, which conditions on both the genetic and environmental risk factor status of the individuals, handling multiple genetic variants is relatively straightforward. In contrast, with "retrospective" methods, which aim to exploit the assumption of gene–environment independence,

the task becomes complicated because all currently existing methods require parametric modeling of the distribution of the genetic or environmental variables.

We propose computationally simple methodology for fitting general logistic regression models to case–control data under the assumption of gene–environment independence, but without requiring any further modeling assumptions about the distributions of the genetic or environmental variables. We extend the Chatterjee–Carroll profile likelihood framework, which originally considered modeling gene–environment interactions using single genetic variants for which genotype status could be specified using parametric multinomial models. The new method relies on the observation that the profile likelihood itself can be estimated based on an empirical genotype distribution that is estimable from a case–control sample. We develop the asymptotic theory of the resulting estimator under a semiparametric inferential framework. Simulations and an example illustrate the properties of the new methodology.

2.2 Model, Method and Theory

2.2.1 Background, model and method

In the following, we use notation similar to that of Chatterjee and Carroll (2005). We will denote disease status, genetic information and environmental risk factors by D, G and X, respectively. Here G may correspond to a complex multivariate genotype associated with multiple genetic variants or a continuous polygenic risk score that is defined a priori based on known associations of the genetic variants with the disease. We assume the risk of the disease given genetic and environmental factors in the underlying population can be specified using a model of the form

$$\text{pr}(D = 1 \mid G, X) = H\{\alpha_0 + m(G, X, \beta)\}, \tag{2.1}$$

where $H(x) = \{1 + \exp(-x)\}^{-1}$ is the logistic distribution function and $m(G, X, \beta)$ is a parametrically specified function that defines a model for the joint effect of G and X on the logistic-risk scale. The goal of the gene–environment interaction study is to make inference on the parameters β in (2.1), including interaction parameters.

Let $F(G, X)$ denote the joint distribution of G and X in the underlying population. The key assumption that genetic, G, and environmental factors, X, are independently distributed in the underlying population can be mathematically stated as

$$dF(G, X) = dF_G(G) \times dF_X(X),$$

where F_G and F_X denote the underlying marginal distributions of G and X, respectively. In the *Supplementary Material* we discuss how to weaken this assumption by suitable conditioning on additional stratification factors. In contrast to the existing literature, here we assume that the marginal distributions $F_G(G)$ and $F_X(X)$ are both completely unspecified.

We consider a population-based case–control study, in which (G, X) are sampled independently from those with the disease, called cases, and those without the disease, called controls. Suppose there are n_1 cases and n_0 controls. Standard prospective logistic regression analysis, which is equivalent to maximum likelihood estimation when $F(G, X)$ is allowed to be completely unspecified, yields consistent estimates of β (Prentice and Pyke, 1979).

The retrospective likelihood is the probability of observing the genetic and environmental variables, given the subject's disease status. Under gene–environment independence in the underlying population, the retrospective likelihood is

$$\text{pr}(G = g, X = x|D = d) = \text{pr}(D = d|G = g, X = x)\text{pr}(G = g)\text{pr}(X = x)/\text{pr}(D = d).$$

Let $f_G(\cdot)$ and $f_X(\cdot)$ represent the density/mass functions of G and X, respectively. The retrospective likelihood is

$$\frac{f_G(g)f_X(x)\exp[d\{\alpha_0 + m(g, x, \beta)\}]/[1 + \exp\{\alpha_0 + m(g, x, \beta)\}]}{\int f_G(u)f_X(v)\exp[d\{\alpha_0 + m(u, v, \beta)\}]/[1 + \exp\{\alpha_0 + m(u, v, \beta)\}]dudv}. \tag{2.2}$$

Chatterjee and Carroll (2005) profiled out $f_X(\cdot)$ by treating it as discrete on the set of distinct observed values $(x_1, \ldots, x_m)$ of X with probabilities $\delta_i = \text{pr}(X = x_i)$, and then maximized (2.2)

over $(\delta_1, \ldots, \delta_m)$, leading eventually to the semiparametric profile likelihood described as follows. Define $\kappa = \alpha_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$, where $\pi_1 = 1 - \pi_0 = \text{pr}(D = 1)$ is defined as the probability of the disease in the underlying population. Define $\Omega = (\kappa, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$. Also define

$$S(d, g, x, \Omega) = \frac{\exp[d\{\kappa + m(g, x, \beta)\}]}{1 + \exp\{\kappa + \log(\pi_1/\pi_0) - \log(n_1/n_0) + m(g, x, \beta)\}}.$$

Then, with this notation, the semiparametric profile likelihood is

$$L(D, G, X, \Omega, f_G) = f_G(G) \frac{S(D, G, X, \Omega)}{\sum_{d=0}^{1} \int f_G(v) S(d, v, X, \Omega) dv}. \tag{2.3}$$

While the representation in (2.3) does not involve the unknown density of X, it does involve the unknown density of G. This is a major reason that the current literature specifies a parametric distribution for G. Our task in this paper is to dispense with the need to give a parametric form for the distribution function of G, so that analysis can be performed with respect to potentially complex multivariate genotype data for which parametric modeling can be difficult and cumbersome.

Here is our key insight, which we discuss first in the context that π_1 is known or at least can be estimated well. For case–control studies that are conducted within well defined populations, relevant probabilities of the disease can be ascertained based on population-based disease registries. When case–control studies are conducted by sampling of subjects within a larger cohort study, the probability of the disease in the underlying population can be estimated using the disease incidence rate observed in the cohort.

Our key insight in treating the distribution of G as nonparametric concerns the term in the denominator of (2.3), defined as

$$R(x, \Omega) = \sum_{r=0}^{1} \int f_G(v) S(r, v, x, \Omega) dv.$$

This is simply the expectation, in the source population, of $\sum_{r=0}^{1} S(r, G, x, \Omega)$. That is, $R(x, \Omega) = E_{\text{pop}}\{\sum_{r=0}^{1} S(r, G, x, \Omega)\}$, where the subscript pop emphasizes that the expectation is in the source

population, not the case–control study. However, crucially,

$$R(x,\Omega) = \pi_1 E\{\textstyle\sum_{r=0}^{1} S(r,G,x,\Omega) \mid D=1\} + \pi_0 E\{\textstyle\sum_{r=0}^{1} S(r,G,x,\Omega) \mid D=0\}. \quad (2.4)$$

Of course, $R(x,\Omega)$ is unknown, but we estimate it unbiasedly and nonparametrically by

$$\widehat{R}(x,\Omega) = \textstyle\sum_{j=1}^{J}\sum_{r=0}^{1}\sum_{d=0}^{1}(\pi_d/n_d) I(D_j=d) S(r,G_j,x,\Omega). \quad (2.5)$$

In the *Supplementary Material*, we show that $\widehat{R}(x,\Omega)$ is an unbiased estimate of $R(x,\Omega)$, that is $n^{1/2}$-consistent, and that it is asymptotically normally distributed.

Ignoring the leading term $f_G(G)$ in (2.3), which is not estimated, and taking logarithms, leads us to an estimated loglikelihood in Ω across the data as

$$\mathcal{L}(\Omega) = \textstyle\sum_{i=1}^{n} \log S(D_i,G_i,X_i,\Omega) - \sum_{i=1}^{n} \log \widehat{R}(X_i,\Omega). \quad (2.6)$$

Define $S_\Omega(d,g,x,\Omega) = \partial S(d,g,x,\Omega)/\partial\Omega$ and similarly for $\widehat{R}_\Omega(x,\Omega)$. Then the estimated score function, a type of estimated estimating equation, is

$$\widehat{\mathcal{S}}_n(\Omega) = n^{-1/2} \sum_{i=1}^{n} \left\{ \frac{S_\Omega(D_i,G_i,X_i,\Omega)}{S(D_i,G_i,X_i,\Omega)} - \frac{\widehat{R}_\Omega(X_i,\Omega)}{\widehat{R}(X_i,\Omega)} \right\}. \quad (2.7)$$

Define

$$\mathcal{S}_n(\Omega) = n^{-1/2} \sum_{i=1}^{n} \left\{ \frac{S_\Omega(D_i,G_i,X_i,\Omega)}{S(D_i,G_i,X_i,\Omega)} - \frac{R_\Omega(X_i,\Omega)}{R(X_i,\Omega)} \right\},$$

which is the profile loglikelihood score function when the distribution of G is known. Since the profile loglikelihood score of Chatterjee and Carroll (2005) would have mean zero if the distribution of G were known, it follows that

$$E\{\mathcal{S}_n(\Omega)\} = 0, \quad (2.8)$$

where the expectation in (2.8) is taken in the case–control study, not in the source population. Thus, since $\widehat{R}(x,\Omega)$ and $\widehat{R}_\Omega(x,\Omega)$ converge in probability to $R(x,\Omega)$ and $R_\Omega(x,\Omega)$, respectively, a consistent estimate of Ω can be obtained by solving $\widehat{\mathcal{S}}_n(\Omega)=0$. This estimate $\widehat{\Omega}$, which maximizes the semiparametric pseudolikelihood (2.6), will be referred to as the semiparametric pseudolikelihood estimator.

2.2.2 Rare diseases when π_1 is unknown

When the probability of disease in the source population is unknown, one can invoke a rare disease assumption which is often reasonable for case–control studies (Epstein and Satten, 2003; Kwee et al., 2007; Lin and Zeng, 2006; Modan et al., 2001; Piegorsch et al., 1994; Zhao et al., 2003). If we assume that $\pi_1 \approx 0$, then $S(d,g,x,\Omega) \approx \exp[d\{\kappa + m(g,x,\beta\}]$, and the expectation involved in calculation of $R(X,\Omega)$ can be evaluated based on the sample of controls, with $D=0$, only. In this case, the estimates of Ω converge not to Ω itself, but instead to Ω_*, the solution to (2.8) with $\pi_1 = 0$. Typically, except when the sample size is very large and hence standard errors are unusually small, the small possible bias of the rare disease approximation is of little consequence and coverage probabilities of confidence intervals remain near nominal, see §2.3 for examples. The asymptotic theory of §2.2.3 below is then unchanged.

In the *Supplementary Material*, we show that the score and the Hessian take on simple forms in this case, and that the Hessian is negative semidefinite. Computation is thus very efficient.

2.2.3 Asymptotic theory

To state the asymptotic results, we first make the definitions

$$\begin{aligned}
\Gamma_1 &= \textstyle\sum_{d=0}^{1}(n_d/n)E\left\{\dfrac{\partial S_\Omega(D,G,X,\Omega)/S(D,G,X,\Omega)}{\partial\Omega^{\mathrm{T}}}\middle| D=d\right\};\\
\Gamma_2 &= \textstyle\sum_{d=0}^{1}(n_d/n)E\left\{\dfrac{\partial R_\Omega(X,\Omega)/R(X,\Omega)}{\partial\Omega^{\mathrm{T}}}\middle| D=d\right\}.
\end{aligned}$$

In addition, define $c_d = n_d/n$, $Z_i = (D_i,G_i,X_i)$, $P_1(X_i,\Omega) = 1/R(X_i,\Omega)$ and $P_2(X_i,\Omega) = R_\Omega(X_i,\Omega)/R^2(X_i,\Omega)$.

We use the notational convention that for arbitrary functions (P, T), $T_E(r, d, x) = E\{T(r, G, x) \mid D = d\}$. Also, we use the convention that

$$\begin{aligned} &E\left[P(X)\left\{T(r, g_i, X) - T_E(r, d, X)\right\} \mid D = t\right] \\ &\quad = E\left[P(X)\left\{T(r, g, X) - T_E(r, d, X)\right\} \mid D = t\right]_{g=G_i}. \end{aligned}$$

Define

$$\begin{aligned} \zeta(Z_i, \Omega) &= \frac{S_\Omega(Z_i, \Omega)}{S(Z_i, \Omega)} - \frac{R_\Omega(X_i, \Omega)}{R(X_i, \Omega)} \\ &\quad - \sum_{d=0}^{1}\sum_{r=0}^{1} \frac{c_d \pi_{d_i}}{c_{d_i}} E\left[\{P_1(X, \Omega) S_\Omega(r, g_i, X) - P_2(X, \Omega) S(r, g_i, X)\} \mid D = d\right]. \end{aligned}$$

Finally, define $\zeta_*(Z_i, \Omega) = \zeta(Z_i, \Omega) - E\{\zeta(Z, \Omega) \mid D = D_i\}$.

Theorem 1. *Suppose $n_d/n \to c_d$, where $0 < c_d < \infty$, and that π_1 is known. Then*

$$n^{1/2}(\widehat{\Omega} - \Omega) = -(\Gamma_1 - \Gamma_2)^{-1} n^{-1/2} \textstyle\sum_{i=1}^{n} \zeta_*(Z_i, \Omega) + o_p(1). \tag{2.9}$$

Thus, since the Z_i are independent and $E\{\zeta_*(Z, \Omega) \mid D_i\} = 0$, as $n \to \infty$, in distribution,

$$\begin{aligned} n^{1/2}(\widehat{\Omega} - \Omega) &\to \textit{Normal}\left[0, (\Gamma_1 - \Gamma_2)^{-1} \Sigma \{(\Gamma_1 - \Gamma_2)^{-1}\}^{\mathrm{T}}\right]; \\ \Sigma &= \textstyle\sum_{d=0}^{1} (n_d/n) \mathrm{cov}\{\zeta_*(D, X, G, \Omega) \mid D = d\} \\ &= \textstyle\sum_{d=0}^{1} (n_d/n) \mathrm{cov}\{\zeta(D, X, G, \Omega) \mid D = d\}. \end{aligned}$$

In §2.2.2, when π_1 is unknown and the disease is relatively rare, the same result holds by setting $\pi_1 = 0$.

2.3 Simulations

2.3.1 Overview

In our simulations, $m(G, X, \beta) = G^{\mathrm{T}}\beta_G + X\beta_X + (GX)^{\mathrm{T}}\beta_{GX}$ and the value of X is binary with population frequency 0.5. There are either three or five correlated single nucleotide polymorphisms within a region: we report the latter case, but the results for the former are similar. Each single nucleotide polymorphism takes on the values 0, 1 or 2 following a trinomial distribution that follows Hardy–Weinberg equilibrium, i.e., the jth component of G equals 0, 1, 2 with probabilities $\{(1-p_j)^2, 2p_j(1-p_j), p_j^2\}$. The values of the p_j are described below.

To generate correlation among the single nucleotide polymorphisms, we first generated a 3 or 5-variate multivariate normal variate, each with mean 0 and standard deviation 1, and a correlation matrix with correlation between the jth and kth component $= \rho^{|j-k|}$, where $\rho = 0.7$. After generating these random variables, we trichotomized them with appropriate thresholds so that frequency of 0, 1 and 2 matched those specified by the allele frequency p_j and Hardy–Weinberg equilibrium.

Table 2.1: Results of 1000 simulations as described in §2.3, with mean bias, coverage probabilities of a 95% nominal confidence interval, and mean squared error efficiency of our semiparametric pseudolikelihood estimator compared to ordinary logistic regression. The simulations were performed with 1000 cases and 1000 controls.

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
True	0.18	0.18	0.00	0.18	0.00	0.41	0.26	0.00	0.00	0.26	0.00
					Logistic: 1000 cases						
Bias	0.00	0.01	0.00	0.01	−0.01	0.01	0.01	−0.01	0.00	0.00	0.01
CI (%)	94.3	95.2	95.7	95.1	94.7	94.6	94.9	94.2	94.5	96.0	94.2
					SPMLE, Rare: 1000 cases						
Bias	0.01	0.00	0.00	0.02	−0.01	0.02	−0.02	−0.01	0.01	−0.02	0.01
CI (%)	95.2	95.4	96.4	95.8	95.3	95.1	95.4	94.8	96.1	95.5	94.9
Avg MSE Eff		All G: 1.28			All X: 1.26				All $G*X$: 2.18		
					SPMLE, π_1 known: 1000 cases						
Bias	0.00	0.00	0.00	0.01	−0.01	0.01	0.00	−0.01	0.01	−0.01	0.01
CI (%)	95.1	95.5	96.4	95.8	95.0	95.5	95.6	94.6	95.9	95.2	94.5
Avg MSE Eff		All G: 1.28			All X: 1.28				All $G*X$: 2.07		

Logistic is ordinary logistic regression; ***SPMLE, Rare*** is our estimator using the rare disease approximation with unknown π_1 (§2.2.2); ***SPMLE, π_1 known*** is our estimator when π_1 is known in the source population (§2.2.1); ***CI (%)*** is the coverage in percent of a nominal 95% confidence interval (calculated using the asymptotic standard error); ***Avg MSE Eff*** is the mean squared error efficiency of our method compared to logistic regression averaged over G (All G), over X (All X) and over the $G*X$ (All $G*X$) interactions.

In both simulations, the logistic intercept α_0 was chosen so that the population disease rate $\pi_1 = 0.03$. However additional simulations with $\pi_1 = 0.01$ yielded very similar results with regards to coverage, efficiency gains, and unbiasedness. See also §2.3.3 for a discussion of additional simulations, and the *Supplementary Material*. In the simulation reported here, $(p_1, p_2, p_3, p_4, p_5) = (0.1, 0.3, 0.3, 0.3, 0.1)$, $\beta_X = \log(1.5)$, $\beta_G = \{\log(1.2), \log(1.2), 0.0, \log(1.2), 0.0\}$, and $\beta_{GX} = \{\log(1.3), 0.0, 0.0, \log(1.3), 0.0\}$. Here the value of $\alpha_0 = -4.14$.

2.3.2 Results

The standard error estimators used in our simulation were based on the asymptotic theory described in Theorem 1: we also used the bootstrap, with very similar results. The appropriate

bootstrap in a case–control study is to resample the cases and controls separately, thus maintaining the sample sizes for each.

The simulation results are presented in Table 2.1. Our semiparametric pseudolikelihood estimator has little bias and coverage percentages near the nominal level. Both with a rare disease approximation and with π_1 known, our semiparametric pseudolikelihood estimator achieves approximately a 25% increase in mean squared error efficiency over ordinary logistic regression for the main effects in both G and X.

Strikingly, the mean squared error efficiency of our semiparametric pseudolikelihood estimators compared to ordinary logistic regression is approximately 2.00 for all the interaction terms, thus demonstrating that our methods, which do not model the distribution of either G or X, achieve numerically significant increases in efficiency.

2.3.3 Additional simulations

The *Supplementary Material* presents a series of additional simulations. These include the results of a simulation to evaluate the robustness of our method to misspecification of the population disease rate, where we found a surprising robustness to disease rate misspecification. Additionally, there are simulations to examine the robustness of our method to violations of the gene–environment independence assumption. Those simulation studies show that there will be bias in the estimate of gene–environment interaction parameters for the specific single nucleotide polymorphisms under violation of gene–environment independence, but average mean square error for parameter estimates across all the different single nucleotide polymorphisms could be still substantially lower than that obtained from prospective logistic regression analysis. We also show there how to remove this bias when G and E are independent conditional on a discrete stratification variable. Mukherjee and Chatterjee (2008) and Chen et al. (2009) show how to use empirical-Bayes methods to provide additional robustness to violations of the gene–environment independence assumption.

2.4 Data Analysis

In this section, we apply our methodology to a case–control study for breast cancer arising from a large prospective cohort at the National Cancer Institute: the Prostate, Lung, Colorectal and Ovarian cancer screening trial (Canzian et al., 2010). The design of this study is described in detail by Prorok et al. (2000) and Hayes et al. (2000). The cohort data consisted of $622,449$ women, of whom 3.56% developed breast cancer (Pfeiffer et al., 2013). The case–control study analyzed here consists of 753 controls and 658 cases. Although π_1 is known in this population, we analyze the data both with π_1 known and with π_1 unknown but with a rare disease approximation.

We had data available on genotypes for 21 single nucleotide polymorphisms that have been previously associated with breast cancer based on large genome-wide association studies. The polygenic risk score was defined by a weighted combination of the genotypes, with the weights defined by log-odds-ratio coefficients reported in prior studies. We examined the interaction of the polygenic risk score with age at menarche (X), a known risk factor for breast cancer, defined as the binary indicator of whether the age at menarche exceeds 13 or not. We also adjust the model for age as a continuous variable, denoted here as Z, so that the model fitted is

$$\text{pr}(D=1) \;\; = \;\; H(\beta_0 + \beta_G G + \beta_X X + \beta_{GX} GX + \beta_Z Z). \tag{2.10}$$

Results in which age was categorized as < 35, 35-40, 40-45,...,> 75 were similar.

We also performed analyses to check the gene–environment independence assumption. Since X is binary, we ran a t-test of the polygenic risk score against the levels of X, of course among the controls only. The p-value was 0.91, indicating almost no genetic effect. We also ran chisquared tests for the 21 individual genes, finding no significant association after controlling the false discovery rate: the minimum q–value was 0.09. We also checked for correlation, known as linkage disequilibrium, between the 21 loci used to create the polygenic risk score and 32 loci that are known to influence age at menarche (Elks et al., 2010). The data available to us do not have the necessary information to analyze linkage disequilibrium between the two sets of loci.

Table 2.2: Results of the analysis of the Prostate, Lung, Colorectal and Ovarian cancer screening trial data

	β_Z	β_G	β_X	β_{GX}
Logistic				
Estimate	0.018	0.297	-0.165	0.124
std err	0.054	0.064	0.132	0.068
p–value	7.45×10^{-1}	3.19×10^{-6}	2.10×10^{-1}	6.87×10^{-2}
SPMLE, Rare				
Estimate	0.024	0.321	-0.175	0.138
std err (asymptotic)	0.054	0.067	0.134	0.055
p–value (asymptotic)	6.60×10^{-1}	1.62×10^{-6}	1.91×10^{-1}	1.16×10^{-2}
SPMLE, π_1 known				
Estimate	0.022	0.313	-0.174	0.141
std err (asymptotic)	0.054	0.065	0.133	0.055
p–value (asymptotic)	6.78×10^{-1}	1.64×10^{-6}	1.93×10^{-1}	1.13×10^{-2}

Logistic is ordinary logistic regression; *SPMLE, Rare* is our method using the rare disease approximation with unknown π_1; *SPMLE, π_1 known* is our method when the disease rate is known in the source population ($\pi_1 = 3.56\%$); *std err* is the asymptotic standard error estimate; β_Z is the main effect for age; β_G and β_X are the main effects for the polygenic risk score (G) and the environmental variable X (age at menarche $>$ 13), respectively; β_{GX} is the gene–environment interaction.

Using phased haplotypes from subjects of European descent from *1000 Genomes* (Consortium, 2015) and *HapMap* (Gibbs et al., 2003), no evidence of linkage disequilibrium was found: the maximum R^2 was 0.1 and the minimum q–value was 0.85. Finally, a 2014 study examined the relationship between age at menarche and 10 of the 21 SNPs used to create our polygenic risk score, none of which were found to influence age at menarche (Andersen et al., 2014).

Table 2.2 presents the results for the cases that π_1 is unknown and known, respectively: as remarked upon previously, the results are very similar. To provide a basis for comparison because of the very different scales of the variables, the variable age at baseline was standardized to have mean zero and standard deviation one. In addition, we standardized some of the coefficient estimates so that β_G was multiplied by the standard deviation of the polygenic risk score, and β_{GX} was multiplied by the standard deviation of X times the polygenic risk score.

As expected from the known association of the single nucleotide polymorphisms with risk of breast cancer, the polygenic risk score was strongly associated with breast cancer status of the women in the study. Standard logistic regression analysis reveals some evidence for interaction of the polygenic risk score with age-at-menarche, but the result was not statistically significant at the 0.05 level. When the analysis was done under the gene–environment independence assumption, the evidence of interaction appeared to be stronger.

The coefficient estimate for the interaction term is slightly larger for our semiparametric methods than that for logistic regression. Also, the asymptotic standard error estimate of logistic regression is approximately 23% larger than our methods, indicating a variance increase of $\approx 50\%$. Although not listed here, the bootstrap mentioned in §2.3.2 has very similar standard error estimates. In that bootstrap, 33% of the time, the logistic interaction estimate was actually greater than that of the disease rate known estimate.

2.5 Discussion and Extensions

We have proposed a general method for using retrospective likelihoods for studying gene–environment interactions involving multiple markers, a method that does not require any distributional assumption of the multivariate genotype distribution. Sometimes, one may consider modeling multi–marker gene–environment interactions using an underlying polygenic risk score, which is a weighted combination of numerous genetic markers where the weights are pre-determined from previous association studies. In such situations, the polygenic risk score might be assumed to follow approximately a normal distribution in the underlying population and the profile likelihood method of Chatterjee and Carroll (2005) can be used with appropriate modification by replacing the parametric multinomial distribution for a single nucleotide polymorphism genotype by a parametric normal distribution for the polygenic risk score, see also Chen et al. (2008) and Lin and Zeng (2009). In general, however, when an investigator desires to explore complex models for multivariate gene–environment interactions retaining separate parameters for distinct single nucleotide polymorphisms or for distinct genetic profiles defined by combinations of correlated single nucleotide polymorphisms, then one cannot avoid dealing with complex multivariate genotype

distributions, something that is not easy to specify through parametric models.

Our methods are types of semiparametric plug-in estimators, and thus have certain features in common with the work of Newey (1994), namely that the profile likelihood has the nonparametric component $R(x, \Omega)$ in (2.4) that is estimated by (2.5). Generally, however, such plug-in estimators are not semiparametric efficient. We believe it will be possible to create an efficient semiparametric estimator by modifying the work of Ma (2010): we are exploring this and its computational aspects, which may be daunting.

Supplementary Material

Supplementary Material available in the Appendix and at *Biometrika* online contains proofs, skewness and kurtosis and q–q plots for the simulation in Table 1, how to modify our methods to account for strata, additional simulations and software written in R. The data used in §2.4 are available from the National Cancer Institute via a data transfer agreement.

3. IMPROVED SEMIPARAMETRIC ANALYSIS OF POLYGENIC GENE-ENVIRONMENT INTERACTIONS IN CASE-CONTROL STUDIES

Standard logistic regression analysis of case-control data has low power to detect gene-environment interactions, but until recently it was the only method that could be used on complex polygenic data for which parametric distributional models are not feasible. Under the assumption of gene-environment independence in the underlying population, Stalder et al. (2017, *Biometrika*, **104**, 801-812) developed a retrospective method that treats both genetic and environmental variables nonparametrically. We propose an improvement to the method of Stalder et al. that increases the efficiency of the estimates with no additional assumptions and modest computational cost. This improvement is achieved by treating the genetic and environmental variables symmetrically to generate two sets of parameter estimates that are combined to generate a more efficient estimate. We employ a semiparametric framework to develop the asymptotic theory of the estimator, and evaluate its performance via simulation studies. The method is illustrated using data from a case-control study of breast cancer.

3.1 Introduction

Genetic epidemiologists have identified both genetic and environmental factors that influence the incidence of complex diseases such as cancers, heart diseases, depression, and diabetes (Gustavsson et al., 2016; Krischer et al., 2017; Mullins et al., 2016; Nickels, 2013; Rudolph, 2015). As new studies identify additional genetic variants associated with a disease, attention turns to exploring the interaction between genetic susceptibility and environmental risk factors.

Researchers studying gene-environment interactions often adopt a case-control study design, wherein diseased cases and healthy control subjects are identified and their covariate information is collected retrospectively. When the disease is rare, sampling cases and controls separately provides substantial cost and time savings over a prospective cohort study, but it makes statistical inference more complicated.

Prentice and Pyke (1979) demonstrated that standard prospective logistic regression of case-control data, which ignores the retrospective sampling scheme, nevertheless yields consistent estimates of all parameters except the logistic intercept. Logistic regression is equivalent to maximum likelihood estimation under a model that places no assumptions on the joint distribution of the genetic and environmental variables, and it achieves the variance lower bound under this model (Breslow et al., 2000).

To improve estimation efficiency, studies of gene-environment interactions often take advantage of the relatively mild assumption that the genetic and environmental variables are independently distributed in the source population. This assumption is easy to test, is frequently valid, and enables the use of specialized methods for the analysis of case-control data. Piegorsch et al. (1994) proposed a case-only approach that efficiently estimates multiplicative interactions (but not main effects) under the assumptions of gene-environment independence and rare disease. Chatterjee and Carroll (2005) exploited the gene-environment independence assumption to develop a semiparametric retrospective profile likelihood framework that treats environmental variables nonparametrically but assumes that the genetic variables have a known, discrete distribution. Further developments have yielded additional retrospective methods based on parametric modeling of the distribution of genetic variables given the environmental variables, see for example Han et al. (2012); Lobach et al. (2008); Ma (2010).

Genome-wide association studies have shown that genetic predisposition to a single disease tends to be highly polygenic, with many genetic variants influencing disease risk (Chatterjee et al., 2016; Fuchsberger et al., 2016). To provide a more complete picture of genetic risk and gene-environment interactions, it is often advantageous to include multiple genetic loci in the disease model (Chatterjee et al., 2006; Jiao et al., 2013; Lin et al., 2015). In the interest of parsimony, many studies have focused on developing polygenic risk scores through a weighted combination of all known genetic variants associated with a disease (Chatterjee et al., 2013; Dudbridge, 2013). Handling multiple genetic variants, polygenic risk scores, or a combination of both is straightforward with prospective logistic regression, but can be unwieldy or even impossible when using

retrospective methods that exploit gene-environment independence to gain efficiency but require a parametric model for the distribution of the genetic component.

The method of Stalder et al. (2017) extends the Chatterjee-Carroll retrospective profile likelihood framework by treating both the genetic and environmental variables nonparametrically, requiring only the assumption of gene-environment independence in the source population. This assumption of independence can be weakened if a discrete stratification variable is found such that genes and environment are independent within strata of the source population.

Here we propose an improvement to the method of Stalder et al. (2017) that increases the efficiency of the estimates with no additional assumptions and modest computational cost. This development relies on the observation that the method of Stalder et al. removes dependence on the distribution of the genetic and environmental variables in two different fashions; by treating the genetic and environmental variables symmetrically we generate two sets of parameter estimates that are combined to generate a more efficient estimate. We employ a semiparametric framework to develop the asymptotic theory of the estimator. The properties of the new method are illustrated through simulations in Section 3.3, and an example in Section 3.4.

3.2 Methodology and Theory

3.2.1 Background

We adopt notation similar to that of Stalder et al., with disease status, genetic information, and environmental risk factors denoted by D, G, and X, respectively. Both G and X are potentially multivariate and can contain both discrete and continuous components. For a given case-control study, n_1 is the number of cases ($D = 1$) and n_0 is the number of controls ($D = 0$), while $\pi_1 = \text{pr}(D = 1)$ is the disease rate in the source population and $\pi_0 = 1 - \pi_1$. We maintain the assumption in Stalder et al. that π_1 is either known or can be estimated well.

The assumption of gene-environment independence in the source population can be written as $f_{GX}(g, x) = f_G(g) \times f_X(x)$, where $f_{GX}(\cdot, \cdot)$ is the joint density or mass of G and X in the underlying population, and $f_G(\cdot)$ and $f_X(\cdot)$ are the marginal density or mass functions of G and X,

respectively, in the underlying population. We assume $f_X(x)$ and $f_G(g)$ are completely unspecified.

Given the genetic and environmental covariates, we assume the risk of disease in the underlying population follows the model $\text{pr}(D = 1 \mid G, X) = H\{\alpha_0 + m(G, X, \boldsymbol{\beta})\}$, where $H(x) = \{1 + \exp(-x)\}^{-1}$ is the logistic distribution function and $m(G, X, \boldsymbol{\beta})$ is a function that describes the joint effect of G and X and is known up to the unspecified parameters of interest $\boldsymbol{\beta}$.

Given the subject's disease status, the retrospective likelihood is the probability of observing the genetic and environmental variables. Under gene-environment independence in the source population, the retrospective likelihood is

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

The logistic intercept α_0, typically of little scientific interest, is not consistently estimated using prospective logistic regression, which instead converges to $\kappa = \alpha_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$ (Prentice and Pyke, 1979). For convenience, we parameterize everything in terms of κ, and we define $\Omega = (\kappa, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$. Chatterjee and Carroll (2005) profiled out $f_X(\cdot)$ to create a semiparametric profile likelihood

$$L_X(D, G, X, \Omega, f_G) = f_G(G) \frac{S(D, G, X, \Omega)}{R_X(X, \Omega)}, \tag{3.1}$$

where

$$\begin{aligned} S(d, g, x, \Omega) &= \frac{\exp[d\{\kappa + m(g, x, \beta)\}]}{1 + \exp\{\kappa - \log(n_1/n_0) + \log(\pi_1/\pi_0) + m(g, x, \beta)\}}; \\ R_X(x, \Omega) &= \textstyle\sum_{r=0}^{1} \int f_G(v) S(r, v, x, \Omega) dv. \end{aligned} \tag{3.2}$$

The key insight of Stalder et al. (2017) was to develop an unbiased estimator of $R_X(x, \Omega)$ that treats $f_G(\cdot)$ nonparametrically, defined as

$$\widehat{R}_X(x, \Omega) = \textstyle\sum_{j=1}^{n} \sum_{r=0}^{1} \sum_{d=0}^{1} (\pi_d/n_d) I(D_j = d) S(r, G_j, x, \Omega). \tag{3.3}$$

The leading term $f_G(G)$ in eq. (3.1) is constant with respect to Ω, and can be ignored for the purpose of estimation. Replacing $R_X(x, \Omega)$ with $\widehat{R}_X(x, \Omega)$ and taking the logarithm yields the estimated profile loglikelihood of Ω given the data as

$$\widehat{\mathcal{L}}_X(\Omega) = \sum_{i=1}^{n} \log\{S(D_i, G_i, X_i, \Omega)\} - \sum_{i=1}^{n} \log\{\widehat{R}_X(X_i, \Omega)\}. \tag{3.4}$$

Define $S_\Omega(d, g, x, \Omega) = \partial S(d, g, x, \Omega)/\partial\Omega$ and $\widehat{R}_{X\Omega}(x, \Omega) = \partial \widehat{R}_X(x, \Omega)/\partial\Omega$. The profile likelihood score function, $\mathcal{S}_X(\Omega)$, is unknown but can be estimated consistently by

$$\widehat{\mathcal{S}}_X(\Omega) = n^{-1/2} \sum_{i=1}^{n} \left\{ \frac{S_\Omega(D_i, G_i, X_i, \Omega)}{S(D_i, G_i, X_i, \Omega)} - \frac{\widehat{R}_{X\Omega}(X_i, \Omega)}{\widehat{R}_X(X_i, \Omega)} \right\}. \tag{3.5}$$

By solving $\widehat{\mathcal{S}}_X(\Omega) = 0$, we obtain a consistent estimate of Ω, which we will denote as $\widehat{\Omega}_X$ and which is called the SPMLE by Stalder et al. (2017).

3.2.2 Symmetric Combination Estimator

The above equations are equivalent to those found in Stalder et al. (2017) with the addition of the subscript X in eqs. (3.1) to (3.5) to emphasize that the density of X has been profiled out, leaving the density of G to be treated nonparametrically. Because our only assumption about G and X is their independence in the source population, we could just as well have interchanged them and profiled out the distribution of G. The notation in this symmetric case is analogous to the above, but with subscript G instead of X. It follows that the analogous estimated score function

$$\widehat{\mathcal{S}}_G(\Omega) = n^{-1/2} \sum_{i=1}^{n} \left\{ \frac{S_\Omega(D_i, G_i, X_i, \Omega)}{S(D_i, G_i, X_i, \Omega)} - \frac{\widehat{R}_{G\Omega}(G_i, \Omega)}{\widehat{R}_G(G_i, \Omega)} \right\}$$

can be used to obtain $\widehat{\Omega}_G$, another consistent estimate of Ω.

The optimal combination of symmetric estimators $\widehat{\Omega}_X$ and $\widehat{\Omega}_G$ follows the principle of generalized least squares. Suppose the dimension of Ω is p. Let I_p be the $p \times p$ identity matrix and define $\mathcal{X} = (I_p, I_p)^{\mathrm{T}}$. Define $\mathcal{Y} = (\widehat{\Omega}_X^{\mathrm{T}}, \widehat{\Omega}_G^{\mathrm{T}})^{\mathrm{T}}$ and $\Lambda_{\text{all}} = \text{cov}(\mathcal{Y})$. Theorem 2, in Section 3.2.3, shows

$\mathcal{Y} \to \text{Normal}(\mathcal{X}\Omega, \Lambda_{\text{all}})$.

Treating this as a generalized least squares problem, we can rewrite it as $\mathcal{Y} = \mathcal{X}\Omega + \epsilon$, where $\epsilon \sim \text{Normal}(0, \Lambda_{\text{all}})$. The Symmetric Combination Estimator is the solution to the linear model, namely

$$\widehat{\Omega}_{\text{Symm}} = (\mathcal{X}^{\text{T}}\Lambda_{\text{all}}^{-1}\mathcal{X})^{-1}\mathcal{X}^{\text{T}}\Lambda_{\text{all}}^{-1}\mathcal{Y}. \tag{3.6}$$

An alternative method of combining the two estimates is to average the two estimated profile likelihoods into a single composite likelihood. The resulting Composite Likelihood Estimator yields minimal efficiency gains over the SPMLE from Stalder et al., and is presented in Section B.1 of the *Supplementary Material*.

3.2.3 Asymptotic Theory

In this subsection we first demonstrate that the joint distribution of $\widehat{\Omega}_X$ and $\widehat{\Omega}_G$ is asymptotically normal. We then show the asymptotic results for the Symmetric Combination Estimator. In practice, asymptotic standard errors for the Symmetric Combination Estimator proved unreliable due to slow convergence, so bootstrap standard errors are used instead.

To state the asymptotic results, let $Z_i = (D_i, G_i, X_i)$ then define

$$
\begin{aligned}
\Gamma_X &= \sum_{d=0}^{1} \frac{n_d}{n} E\left[\frac{\partial}{\partial \Omega^{\mathrm{T}}}\left\{\frac{S_\Omega(Z,\Omega)}{S(Z,\Omega)} - \frac{R_{X\Omega}(X,\Omega)}{R_X(X,\Omega)}\right\}\middle| D=d\right];\\
\Gamma_G &= \sum_{d=0}^{1} \frac{n_d}{n} E\left[\frac{\partial}{\partial \Omega^{\mathrm{T}}}\left\{\frac{S_\Omega(Z,\Omega)}{S(Z,\Omega)} - \frac{R_{G\Omega}(X,\Omega)}{R_G(X,\Omega)}\right\}\middle| D=d\right];\\
\zeta_X(Z_i,\Omega) &= \frac{S_\Omega(Z_i,\Omega)}{S(Z_i,\Omega)} - \frac{R_{X\Omega}(X_i,\Omega)}{R_X(X_i,\Omega)}\\
&\quad - \sum_{d=0}^{1}\sum_{r=0}^{1} \frac{n_d \pi_{d_i}}{n_{d_i}} E\left\{\frac{S_\Omega(r,g,X,\Omega)}{R_X(X,\Omega)} - \frac{R_{X\Omega}(X,\Omega)S(r,g,X,\Omega)}{R_X^2(X,\Omega)}\middle| D=d\right\}_{g=G_i};\\
\zeta_G(Z_i,\Omega) &= \frac{S_\Omega(Z_i,\Omega)}{S(Z_i,\Omega)} - \frac{R_{G\Omega}(G_i,\Omega)}{R_G(G_i,\Omega)}\\
&\quad - \sum_{d=0}^{1}\sum_{r=0}^{1} \frac{n_d \pi_{d_i}}{n_{d_i}} E\left\{\frac{S_\Omega(r,G,x,\Omega)}{R_G(G,\Omega)} - \frac{R_{G\Omega}(G,\Omega)S(r,G,x,\Omega)}{R_G^2(G,\Omega)}\middle| D=d\right\}_{x=X_i};\\
\zeta_{X*}(Z_i,\Omega) &= \zeta_X(Z_i,\Omega) - E\{\zeta_X(Z,\Omega)|D=D_i\};\\
\zeta_{G*}(Z_i,\Omega) &= \zeta_G(Z_i,\Omega) - E\{\zeta_G(Z,\Omega)|D=D_i\}.
\end{aligned}
$$

By profiling X and G out separately, we have the following two equations

$$n^{1/2}(\widehat{\Omega}_X - \Omega) = -\Gamma_X^{-1} n^{-1/2} \textstyle\sum_{i=1}^{n} \zeta_{X*}(Z_i,\Omega) + o_p(1); \tag{3.7}$$

$$n^{1/2}(\widehat{\Omega}_G - \Omega) = -\Gamma_G^{-1} n^{-1/2} \textstyle\sum_{i=1}^{n} \zeta_{G*}(Z_i,\Omega) + o_p(1). \tag{3.8}$$

Equation (3.7) is proved in Theorem 1 of Stalder et al. (2017), and the proof of the symmetric case in eq. (3.8) is analogous.

To demonstrate the asymptotic properties of the Symmetric Combination Estimator, denote the block-diagonal matrix $\Gamma_{\text{all}}^{-1} = \text{diag}(\Gamma_X^{-1}, \Gamma_G^{-1})$.

Theorem 2. *Suppose that* $0 < \lim_{n\to\infty} n_d/n < 1$*, and* π_1 *is known. Then*

$$n^{1/2}(\mathcal{Y} - \mathcal{X}\Omega) = -\Gamma_{\text{all}}^{-1}\, n^{-1/2} \sum_{i=1}^{n} \begin{Bmatrix} \zeta_{X*}(Z_i,\Omega) \\ \zeta_{G*}(Z_i,\Omega) \end{Bmatrix} + o_p(1).$$

The Z_i are independent and $E\{\zeta_{X*}(Z_i,\Omega)|D_i\} = E\{\zeta_{G*}(Z_i,\Omega)|D_i\} = 0$, so as $n \to \infty$,

$$n^{1/2}(\mathcal{Y} - \mathcal{X}\Omega) \to \textit{Normal}(0, \Lambda_{\text{all}}) \tag{3.9}$$

in distribution, where

$$\begin{aligned}
\Lambda_{\text{all}} &= \Gamma_{\text{all}}\, \Sigma_{\text{all}}\, \Gamma_{\text{all}}^{\text{T}}; \\
\Sigma_{\text{all}} &= \text{cov}\left\{\begin{matrix}\zeta_{X*}(Z,\Omega) \\ \zeta_{G*}(Z,\Omega)\end{matrix}\right\} = \text{cov}\left\{\begin{matrix}\zeta_{X}(Z,\Omega) \\ \zeta_{G}(Z,\Omega)\end{matrix}\right\}.
\end{aligned}$$

The proof of Theorem 2 follows directly from the proofs of eqs. (3.7) and (3.8) and the properties of M-estimators $\widehat{\Omega}_X$ and $\widehat{\Omega}_G$.

Remark 1. *In Section 3.2.2, we constructed a linear model from eq.* (3.9) *and used generalized least squares to calculate $\widehat{\Omega}_{\text{Symm}}$. The asymptotic properties of GLS estimators inform us that as $n \to \infty$,*

$$n^{1/2}(\widehat{\Omega}_{\text{Symm}} - \Omega) \to \textit{Normal}\{0, (\mathcal{X}^{\text{T}}\Lambda_{\text{all}}^{-1}\mathcal{X})^{-1}\}.$$

In practice, $\widehat{\Omega}_X$ and $\widehat{\Omega}_G$ are highly correlated, which slows convergence to the asymptotic covariance matrix. Asymptotic estimates of standard errors proved unreliable in simulations, and are not recommended. Instead, we estimate $\text{cov}(\widehat{\Omega}_{\text{Symm}})$ *using a balanced bootstrap, where cases and controls are resampled separately, thus maintaining their respective sample sizes.*

3.2.4 Rare Diseases When π_1 is Unknown

Due to its sampling efficiency, the case-control design is typically used to study relatively rare diseases. If the true disease rate in the source population is unknown, it is common to assume that $\pi_1 \approx 0$ (Kwee et al., 2007; Lin and Zeng, 2006; Modan et al., 2001; Piegorsch et al., 1994). Under this rare disease assumption, $\widehat{\Omega}_X$ and $\widehat{\Omega}_G$ converge not to Ω, but to Ω_{X*} and Ω_{G*}, the solutions to their respective score equations with $\pi_1 = 0$. Using estimates of Ω_{X*} and Ω_{G*} to calculate $\widehat{\Omega}_{\text{Symm}}$ runs the risk of introducing bias, but in practice the small potential bias is typically inconsequential

unless the sample size is very large and standard errors unusually small. Examples in Section B.2.2 of the *Supplementary Material* demonstrate that under the rare disease approximation, coverage intervals remain near nominal until the true disease rate exceeds 8%.

3.3 Simulations

3.3.1 Scenario

To investigate the performance of the Symmetric Combination Estimator, we adopt the same simulation settings as reported in Stalder et al. (2017). Environmental variable X is binary with population frequency 0.5, and G consists of five correlated single nucleotide polymorphisms (SNPs). The SNPs follow a trinomial distribution in Hardy-Weinberg equilibrium, wherein SNP G_j takes values $(0, 1, 2)$ with probabilities $\{(1-p_j)^2,\ \ 2p_j(1-p_j),\ \ p_j^2\}$, respectively.

To generate correlated SNPs, we first simulated a 5-variate normal random variable with mean 0 and covariance between the jth and kth components equal to $0.7^{|j-k|}$. We then trichotomized the variates with appropriate thresholds so that the frequency of 0, 1, and 2 followed Hardy-Weinberg equilibrium with minor allele frequencies $(p_1, p_2, p_3, p_4, p_5) = (0.1, 0.3, 0.3, 0.3, 0.1)$.

Disease status was simulated according to the risk model $H\{\alpha_0 + m(G, X, \boldsymbol{\beta})\}$, with $m(G, X, \beta) = G^{\mathrm{T}}\beta_G + X\beta_X + (GX)^{\mathrm{T}}\beta_{GX}$. Here $\beta_G = \{\log(1.2), \log(1.2), 0, \log(1.2), 0\}$, $\beta_X = \log(1.5)$, and $\beta_{GX} = \{\log(1.3), 0, 0, \log(1.3), 0\}$. We set the logistic intercept $\alpha_0 = -4.165$ to yield a population disease rate $\pi_1 = 0.03$.

A sample of 1000 cases and 1000 controls was drawn from the simulated population, and parameters were estimated using logistic regression, the SPMLE of Stalder et al. (2017), the Symmetric Combination Estimator with known π_1, and the Symmetric Combination Estimator with a rare disease approximation. Standard error estimates for both logistic regression and the SPMLE were based on asymptotic theory, while those for the Symmetric Combination Estimator were calculated using 200 balanced bootstrap samples, as described in Remark 1.

3.3.2 Results

Table 3.1 presents the results of 1000 simulations comparing standard logistic regression, the SPMLE proposed by Stalder et al. with known π_1, our proposed Symmetric Combination Estimator with known π_1, and the Symmetric Combination Estimator using the rare disease approximation. Standard error estimates for logistic regression and the SPMLE were calculated using asymptotic theory, and the standard error estimates for both versions of the Symmetric Combination Estimator were calculated using 200 bootstrap samples as described in Remark 1.

The Symmetric Combination Estimator, both with known π_1 and when using the rare disease approximation, shows negligible bias and has coverage percentages near the nominal level. Like the SPMLE, both versions of our Symmetric Combination Estimator provide slightly more than 25% improvement in mean squared error efficiency over ordinary logistic regression for the main effect of X.

More impressively, our estimator nearly doubles the mean squared error efficiency of logistic regression for the main effects of G, and nearly triples the mean squared error efficiency for the interaction terms. This is a marked improvement even over the performance of the SPMLE, and it is accomplished without modeling the distribution of either G or X.

3.3.3 Further Simulations

Further simulations were conducted with multiple correlated SNPs and a binary environmental risk factor, but with changes to the number of SNPs (3 or 8), the population disease rate (1% or 5%), or the sample size (500 or 3000 cases & controls). All such simulations yielded results similar to those in Table 3.1 with regards to coverage, efficiency gains, and unbiasedness, and are thus not reported.

Section B.2 of the *Supplementary Material* contains the results of simulations examining the behavior of the Symmetric Combination Estimator in a variety of settings. Section B.2.1 contains an unabridged version of Table 3.1 that includes the SPMLE_G ($\widehat{\Omega}_G$) and the Composite Likelihood Estimator, neither of which approach the MSE efficiency of the Symmetric Combination

Table 3.1: Results of 1000 simulations as described in Section 3.3.1, comparing the bias, coverage, and efficiency of four estimators: ordinary logistic regression, the SPMLE of Stalder et al. with known π_1, our proposed Symmetric Combination Estimator with known π_1, and the Symmetric Combination Estimator using the rare disease approximation

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{XG1}	β_{XG2}	β_{XG3}	β_{XG4}	β_{XG5}
True	0.18	0.18	0.00	0.18	0.00	0.41	0.26	0.00	0.00	0.26	0.00
Logistic Regression											
Bias	0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
CI(%)	95.2	95.5	94.4	94.7	95.3	95.8	94.5	95.9	94.7	94.6	95.3
SPMLE, known π_1											
Bias	0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.01
CI(%)	95.4	95.8	94.8	96.1	96.3	95.6	94.6	96.0	94.3	95.6	95.1
MSE Eff	All G: 1.29					1.27	All $G*X$: 1.98				
Symmetric Combination Estimator, known π_1											
Bias	0.00	-0.03	0.00	0.00	-0.01	0.01	-0.03	0.02	0.00	-0.02	0.01
CI*(%)	96.7	95.7	96.7	96.5	97.8	95.4	94.8	96.7	96.2	96.6	97.2
MSE Eff	All G: 1.92					1.31	All $G*X$: 2.90				
Symmetric Combination Estimator, rare											
Bias	0.01	-0.02	0.00	0.01	-0.01	0.02	-0.05	0.02	0.00	-0.03	0.00
CI*(%)	96.4	95.7	95.7	96.3	98.1	94.9	94.0	97.0	96.4	95.5	97.5
MSE Eff	All G: 1.86					1.27	All $G*X$: 2.96				

CI: coverage of a 95% nominal confidence interval, calculated using asymptotic standard error. *CI**: coverage of a 95% nominal confidence interval, calculated using 200 bootstrap samples. *MSE Eff*: mean squared error efficiency when compared to logistic regression, averaged over G (All G) or $G*X$ interactions (All $G*X$).

Estimator. Section B.2.2 presents the results of simulations with misspecified population disease rate; we found the Symmetric Combination Estimator fairly robust to the misspecification of the disease rate. Section B.2.3 contains simulation studies examining the robustness of our method with respect to violations of the gene-environment independence assumption. Those simulations demonstrate that there will be bias in the estimated interaction parameter between a specific gene and a correlated environmental variable, but the rest of the parameter estimates continue unbiased, and the average mean squared error for all $G*X$ interactions can still be substantially lower than that obtained from prospective logistic regression. Section B.2.4 presents the results of simulations with different distributions for G and X.

3.4 Data Analysis

3.4.1 Data

Here we apply the proposed methodology to a case-control study of breast cancer. This case-control sample is taken from a large prospective cohort at the National Cancer Institute: the Prostate, Lung, Colorectal and Ovarian cancer screening trial (Canzian et al., 2010). This cohort enrolled 64,440 non-Hispanic, white women aged 55 to 74, of whom 3.72% developed breast cancer (Pfeiffer et al., 2013). The case-control study analyzed here consists of 658 cases and 753 controls.

Each of the 1411 subjects was genotyped for 21 SNPs that have been previously associated with breast cancer based on large genome-wide association studies. These SNPs were weighted by their log-odds-ratio coefficients and summed to define a polygenic risk score. A scaled version of this polygenic risk score, with mean zero and standard deviation one, was used as the genetic risk factor G. The individual SNPs and their coefficients can be found in Section B.2.5 of the *Supplementary Material*.

Early menarche is a known risk factor for breast cancer (Anderson et al., 2007), and environmental variable X is a binary indicator of whether the age at menarche is less than 14. The interaction between age at menarche and genetic breast cancer risk is a topic of interest, but the power to detect such interactions in previous studies has been limited (Gail, 2008).

The model fitted is $\text{pr}(D = 1) = H(\beta_0 + \beta_G G + \beta_X X + \beta_{GX} GX)$. While π_1 is known in this population, we apply our method using both the known disease rate and the rare disease approximation.

3.4.2 Verifying Gene-Environment Independence

Before applying our approach, we performed analyses to check the assumption of gene-environment independence in the population. Using the 753 controls, we ran a t-test of the polygenic risk score against the levels of X. The p-value was 0.91, indicating no evidence of correlation between G and X. We also ran chi-squared tests for each of the 21 individual genes and found no significant

association after controlling the false discovery rate: the minimum q-value was 0.09.

We also checked for correlation, known as linkage disequilibrium, between the 21 SNPs used to create the polygenic risk score and 32 SNPs known to influence age at menarche (Elks et al., 2010). Using phased haplotypes from subjects of European descent from *1000 Genomes* (Consortium, 2015) and *HapMap* (Gibbs et al., 2003), we were able to analyze 651 of the 672 possible linkages, and no evidence of linkage disequilibrium was found: the maximum R^2 was 0.1 and the minimum q-value was 0.85. Finally, a recent study of breast cancer susceptibility loci examined the relationship between age at menarche and 10 of the 21 SNPs used to create our polygenic risk score, none of which were found to influence age at menarche (Andersen et al., 2014).

3.4.3 Results

Table 3.2: Results of the analysis of the Prostate, Lung, Colorectal and Ovarian cancer screening trial data

	β_G	β_X	β_{GX}
Logistic Regression			
Estimate	0.539	0.124	-0.242
Standard Error (asymptotic)	0.117	0.128	0.133
p-value (asymptotic)	$< 1e$-4	0.331	0.068
Symmetric Combination, known $\pi_1 = 3.72\%$			
Estimate	0.495	0.093	-0.215
Standard Error (bootstrap)	0.094	0.133	0.089
p-value (bootstrap)	$< 1e$-4	0.484	0.016
Symmetric Combination, rare disease approximation			
Estimate	0.538	0.116	-0.237
Standard Error (bootstrap)	0.089	0.124	0.099
p-value (bootstrap)	$< 1e$-4	0.352	0.016

β_G and β_X are the main effects for the polygenic risk score G and the environmental variable X (age at menarche < 14), and β_{GX} is the gene-environment interaction.

Table 3.2 presents the results of our analysis with known π_1 and under a rare disease approximation. In both cases, standard errors for the Symmetric Combination Estimator were calculated using 500 bootstrap samples. The two estimates yield very similar results, indicating that a valid analysis can be conducted even if π_1 is not known.

The polygenic risk score was strongly associated with breast cancer status of the women in the

study, which is to be expected given that each of its component SNPs has a known association with breast cancer risk. Standard logistic regression analysis provides some indication of an interaction between the polygenic risk score and age at menarche, but the result is not statistically significant at the 0.05 level. Using the assumption of gene-environment independence in the population, the Symmetric Combination Estimator finds stronger evidence of this interaction. The improved power to detect this interaction is due to the much smaller standard error estimates of the Symmetric Combination Estimators. Using logistic regression, the estimated standard error of β_{GX} is 49% larger than with our method, indicating a variance increase of 121% (when applying the rare disease approximation, the variance increase is 81%).

3.5 Discussion and Extensions

Researchers investigating gene-environment interactions in case-control studies have traditionally had two broad options for analysis: logistic regression, which is flexible but has low power to detect interactions, or less flexible methods that exploit the assumption of gene-environment independence for increased efficiency. Improved understanding of genetic risk factors has led to the need for efficient estimators that can model complex gene-environment interactions. Stalder et al. (2017) proposed a retrospective profile method that exploits the assumption of gene-environment independence while treating the genetic and environmental variables nonparametrically. By obviating the need for a parametric model of genotype distributions, their method is well suited for the analysis of multimarker genetic data and polygenic risk scores.

We proposed an improvement to the method of Stalder et al. (2017) that increases the efficiency of the estimates with modest computational cost and no additional assumptions, making it applicable anywhere that the method of Stalder et al. can be used. The proposed Symmetric Combination Estimator places no distributional assumptions on the genetic or environmental variables, but it does rely on three assumptions. The first assumption, that the logistic risk model $H\{\alpha_0 + m(G, X, \boldsymbol{\beta})\}$ is known up to parameters α_0 and $\boldsymbol{\beta}$, is minimally restrictive because a flexible function, such as a function of b-splines, can be defined for $m(G, X, \boldsymbol{\beta})$. The second assumption, that π_1 is known or can be well estimated, can be relaxed by using the rare disease ap-

proximation of Section 3.2.4. Even if the true disease rate is not rare, the Symmetric Combination Estimator is generally robust to the misspecification of π_1, as demonstrated in Section B.2.2 of the *Supplementary Material*.

The final assumption is gene-environment independence in the source population. In Section B.2.3 of the *Supplementary Material*, we present the results of simulations demonstrating that bias is introduced in the estimated interaction parameter between correlated genetic and environmental variables, but the rest of the parameter estimates are unbiased. We recommend that researchers verify gene-environment independence before applying the Symmetric Combination Estimator, as we did in Section 3.4.2. To relax the gene-environment independence assumption, it should be possible to adapt the Symmetric Combination Estimator to the case where G and X are conditionally independent within the strata of an observed factor, as demonstrated in the *Supplementary Material* of Stalder et al. (2017). If suitable strata cannot be found, another possibility is to construct an empirical Bayes-type estimator like that of Mukherjee and Chatterjee (2008), which shrinks the Symmetric Combination Estimator back to the standard logistic regression estimate when the gene-environment independence assumption is violated.

4. SEMIPARAMETRIC ANALYSIS OF POLYGENIC GENE-ENVIRONMENT INTERACTIONS IN CASE-CONTROL STUDIES WITH CASECONTROLGE

Gene-environment interactions can be efficiently estimated in case-control data by methods that assume gene-environment independence in the source population, but until recently such techniques required parametric modelling of the genetic variables. The caseControlGE package implements the methods of Stalder et. al. (2017, *Biometrika*, **104**, 801-812) and Wang et. al. (2018, unpublished), which exploit the assumption of gene-environment independence without placing any assumptions on the marginal distributions of the genetic or environmental variables. These methods are ideally suited for analyzing complex polygenic data for which parametric distributional models are not feasible. In addition to the two estimators, the package also supplies a function to simulate case-control data and several helper functions for use on model objects. Use of this package is illustrated by simulating and analyzing data from a case-control study of breast cancer.

4.1 Introduction

4.1.1 caseControlGE package

The **caseControlGE** package (Asher, 2018) contains tools for the analysis of case-control data using `R` (R Core Team, 2018). It implements the methods of Stalder et al. (2017) and Wang et al. (2018), both of which fall under the class of semiparametric retrospective profile likelihood estimators. These methods are the first available to exploit the assumption of gene-environment independence while treating the genetic component nonparametrically. As such, they are well suited to replace logistic regression as the preferred method in situations where parametric distributional models are not feasible, such as in the analysis of complex polygenic data.

caseControlGE contains three main functions: `simulateCC`, `spmle`, and `spmleCombo`, as well as several helper functions. Section 4.2 of this paper introduces `simulateCC` in the context of simulating case-control data analogous to the data analyzed in Wang et al. (2018). Section 4.3 introduces `spmle` as a tool to analyze the simulated data, and Section 4.4 introduces

`spmleCombo` to conduct a more efficient analysis of the simulated data.

4.1.2 Background

Case-control studies are retrospective observational studies in which the sample consists of a group of healthy subjects and a group of diseased subjects. A crucial aspect of the case-control design is that the outcome, disease status, is known *before* sampling. The ability to deliberately oversample diseased subjects makes the case-control design cost effective, which is why it is widely popular in studies of gene-environment interactions.

Given the genetic and environmental covariates G and E, we assume the risk of disease D in the underlying population follows the model

$$\text{pr}(D = 1 \mid G, X) = H\{\beta_0 + m(G, X, \boldsymbol{\beta})\}, \tag{4.1}$$

where $H(x) = \{1 + \exp(-x)\}^{-1}$ is the logistic distribution function and $m(G, X, \boldsymbol{\beta})$ is a function that describes the joint effect of G and X and is known up to the unspecified parameters of interest $\boldsymbol{\beta}$.

Given the retrospective nature of case-control sampling, it is surprising that standard prospective logistic regression can be used to obtain unbiased estimates of $\boldsymbol{\beta}$ (Prentice and Pyke, 1979). Logistic regression requires no assumptions about the joint distribution of G and E, but it suffers from low power when estimating $G * E$ interaction effects. To gain efficiency, Chatterjee and Carroll (2005) exploited the assumption of gene-environment independence in the source population to maximize the retrospective likelihood while profiling out the distribution of E. Their method is available as the function `snp.logistic` in the *Bioconductor* package **CGEN** (Bhattacharjee et al., 2012).

The method of Chatterjee and Carroll, and subsequent methods utilizing the same retrospective profile likelihood framework, require a parametric model for the distribution of G given E. This becomes difficult as the number and complexity of genetic variables in the model grows. Capitalizing on advances in high-throughput genomics, genome-wide association studies have iden-

tified scores of SNPs associated with complex diseases such as cancers and diabetes. Modern case-control studies of gene-environment interactions need efficient methodology that allows for a flexible and arbitrarily complex genetic component, such as multiple correlated SNPs and/or continuous polygenic risk scores.

The SPMLE method of Stalder et al. (2017) extends the retrospective profile likelihood framework of Chatterjee and Carroll, dispensing with the need to model G parametrically. When the population disease rate π_1 is known, the retrospective profile loglikelihood can be estimated (up to an additive constant) using just the case-control sample and without modeling the distribution of G. When π_1 is unknown but the disease is rare, estimates can be obtained using the *rare disease approximation* that $\pi_1 \approx 0$, which typically introduces negligible bias (Stalder et al., 2017).

Wang et al. (2018) proposed an improvement to the method of Stalder et al. that increases the efficiency of the estimates with no additional assumptions. This development relies on the observation that the method of Stalder et al. removes dependence on the distribution of the genetic and environmental variables in two different fashions; by treating the genetic and environmental variables symmetrically Wang et al. generate two sets of parameter estimates that are combined to generate a more efficient estimate.

4.1.3 Implementation

The semiparametric method of Stalder et al. (2017) is implemented as the function `spmle` in **caseControlGE**, detailed in Section 4.3. Estimating the semiparametric profile likelihood is a computationally intensive process, and significant effort was invested in speeding up calculations. Estimation functions, including the analytic gradient and hessian, are written in C++ and compiled using **Rcpp** (Eddelbuettel, 2013), providing a tremendous speedup over native `R` code. Extensive benchmarking and code profiling was conducted, and estimation functions were written to apply matrix operations to contiguous blocks of memory whenever possible, reducing memory latency and allowing modern processors to exploit data level parallelism and perform the same operation on multiple data points simultaneously.

The estimated semiparametric likelihood is maximized using the quasi-Newton optimizer **ucminf**

(Nielsen and Mortensen, 2016) using starting values from logistic regression. `ucminf` is particularly well suited for this application because it allows us to precondition the optimization with the analytic hessian, and it evaluates the gradient after each call to the objective function. Calculating the gradient along with the likelihood adds negligible computational complexity, so we call a single C++ function to compute them both, then return them separately to `ucminf`. This leads `ucminf` to converge in roughly half the time of the next-fastest optimizers (several of the various `R` implementations of the BFGS algorithm tie for second place). The unmatched speed of `ucminf` means we are willing to tolerate its bugs, which include occasionally declaring convergence before actually converging. To address this, `spmle` checks the gradient at the reported optimum and restarts the optimization if necessary (with different starting values).

Computational complexity of the asymptotic covariance estimation, which contains a sum of the form $\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \partial \mathcal{L}_{ijk}(\Omega)/\partial\Omega$, was reduced from $O(n^3)$ to $O(n^2)$ by storing intermediate values in a three-dimensional array. This increases speed at the cost of memory usage, which climbs from $O(n)$ to $O(n^2)$, setting a practical limit on sample size in the low tens of thousands for average personal computers. This is sufficient to analyze all but the largest case-control studies; covariance estimates for larger studies should be computed using the bootstrap.

Asymptotic covariance estimates for the Symmetric Combination Estimator of Wang et al. converge slowly and unreliable in practice, often providing poor coverage. Wang et al. recommend a balanced bootstrap, with cases and controls resampled separately, to estimate covariance. **caseControlGE** offers users with multicore computers the option to speed up computation by using multiple processors. Parallelization is implemented using the `R` base package **parallel**, which is installed by default on all operating systems. Parallelization on computers running Linux or macOS is done by forking the active `R` session, saving time and memory. This option is unavailable in Windows, so parallelization is fractionally slower because a PSOCK cluster is created with a new instance of `R` running on each core.

4.2 Simulating case-control data with simulateCC

4.2.1 Data description

Wang et al. (2018) demonstrate the utility of their method by analyzing data from a case-control study of breast cancer. This case-control sample is taken from a large prospective cohort at the National Cancer Institute: the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial (Canzian et al., 2010). The case-control study analyzed by Wang et al. consists of 658 cases and 753 controls sampled from a cohort of 64,440 non-Hispanic, white women aged 55 to 74, of whom 3.72% developed breast cancer (Pfeiffer et al., 2013). The data are available from the National Cancer Institute via a data transfer agreement, but cannot be distributed with the **caseControlGE** package. Fortunately, we can use the **caseControlGE** function `simulateCC` to generate a similar data set for analysis.

Each of the 1411 subjects in the PLCO sample was genotyped for 21 SNPs that have been previously associated with breast cancer based on large genome-wide association studies. These SNPs were weighted by their log-odds-ratio coefficients and summed to define a polygenic risk score (PRS). A standardized version of this PRS, with mean zero and standard deviation one, was used as the genetic risk factor G by Wang et al.. Early menarche is a known risk factor for breast cancer, and Wang et al. used a binary indicator of whether the subject underwent early menarche as E (age at menarche < 14). Several environmental variables were recorded as part of the PLCO study, including body mass index (BMI). There is some evidence that obese women have a reduced risk of breast cancer, so in our simulation we will consider BMI in addition to the variables modeled by Wang et al..

4.2.2 Data simulation

Genetic variables generated by `simulateCC` include SNPs and three distributions of continuous PRS: Normal(0,1), Gamma(shape=20, scale=20), and bimodal. Environmental variables can be binary or Normal(0,1). To simulate case-control data with `simulateCC`, we specify distributions for G and E and provide regression coefficients $\boldsymbol{\beta}$ and intercept β_0 from eq. (4.1). The

function `simulateCC` generates values of G and E for a simulated population, then simulates binary D from its conditional distribution $(D \mid G, X, \beta_0, \boldsymbol{\beta})$. A sample of n_1 cases and n_0 controls is taken from this simulated population.

To determine the appropriate distributions to use when simulating G and E, we examine the PLCO data. In doing so, it is important to keep in mind that the case control sample is not representative of the source population. Case-control studies deliberately oversample cases, so the distribution of G and E in the sample may be quite different from the distribution of G and E in the population (especially for variables that are strongly correlated with disease status). To accurately simulate the genetic and environmental variables from the PLCO study, we need to estimate their distributions *in the source population*.

Wang et al. report $\beta_G = 0.459$ with $p < 1e-4$, but they standardized G to mean zero and standard deviation one *in the case-control sample*. G has a strong positive effect on disease risk, indicating that the distribution of $(G|D = 0)$ is meaningfully different from the distribution of $(G|D = 1)$. Specifically, $\mathbf{E}(G|D = 1) > \mathbf{E}(G|D = 0)$. With a population disease rate of 0.0372, this implies $\mathbf{E}_{\text{pop}}(G) \approx \mathbf{E}(G|D = 0) < 0$, where the subscript pop emphasizes that the expectation is in the source population.

This causes no problem for Wang et al., but it presents us with the dilemma that $G \nsim \text{N}(0, 1)$. If we simulate $G \sim \text{N}(0, 1)$ and use $\beta_G = 0.459$ as reported in Wang et al., our simulated $(D \mid G, X, \beta_0, \boldsymbol{\beta})$ will not match the distribution of the actual PLCO data.

If we did not have access to the PLCO data, our best option would be to approximate $\delta = \mathbf{E}(G|D = 0)$, simulate $G \sim \text{N}(\delta, 1)$, and use $\boldsymbol{\beta}$ as reported in Wang et al.. While we cannot distribute the PLCO data, we *can* use it to estimate population parameters, so approximating δ is not necessary. The simplest and most common way to estimate population parameters is to calculate them using just the controls. Case-control designs are typically used to study relatively rare diseases, and the bias introduced by using the cases as a stand-in for the population is usually quite small.

When π_1 is known, it is possible to calculate unbiased estimates by weighting the cases and

controls by π_1 and $(1-\pi_1)$, respectively. (This technique is employed to great effect by Stalder et al., and is the reason that `spmle` requires the user to specify a value for `pi1`.)

We return to the PLCO data to conduct an analysis similar to that of Wang et al., but with two environmental variables: the indicator of early menarche and BMI. We will standardize the two continuous variables due to their very different scales, but to make our lives easier when we conduct subsequent simulations, we standardize them to have mean zero and standard deviation one *in the source population*.

We calculate $\widehat{\mathbf{E}}_{\text{pop}}(\text{PRS})$, $\widehat{\mathbf{E}}_{\text{pop}}(\text{early menarche})$, and $\widehat{\mathbf{E}}_{\text{pop}}(\text{BMI})$ by weighting the means within cases and controls by π_1 and $(1-\pi_1)$, respectively. We calculate $\widehat{\text{sd}}_{\text{pop}}(\text{PRS})$ and $\widehat{\text{sd}}_{\text{pop}}(\text{BMI})$ using the sample standard deviations among the controls only. We have

$$G = \frac{\text{PRS} - \widehat{\mathbf{E}}_{\text{pop}}(\text{PRS})}{\widehat{\text{sd}}_{\text{pop}}(\text{PRS})}, \quad E_1 = \mathbf{I}(\texttt{age at menarche} < 14), \quad E_2 = \frac{\text{BMI} - \widehat{\mathbf{E}}_{\text{pop}}(\text{BMI})}{\widehat{\text{sd}}_{\text{pop}}(\text{BMI})}.$$

After this scaling, the distributions of G and E_2 in the source population can be well approximated by uncorrelated $\text{N}(0,1)$ random variables. Binary environmental variable E_1 is also uncorrelated with G, and has a frequency of 0.745 in the population. We fit a model in these variables to the PLCO data using `spmleCombo`, yielding the rest of the information we need to simulate case control data:

$$\begin{array}{lll} \pi_1 = 0.0372 & n_0 = 753 & n_1 = 658 \\ G \sim \text{N}(0,1) & E_1 \sim \text{Bin}(0.745) & E_2 \sim \text{N}(0,1) \\ \beta_G = 0.450 & \beta_{E_1} = 0.143 & \beta_{E_2} = -0.019 \\ & \beta_{GE_1} = -0.195 & \beta_{GE_2} = -0.040 \end{array}$$

The logistic intercept β_0 is not consistently estimated by logistic regression or either of the semi-parametric methods in **caseControlGE**, however it is typically of little interest. The function `simulateCC` prints the population disease rate each time it runs, so we run `simulateCC` several times with different values of β_0. Using a guess-and-check approach with increasing sample size as we get closer, we manipulate β_0 to match the disease rate observed in the source population.

```
#### Load the castControlGE package and set the random seed
library("caseControlGE")
set.seed(979)

#### Generate data with beta0 = -3 as a starting point
tmp = simulateCC(ncase=1000, ncontrol=1000, beta0=-3,
                 betaG_normPRS=0.450, betaE_bin=0.143,
                 betaE_norm=-0.019, betaGE_normPRS_bin=-0.195,
                 betaGE_normPRS_norm=-0.040, E_bin_freq=0.745)
#>
#> Disease prevalance: 0.0520909090909091
#### Disease rate too high, try beta0 = -4
tmp = simulateCC(ncase=1000, ncontrol=1000, beta0=-4,
                 betaG_normPRS=0.450, betaE_bin=0.143,
                 betaE_norm=-0.019, betaGE_normPRS_bin=-0.195,
                 betaGE_normPRS_norm=-0.040, E_bin_freq=0.745)
#>
#> Disease prevalance: 0.0212909427411371
#### Continue guessing, increasing sample size as we get closer
#### Not run during the vignette (was used when writing it)
## tmp = simulateCC(ncase=1e3, ncontrol=1e3, beta0=-3.5, ...
## tmp = simulateCC(ncase=1e4, ncontrol=1e4, beta0=-3.4, ...
## tmp = simulateCC(ncase=1e5, ncontrol=1e5, beta0=-3.4, ...
## tmp = simulateCC(ncase=1e5, ncontrol=1e5, beta0=-3.41, ...
rm(tmp)
```

After several iterations (commented out for speed), we determined that `beta0 = -3.41` produces a population disease rate of 0.0372. Now we generate our simulated PLCO data.

```
#### Set the random seed for reproducibility
set.seed(70)


#### Generate a synthetic data set that has similar properties
#### to the PLCO data
dat = simulateCC(ncase=658, ncontrol=753, beta0=-3.41,
                 betaG_normPRS=0.450, betaE_bin=0.143,
                 betaE_norm=-0.019, betaGE_normPRS_bin=-0.195,
                 betaGE_normPRS_norm=-0.040, E_bin_freq=0.745)
#>
#> Disease prevalance: 0.0362381630253141
```

4.2.3 Confirming the G-E independence assumption

The function `simulateCC` returns a list with elements `D`, `G`, and `E`, which are numeric vectors or matrices. We combine them into a `data.frame` to print the first 6 rows and tabulate by disease status.

```
#### Examine the simulated data
#### and tabulate the number of cases & controls
kable(list(head(as.data.frame(dat)), table(dat$D, dnn="D")),
      caption="Simulated PLCO data", booktabs=TRUE)
```

Table 4.1: Simulated PLCO data

D	G	E.1	E.2
0	0.1336533	1	0.3866567
0	0.4328288	0	1.4536655
0	0.3389738	1	-1.0088425
0	-1.4542346	1	0.7076604
0	-1.0777144	1	-0.2864078
0	1.4280646	1	1.6047819

D	Freq
0	753
1	658

Case-control data generated by `simulateCC` are sorted by disease status (which is why the first 6 rows do not contain a single case), but we see there are 658 cases and 753 controls. Before we calculate the `spmle`, we will check the assumption of gene-environment independence in the source population. In our case, this check is largely perfunctory because we did not provide the arguments `regress_E_bin_on_G_normPRS` or `regress_E_norm_on_G_normPRS` to `simulateCC`, so the genetic and environmental variates were drawn from independent distributions.

But when analyzing real data, it is crucial to verify this assumption. Violations of the G-E independence assumption can introduce bias in the estimates of interaction parameters between the specific genetic and environmental variables in violation of G-E independence (the other parameters in the model appear unaffected in simulation studies by Stalder et al.).

We do this by checking for dependence between `G` and `E` in the controls. Environmental variable E_1 is binary, so we use a *t*-test of G over the two levels of E_1. To test for dependence between G and E_2, we conduct a correlation test.

```
#### Save the indices of all controls in dat
controls = which(dat$D == 0)


#### t-test of G over the levels of E1
pander(t.test(dat$G[controls] ~ dat$E[controls, 1]),
       split.cells=11)
```

Table 4.2: Welch Two Sample t-test: `dat$G[controls]` by `dat$E[controls, 1]`

Test statistic	df	P value	Alternative hypothesis	mean in group 0	mean in group 1
-0.7291	317	0.4665	two.sided	-0.1163	-0.05508

```
#### correlation test between G and E2
pander(cor.test(dat$G[controls], dat$E[controls, 2]))
```

Table 4.3: Pearson's product-moment correlation: `dat$G[controls]` and `dat$E[controls, 2]`

Test statistic	df	P value	Alternative hypothesis	cor
-0.8129	751	0.4165	two.sided	-0.02965

Now that we are satisfied the assumption of gene-environment independence has not been violated, we can exploit this assumption using the two semiparametric retrospective methods of **caseControlGE**.

4.3 Analyzing case-control data with spmle

4.3.1 Known and rare disease

The function `spmle` is the backbone of **caseControlGE**; it is called on its own to evaluate the SPMLE of Stalder et al., and it is the key component of the Symmetric Combination Estimator of Wang et al.. Calling `spmle` is slightly different from calling other estimation commands like `lm` or `glm` because we do not specify the model formula to `spmle`. Instead we specify which variables are genetic and which are environmental, and `spmle` fits the formula: `D ~ G * E`.

`spmle` returns an `S3` object of class `"spmle"`. **caseControlGE** contains `spmle` methods for all applicable `S3` generics, such as `summary.spmle`, `print.spmle`, `anova.spmle`, `predict.spmle`, `confint.spmle`, etc.

We fit the gene-environment interaction model with `spmle`, and compare it to the estimates from standard logistic regression. We do not need to fit the logistic regression model with a call to `glm` because `spmle` automatically fits a logistic regression model to obtain starting values. This logistic regression model is returned with the fitted `spmle` object.

```
#### Fit the spmle to the simulated PLCO data
spmleFull = spmle(D=D, G=G, E=E, pi1=0.0372, data=dat)
#### Print coefficient estimates from spmle
#### and the logistic model returned by spmle
kable(summary(spmleFull)$coefficients,
      caption="spmle, known pi1", digits=4)
```

Table 4.4: spmle, known pi1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3226	0.1049	-3.0755	0.0021
G	0.5808	0.1033	5.6203	0.0000
E1	0.1762	0.1332	1.3225	0.1860
E2	0.0430	0.0565	0.7604	0.4470
G:E1	-0.2426	0.1064	-2.2807	0.0226
G:E2	-0.0436	0.0432	-1.0091	0.3129

```
kable(summary(spmleFull$glm_fit)$coefficients,
      caption="logistic regression", digits=4)
```

Table 4.5: logistic regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3014	0.1152	-2.6160	0.0089
G	0.5513	0.1111	4.9610	0.0000
E1	0.1549	0.1311	1.1814	0.2375
E2	0.0435	0.0562	0.7730	0.4395
G:E1	-0.2263	0.1279	-1.7695	0.0768
G:E2	-0.0130	0.0574	-0.2258	0.8213

The parameter estimates are extremely similar between the two models, but the `spmle` has smaller standard errors for the interaction terms. Logistic regression uncovers some evidence of a `G:E1` interaction between the PRS and early menarche, but the result is not significant at the 0.05 level. The `spmle` is able to provide stronger evidence of a `G:E1` interaction because the estimated standard error of the `G:E1` coefficient is 20% larger with logistic regression than the `spmle`, giving a variance increase of almost 45%.

In this instance we know the true population disease rate $\pi_1 = 0.0372$. If π_1 were unknown we would calculate the `spmle` under the assumption that $\pi_1 \approx 0$. Calculating the rare disease approximation using `spmle` is as simple as specifying `pi1 = 0` in the function call.

```
kable(summary(spmle(D=D, G=G, E=E, pi1=0, data=dat))$coef,
      caption="spmle, rare disease", digits=4)
```

Table 4.6: spmle, rare disease

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3253	0.1054	-3.0857	0.0020
G	0.5862	0.1054	5.5617	0.0000
E1	0.1791	0.1336	1.3404	0.1801
E2	0.0428	0.0566	0.7562	0.4495
G:E1	-0.2436	0.1051	-2.3175	0.0205
G:E2	-0.0437	0.0416	-1.0492	0.2941

The estimates and standard errors are nearly identical to the model with known π_1, indicating that a valid estimator can be obtained even when the disease rate is unknown.

4.3.2 Reduced model test

Body mass index does not appear to be a significant predictor in this model. The coefficients for the E_2 main effect and the $G * E_2$ interaction are near zero and both terms have large p values, so we fit a reduced model without BMI. To demonstrate the options controlling optimization, we disable hessian preconditioning and supply bad starting values to the optimizer while fitting the reduced model.

```
#### Fit the reduced spmle with bad starting values
spmleRed = spmle(D=D, G=G, E=E[,1], pi1=0.0372, data=dat,
                 startvals=rep(NA, 4), control=list(use_hess=F))
#> ucminf retry 1 of 2
```

With invalid starting values, `ucminf` is unable to converge during its first attempt. `splme` checks whether `ucminf` has converged and, seeing that it has not, prints "`ucminf retry 1 of 2`" and restarts the optimization with different starting values. If optimization had failed to converge on the second try, `spmle` would have issued an error. The number of retries, convergence criterion, and other optimization parameters can be passed as elements of the `control` argument.

`summary.spmle` reports the number of retries, number of `ucminf` iterations, and maximum gradient at the optimum, so we can confirm that the model really did converge while we check the parameter estimates.

```
#### Check convergence and parameter estimates
summary(spmleRed, signif.legend=FALSE)
#>
#> Call:
#> spmle(D = D, G = G, E = E[, 1], pi1 = 0.0372, data = dat,
#>       control = list(use_hess = F), startvals = rep(NA, 4))
#>
#> Pearson Residuals:
#>     Min       1Q   Median       3Q      Max
#> -1.9511  -0.9175  -0.6480   1.0055   2.1736
#>
#> Coefficients:
#>             Estimate Std. Error z value Pr(>|z|)
#> (Intercept)  -0.3221     0.1048  -3.075  0.00211 **
#> G             0.5808     0.1034   5.616 1.96e-08 ***
#> E[, 1]        0.1737     0.1331   1.305  0.19187
#> G:E[, 1]     -0.2400     0.1064  -2.256  0.02408 *
#>
#>     Null deviance: 1949.7  on 1410  degrees of freedom
#> Residual deviance: 1848.1  on 1407  degrees of freedom
#> AIC: 1856.1
#> UCMINF retries: 1, iterations: 14,
#>        max gradient at convergence: 4.917e-08
```

We see that not only did the model converge, but it converged to nearly the same parameter estimates as the full model. To check whether the reduced model is just as good as the model with BMI, we use `anova` to conduct a nested model test.

If the function `anova` is called on `spmle` objects, the method `anova.spmle` is used to

calculate likelihood ratio tests of the models. This is a valid way to test full vs reduced `spmle` models because the loglikelihood reported by `logLik.spmle` is accurate up to an additive constant. However, `anova` should not be used to compare an `spmle` model to a model fit by a different method.

```
#### Likelihood ratio test for reduced vs full model
pander(anova(spmleRed, spmleFull))
```

Table 4.7: Likelihood ratio test

#Df	LogLik	Df	Chisq	Pr(>Chisq)
4	-924	NA	NA	NA
6	-923.3	2	1.387	0.4998

The negligible difference in loglikelihood and large p value of the likelihood ratio test confirms our suspicion that BMI is not an important predictor of breast cancer status in the simulated PLCO data. Now that we have completed variable selection, we will fit the Symmetric Combination Estimator of Wang et al. using early menarche as the sole environmental variable. The Symmetric Combination Estimator is not a maximum (pseudo)likelihood estimator like `spmle`; it is the optimal combination of two such estimators. As such, it has no associated loglikelihood and the function `anova.spmle` cannot be used to compare models fit using the Symmetric Combination Estimator.

4.4 Analyzing case-control data with spmleCombo

4.4.1 Fitting spmleCombo with bootstrap standard error estimates

The SPMLE of Stalder et al. exploits the gene-environment independence assumption to produce an estimator that is substantially more efficient than standard logistic regression. It is remarkable then, that the Symmetric Combination Estimator of Wang et al. is yet more efficient than the SPMLE without requiring any additional assumptions about the data. But there is no free lunch, and the challenge presented by the Symmetric Combination Estimator is that its standard error converges to its asymptotic limit very slowly, making asymptotic standard error estimates imprecise and unreliable in practice.

The **caseControlGE** implementation of the Symmetric Combination Estimator, `spmleCombo`, estimates standard error using the balanced bootstrap recommended by Wang et al., wherein cases and controls are resampled separately to maintain their sample sizes. To speed up the process, `spmleCombo` can run the bootstrap on multiple cores in parallel with the argument `ncores`. By default, `spmleCombo` uses 50 bootstraps to estimate the standard error, executed in series on a single core. This is typically sufficient to obtain a reasonable estimate of standard error without undue computational burden. Users with multiple cores, small data sets, or copious free time can increase the number of bootstraps for more precise estimates, though there are diminishing returns in precision as the number of bootstraps grows large. Here we use 100 bootstraps distributed between 2 cores.

```
#### Set seed for reproducibility,
#### then fit spmleCombo with 100 bootstraps and 2 cores
set.seed(75)
comboRed = spmleCombo(D=D, G=G, E=E.1, pi1=0.0372,
                      data=as.data.frame(dat)[,-4],
                      nboot=100, ncores=2)
```

When we fit the SPMLE `spmleRed` in Section 4.3.2, we removed BMI from the model by

specifying `E=E[,1]`. In the code above we achieve the same effect by coercing `dat` into a `data.frame` and dropping E_2 from the `data` that is passed to `spmleCombo`. We examine the parameter estimates and variance.

```
#### Print coefficient estimates
#### for the Symmetric Combo reduced model
kable(summary(comboRed)$coefficients,
      caption="Symmetric Combo, known pi1", digits=4)
```

Table 4.8: Symmetric Combo, known pi1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3337	0.0949	-3.5177	0.0004
G	0.5426	0.0976	5.5601	0.0000
E.1	0.1887	0.1208	1.5615	0.1184
G:E.1	-0.2410	0.0954	-2.5250	0.0116

```
#### Ratio of variances: logistic regression / Symmetric Combo
pander(t(diag(vcov(comboRed$glm_fit))/diag(vcov(comboRed))),
       caption="Ratio of variances: logistic / Symmetric Combo")
```

Table 4.9: Ratio of variances: logistic / Symmetric Combo

(Intercept)	G	E.1	G:E.1
1.473	1.294	1.176	1.79

Parameter estimates from the Symmetric Combination Estimator are very similar to those of the SPMLE, but the efficiency is greater still. The lack of a viable asymptotic standard error is an inconvenience, but with no additional assumptions and a compute time measured in minutes if not seconds, the Symmetric Combination takes the day. And because the Symmetric Combination is a combination of two SPMLE models, `spmleCombo` includes both SPMLE fits in the `spmle`

object it returns. The first of which, labeled `spmle_E` because its likelihood was profiled over the distribution of E, is the same reduced model we fit with `spmle` in Section 4.3.2. The symmetric counterpart, which maximizes the likelihood profiled over the distribution of G, is returned as `spmle_G`.

```
#### Verify that the "spmle_E" component of comboRed
#### is the same as spmleRed from section 3.2
all.equal(coef(comboRed$spmle_E), coef(spmleRed),
          check.attributes=FALSE)
#> [1] TRUE


#### Print coefficient estimates for the SPMLE profiled over G
kable(summary(comboRed$spmle_G)$coef,
      caption="spmle profiled over G", digits=4)
```

Table 4.10: spmle profiled over G

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3167	0.1042	-3.0389	0.0024
G	0.5586	0.0979	5.7044	0.0000
E.1	0.1695	0.1327	1.2774	0.2015
G:E.1	-0.2322	0.1023	-2.2700	0.0232

The estimates from the SPMLE profiled over G bear a striking similarity to those from version profiled over E. It is this correlation that makes the asymptotic standard error of the Symmetric Combination Estimator so slow to converge to its asymptotic limit. `spmleCombo` will (grudgingly) report the asymptotic standard error estimate if we disable the bootstrap with `nboot = 0`. This triggers a warning, which we will see when we calculate the Symmetric Combination Estimator using the rare disease approximation.

```
#### Force spmleCombo to report the asymptotic SE
#### for the rare disease model
comboRare = spmleCombo(D=D, G=G, E=E.1, pi1=0,
                       data=as.data.frame(dat)[,-4], nboot=0)
#> Warning in spmleCombo(D = D, G = G, E = E.1, pi1 = 0, data =
#> as.data.frame(dat)[, : nboot=0,
#> using asymptotic standard error estimate,
#> which has poor coverage properties
```

If we *only* want point estimates of the parameters, and not their standard errors, we can ignore this warning. But if we want p-values or confidence intervals, we should not put faith in the asymptotic estimates of standard error.

```
#### Print coefficient estimates with asymptotic SE
#### for the rare disease approximation
kable(summary(comboRare)$coefficients,
      caption="spmleCombo, rare : Asymptotic SE", digits=4)
```

Table 4.11: spmleCombo, rare : Asymptotic SE

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3260	0.0954	-3.4180	0.0006
G	0.5494	0.0233	23.5543	0.0000
E.1	0.1807	0.1254	1.4417	0.1494
G:E.1	-0.2395	0.0324	-7.3816	0.0000

The parameter estimates are perfectly reasonable, but the asymptotic standard error estimates for G and $G * E_1$ are completely unbelievable.

4.4.2 Residual analysis

Diagnostic plots of the Pearson residuals from the SPMLE or Symmetric Combination models are interpreted similarly to diagnostic plots from logistic regression. As with logistic regression, we do not expect normally distributed residuals. Instead we check that the expected value of the residual is near zero over the range of fitted values. **caseControlGE** provides the `plot.spmle` function to do just this, and draws a lowess curve through the residuals.

```
#### Plot Pearson residuals vs predicted value
plot(comboRed,
     main="Residuals vs Fitted: Symmetric Combo, known pi1")
```

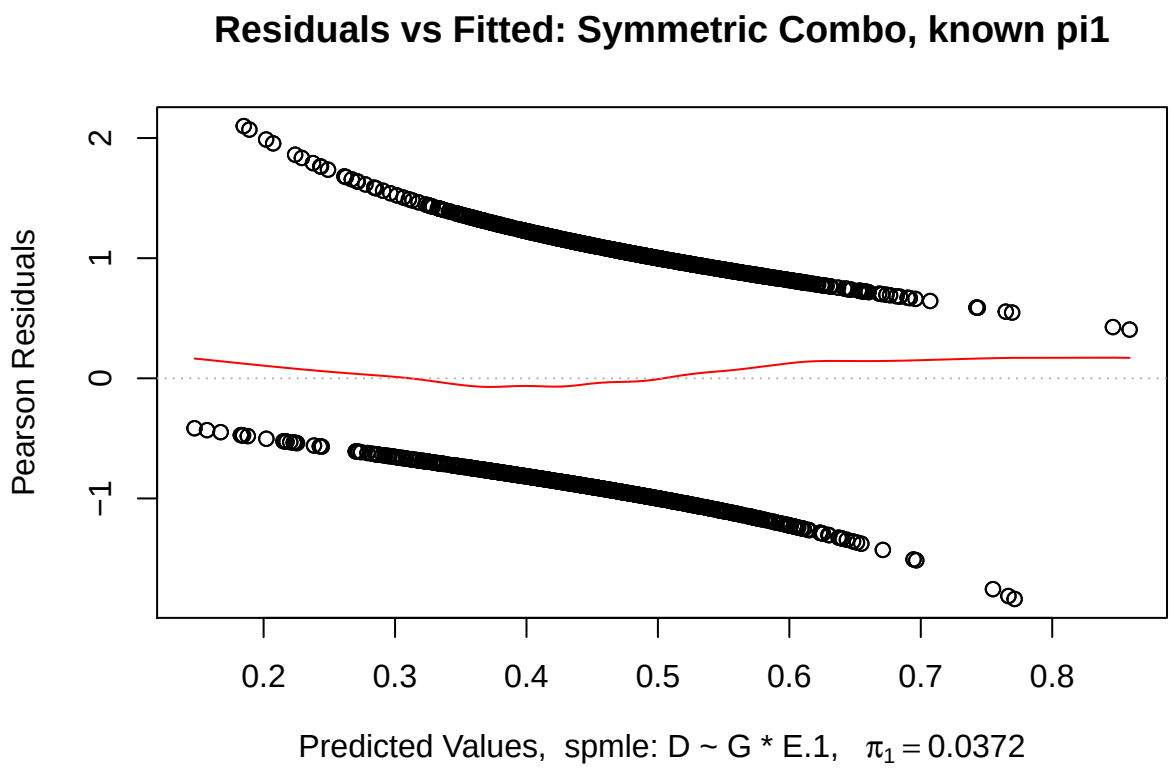

Figure 4.1: Reduced model diagnostics: residuals vs predicted values

The fitted lowess curve is near zero over the range of predicted values, raising no concerns

about model fit. Below we plot residuals vs. independent variables and find no issues

```
#### Plot Pearson residuals vs independent variables
plot(resid(comboRed)~dat$G, ylab="residuals", xlab="PRS")
panel.smooth(x=dat$G, y=resid(comboRed))
boxplot(resid(comboRed) ~ dat$E[,1], notch=TRUE,
        names=c("late menarche", "early menarche"))
```

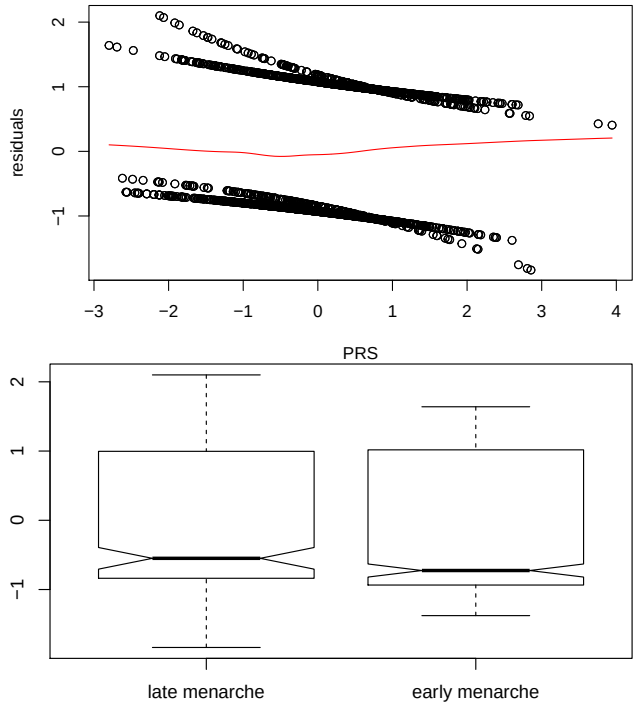

Figure 4.2: Reduced model diagnostics: residuals vs independent variables

Finally, we plot the residuals against BMI, the independent variable we removed from the model after the likelihood ratio test indicated it was not improving the model. We expect to see no pattern to the residuals.

```
#### Plot Pearson residuals vs BMI
plot(resid(comboRed)~dat$E[,2], ylab="residuals", xlab="BMI",
     main="Residuals vs BMI")
panel.smooth(x=dat$E[,2], y=resid(comboRed))
```

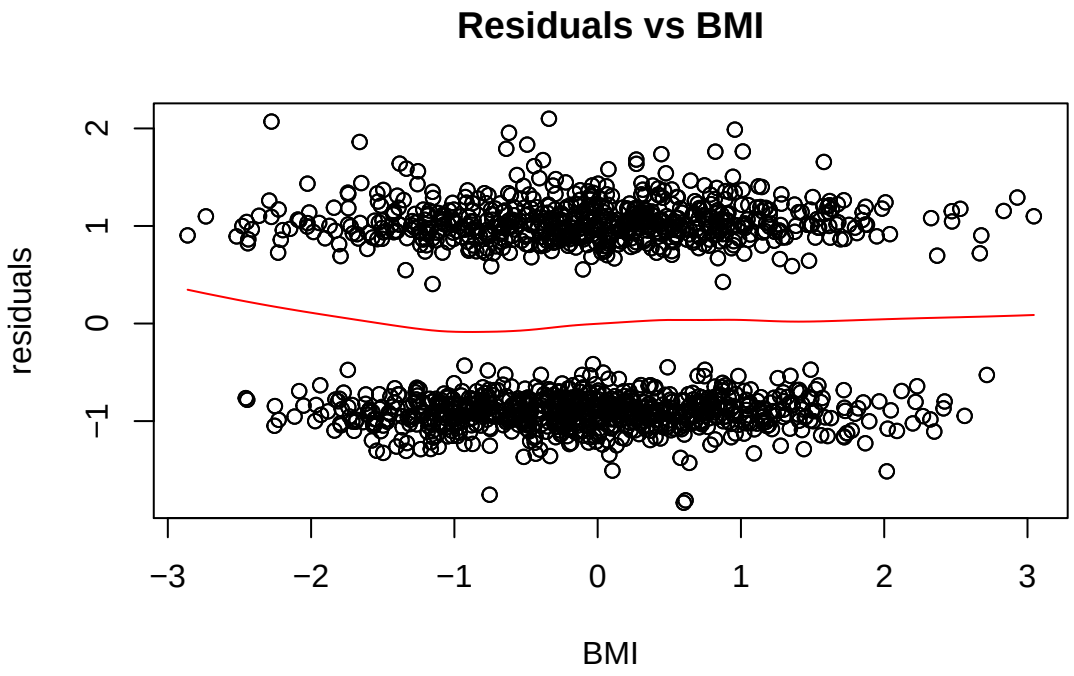

Figure 4.3: Reduced model diagnostics: residuals vs BMI, excluded from the model

4.4.3 Predictions

Satisfied that the model fits well, our final task is to visualize the model by predicting the expected probability of developing breast cancer for subjects who experience early or late menarche, spanning the gamut of PRS values. We use the function `predict.spmle`, which takes the same arguments as the `predict` methods for other classes. The one detail we must keep in mind is that `spmle` and `spmleCombo` allow the `data` argument to be a `data.frame` *or* a `list`. When we fit `comboRed` we coerced `dat` into a `data.frame`, so when we make predictions using

comboRed, we will need to supply newdata as a data.frame with variable names "D", "G", and "E.1".

The following lengthy block of code creates two sets of prediction data: one for women who underwent early menarche and one for women who did not. We plot the curves for both groups of women along with a 95% confidence interval of the mean (predict.spmle does not calculate prediction intervals for an individual because they are not easily interpretable with a binary response).

```
#### Store vectors of polygenic risk scores,
#### separated by disease status and menarche timing
GcaseEarly = dat$G[which(dat$D==1 & dat$E[,1]==1)]
GcaseLate = dat$G[which(dat$D==1 & dat$E[,1]==0)]
GcontrolEarly = dat$G[which(dat$D==0 & dat$E[,1]==1)]
GcontrolLate = dat$G[which(dat$D==0 & dat$E[,1]==0)]

#### Create a grid of PRS values
xg = seq(from=min(dat$G), to=max(dat$G), length=100)

#### Create two sets of new data,
#### one with early menarche, the other without
newEarly = data.frame(G=xg, E.1=1)
newLate = data.frame(G=xg, E.1=0)

#### Predict risk and get CI for both data sets
predEarly = predict(object=comboRed, newdata=newEarly,
                    interval="confidence", type="response")
predLate = predict(object=comboRed, newdata=newLate,
                   interval="confidence", type="response")
```

```
#### Calculate the boundaries of a 95% CI for each data set
CIearly = data.frame(x=c(xg, rev(xg)),
                     y=c(predEarly[,"lwr"], rev(predEarly[,"upr"])))
CIlate = data.frame(x=c(xg, rev(xg)),
                      y=c(predLate[,"lwr"], rev(predLate[,"upr"])))

#### Plot the predictions for women with early age at menarche
plot(x=xg, y=predEarly[,"fit"], xlim=range(xg),
 main="Effect of Polygenic Risk Score and Early Age at Menarche",
 ylim=range(c(predEarly, predLate)), type="l",
 lwd=3, col="blue", xlab="Polygenic Risk Score",
 ylab="Predicted pr(D=1)", xaxs="i")

#### Add the predictions for women with late age at menarche
lines(x=xg, y=predLate[,"fit"], lwd=3, col="red")

#### Shade the 95% CIs
polygon(x=CIearly, col=addTrans("blue", 120), border=NA)
polygon(x=CIlate, col=addTrans("red", 120), border=NA)

#### Add a rug for each data set: Early in blue, Late in red,
#### and cases above, controls below
rug(GcaseEarly, ticksize=0.02, side=3, col="blue")
rug(GcaseLate, ticksize=0.02, side=3, col="red")
rug(GcontrolEarly, ticksize=0.02, side=1, col="blue")
rug(GcontrolLate, ticksize=0.02, side=1, col="red")
```

```
#### Add a vertical line at the population mean PRS
abline(v=0, lty=3, col="darkslategray")


#### Add a legend
legend(x="topleft", inset=0.05, title="Age at Menarche",
       legend=c("Early", "Late"), col=c("blue", "red"), lwd=3)
```

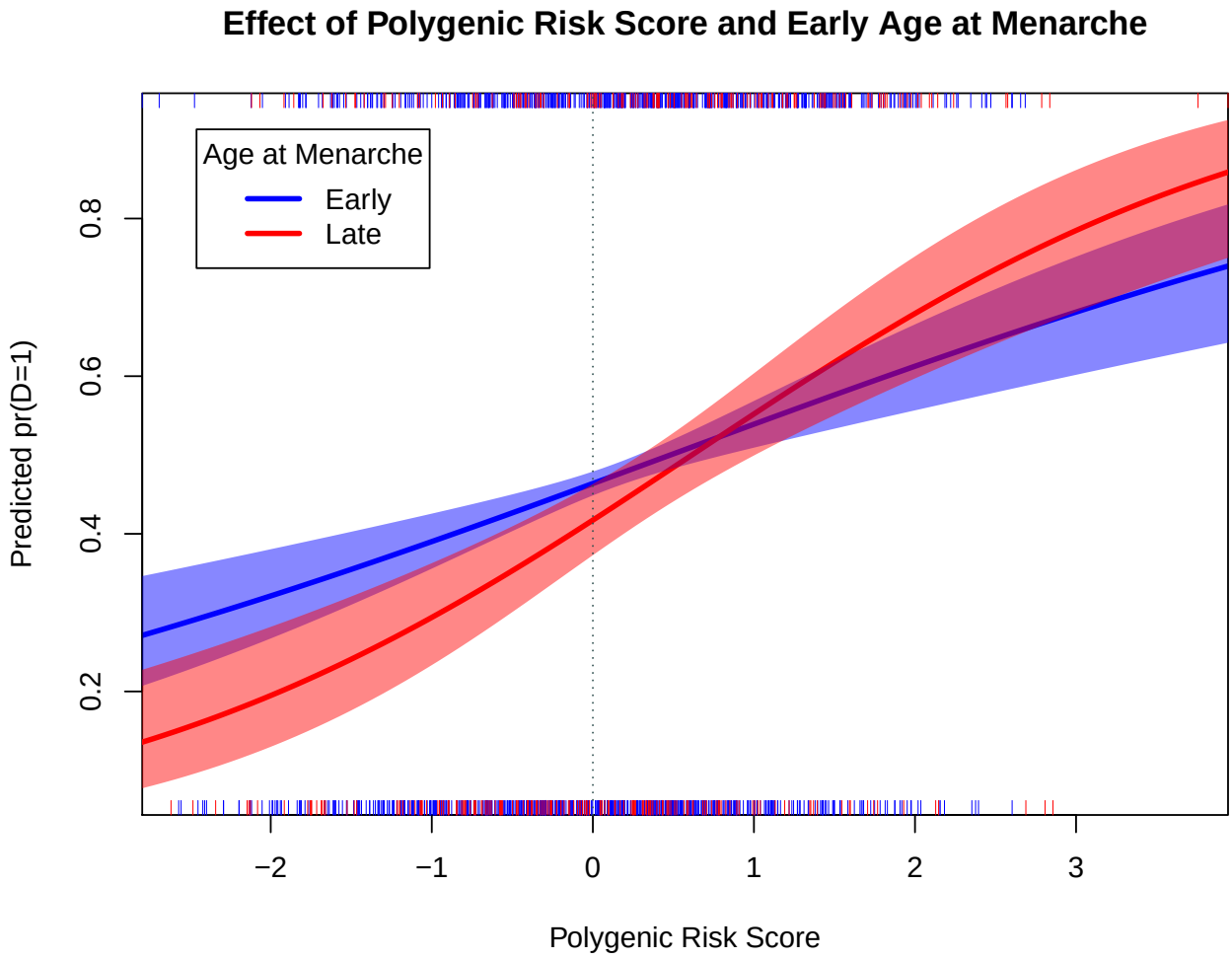

Figure 4.4: Predicted probability of developing breast cancer

Both early menarche and PRS have positive main effects, but the interaction has a negative coefficient. Visualizing the model, we see how this interesting effect plays out: *on average*, women

who experienced early menarche are at higher risk for breast cancer. However, among women with a high PRS, those who underwent early menarche are at lower risk of developing breast cancer than those who underwent late menarche.

The difference in risk at the far right end of the scale appears dramatic, but examining the tic-marks indicating observed values of PRS we see that there is a lone outlier with PRS > 3, so we would be well advised to ignore or crop values of PRS above 3. On the other end of the spectrum, women who had early menarche experience higher risk of breast cancer than women with equally "good" genes who had late menarche.

5. SUMMARY

Researchers investigating gene-environment interactions in case-control studies have traditionally had two broad options for analysis: logistic regression, which is flexible but has low power to detect interactions, or less flexible methods that exploit the assumption of gene-environment independence for increased efficiency. Improved understanding of genetic risk factors has led to the need for efficient estimators that can model complex gene-environment interactions.

We have proposed a general method for using retrospective likelihoods for studying gene-environment interactions involving multiple markers, a method that does not require any distributional assumption of the multivariate genotype distribution (Stalder et al., 2017). By obviating the need for a parametric model of genotype distributions, this method is well suited for the analysis of multimarker genetic data and polygenic risk scores. Additionally, we proposed an improvement to the aforementioned method that increases the efficiency of the estimates with modest computational cost and no additional assumptions.

To make these methods available to the public, we have created a free and open source software package implementing both methods (Asher, 2018). The package is extensively documented and is available for download at `https://github.com/alexasher/caseControlGE/`.

REFERENCES

Andersen, S. W., Trentham-Dietz, A., Gangnon, R. E., Hampton, J. M., Skinner, H. G., Engelman, C. D., Klein, B. E., Titus, L. J., Egan, K. M., and Newcomb, P. A. (2014). Breast cancer susceptibility loci in association with age at menarche, age at natural menopause and the reproductive lifespan. *Cancer Epidemiology*, 38, 62–65.

Anderson, W. F., Matsuno, R. K., Sherman, M. E., Lissowska, J., Gail, M. H., Brinton, L. A., Yang, X. R., Peplonska, B., Chen, B. E., Rosenberg, P. S., Chatterjee, N., Szeszenia-Dabrowska, N., Bardin-Mikolajczak, A., Zatonski, W., Devesa, S. S., and García-Closas, M. (2007). Estimating age-specific breast cancer risks: a descriptive tool to identify age interactions. *Cancer Causes & Control: CCC*, 18, 439 – 447.

Asher, A. (2018). *Semiparametric Gene-Environment Interactions in Case-Control Studies*. R package version 0.2.

Bhattacharjee, S., Chatterjee, N., Han, S., Song, M., and Wheeler, W. (2012). *CGEN: An R package for analysis of case-control studies in genetic epidemiology*. R package version 3.6.2.

Breslow, N. E., Robins, J. M., and Wellner, J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli*, 6, 447–55.

Canzian, F., Cox, D. G., Setiawan, V. W., Stram, D. O., Ziegler, R. G., Dossus, L., Beckmann, L., Blanché, H., Barricarte, A., Berg, C. D., et al. (2010). Comprehensive analysis of common genetic variation in 61 genes related to steroid hormone and insulin-like growth factor-i metabolism and breast cancer risk in the NCI breast and prostate cancer cohort consortium. *Human Molecular Genetics*, 19, 3873–3884.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. *Biometrika*, 92, 399–418.

Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U., and Wacholder, S. (2006). Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *American Journal of Human Genetics*, 79, 1002–1016.

Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, 17, 392–406.

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., and Park, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genetics*, 45, 400–405.

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.

Chen, Y. H., Chatterjee, N., and Carroll, R. J. (2009). Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association*, 104, 220–233.

Consortium, T. . G. P. (2015). A global reference for human genetic variation. *Nature*, 526, 68–74.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, 9, e1003348.

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer, New York. ISBN 978-1-4614-6867-7.

Elks, C. E., Perry, J. R. B., Sulem, P., Chasman, D. I., Franceschini, N., He, C., Lunetta, K. L., Visser, J. A., Byrne, E. M., Cousminer, D. L., et al. (2010). Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nature Genetics*, 42, 1077–1085.

Epstein, M. P. and Satten, G. A. (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics*, 73, 1316–1329.

Fuchsberger, C., Flannick, J., Teslovich, T. M., et al. (2016). The genetic architecture of type 2 diabetes. *Nature (doi: 10.1038/nature18642)*, .

Gail, M. H. (2008). Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *JNCI: Journal of the National Cancer Institute*, 100, 1037–1041.

Gauderman, W. J., Zhang, P., Morrison, J. L., and Lewinger, J. P. (2013). Finding novel genes

by testing G× E interactions in a Genome-Wide Association Study. *Genetic Epidemiology*, 37, 603–613.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. (2003). The International hapmap Project. *Nature*, 426, 789–796.

Gustavsson, J., Mehlig, K., Leander, K., Berg, C., Tognon, G., Strandhagen, E., Björck, L., Rosengren, A., Lissner, L., and Nyberg, F. (2016). Fto gene variation, macronutrient intake and coronary heart disease risk: a gene-diet interaction analysis. *European Journal Of Nutrition*, 55, 247 – 255.

Han, S. S., Rosenberg, P. S., Garcia-Closas, M., Figueroa, J. D., Silverman, D., Chanock, S. J., Rothman, N., and Chatterjee, N. (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.

Han, S. S., Rosenberg, P. S., Ghosh, A., Landi, M. T., Caporaso, N. E., and Chatterjee, N. (2015). An exposure-weighted score test for genetic associations integrating environmental risk factors. *Biometrics*, 71, 596–605.

Hayes, R. B., Reding, D., Kopp, W., Subar, A. F., Bhat, N., Rothman, N., Caporaso, N., Ziegler, R. G., Johnson, C. C., Weissfeld, J. L., et al. (2000). Etiologic and early marker studies in the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial. *Controlled Clinical Trials*, 21, 349S–355S.

Hsu, L., Jiao, S., Dai, J. Y., Hutter, C., Peters, U., and Kooperberg, C. (2012). Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genetic Epidemiology*, 36, 183–194.

Jiao, S., Hsu, L., Bézieau, S., Brenner, H., Chan, A. T., Chang-Claude, J., Le Marchand, L., Lemire, M., Newcomb, P. A., Slattery, M. L., et al. (2013). SBERIA: Set-based gene-environment interaction test for rare and common variants in complex diseases. *Genetic Epidemiology*, 37, 452–464.

Krischer, J. P., Lynch, K. F., Lernmark, A., Hagopian, W. A., Rewers, M. J., She, J.-X., Toppari, J.,

Ziegler, A.-G., and Akolkar, B. (2017). Genetic and environmental interactions modify the risk of diabetes-related autoimmunity by 6 years of age: The teddy study. *Diabetes Care*, 40, 1194 – 1202.

Kwee, L. C., Epstein, M. P., Manatunga, A. K., Duncan, R., Allen, A. S., and Satten, G. A. (2007). Simple methods for assessing haplotype-environment interactions in case-only and case-control studies. *Genetic Epidemiology*, 31, 75–90.

Lin, D. Y. and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association*, 101, 89–104.

Lin, D. Y. and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology*, 33, 256–265.

Lin, X., Lee, S., Christiani, D. C., and Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*, 14, 667–681.

Lin, X., Lee, S., Wu, M. C., Wang, C., Chen, H., Li, Z., and Lin, X. (2015). Test for rare variants by environment interactions in sequencing association studies. *Biometrics*, 72, 156–164.

Lobach, I., Carroll, R. J., Spinka, C., Gail, M. H., and Chatterjee, N. (2008). Haplotype-based regression analysis of case-control studies with unphased genotypes and measurement errors in environmental exposures. *Biometrics*, 64, 673–684.

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

Meigs, J. B., Shrader, P., Sullivan, L. M., McAteer, J. B., Fox, C. S., Dupuis, J., Manning, A. K., Florez, J. C., Wilson, P. W., D'Agostino Sr, R. B., et al. (2008). Genotype score in addition to common risk factors for prediction of type 2 diabetes. *New England Journal of Medicine*, 359, 2208–2219.

Modan, B., Hartge, P., Hirsh-Yechezkel, G., Chetrit, A., Lubin, F., Beller, U., Ben-Baruch, G., Fishman, A., Menczer, J., Struewing, J. P., et al. (2001). Parity, oral contraceptives, and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *New England Journal of Medicine*, 345, 235–240.

Mukherjee, B., Ahn, J., Gruber, S. B., and 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.

Mukherjee, B. and Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case–control studies: An empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, 64, 685–694.

Mullins, N., Power, R. A., Fisher, H. L., Euesden, J., Iniesta, R., Craig, I. W., Farmer, A. E., McGuffin, P., Breen, G., Lewis, C. M., et al. (2016). Polygenic interactions with environmental adversity in the aetiology of major depressive disorder. *Psychological Medicine*, 46, 759–770.

Murcray, C. E., Lewinger, J. P., and Gauderman, W. J. (2009). Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology*, 169, 219–226.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62, 1349–1382.

Nickels, S. e. a. (2013). Evidence of gene-environment interactions between common breast cancer susceptibility loci and established environmental risk factors. *PLoS Genetics*, 9, e1003284.

Nielsen, H. B. and Mortensen, S. B. (2016). *ucminf: General-Purpose Unconstrained Non-Linear Optimization*. R package version 1.1-4.

Pfeiffer, R. M., Park, Y., Kreimer, A. R., Lacey Jr, J. V., Pee, D., Greenlee, R. T., Buys, S. S., Hollenbeck, A., Rosner, B., Gail, M. H., et al. (2013). Risk prediction for breast, endometrial, and ovarian cancer in white women aged 50 y or older: derivation and validation from population-based cohort studies. *PLoS Medicine*, 10, e1001492.

Piegorsch, W. W., Weinberg, C. R., and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Statistics in Medicine*, 13, 153–162.

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

Prorok, P. C., Andriole, G. L., Bresalier, R. S., Buys, S. S., Chia, D., Crawford, E. D., Fogel, R.,

Gelmann, E. P., Gilbert, F., Hasson, M. A., et al. (2000). Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Controlled Clinical Trials*, 21, 273S–309S.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rudolph, A. e. a. (2015). Investigation of gene-environment interactions between 47 newly identified breast cancer susceptibility loci and environmental risk factors. *International Journal of Cancer*, 136, 685–696.

Stalder, O., Asher, A., Liang, L., Carroll, R. J., Ma, Y., and Chatterjee, N. (2017). Semiparametric analysis of complex polygenic gene-environment interactions in case-control studies. *Biometrika*, 104, 801–812.

Umbach, D. M. and Weinberg, C. R. (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine*, 16, 1731 – 1743.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wacholder, S., Hartge, P., Prentice, R., Garcia-Closas, M., Feigelson, H. S., Diver, W. R., Thun, M. J., Cox, D. G., Hankinson, S. E., Kraft, P., et al. (2010). Performance of common genetic variants in breast-cancer risk models. *New England Journal of Medicine*, 362, 986–993.

Wang, T., Asher, A., and Carroll, R. J. (2018). Improved semiparametric analysis of polygenic gene-environment interactions in case-control studies. *To Appear*, .

Wei, J., Carroll, R. J., Muller, U., Van Keilegom, I., and Chatterjee, N. (2013). Locally efficient estimation for homoscedastic regression in the secondary analysis of case-control data. *Journal of the Royal Statistical Society, Series B*, 75, 185–206.

Zhao, L. P., Li, S. S., and Khalid, N. (2003). A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *American Journal of Human Genetics*, 72, 1231–1250.

APPENDIX A

APPENDIX TO SECTION 2*

A.1 Proof of Theorem 1

Sketch of Technical Arguments

Necessary U-Statistic theory

Consider the case of one sample. Let $Z_1,\ldots,Z_n$ be independent and identically distributed. Let $h_*(\cdot)$ be a function such that $E\{h_*(Z_1, Z_2)\} = 0$. Define

$$U_{n*} = \sum_{i=1}^{n}\sum_{j\neq i}^{n} h_*(Z_i, Z_j)/\{n(n-1)\} = \sum_{i=1}^{n}\sum_{j<i} h_*(Z_i, Z_j)/\{n(n-1)/2\}.$$

If $h_*(z_1, z_2) \neq h_*(z_2, z_1)$, we make it symmetric in its arguments by noticing that if

$$h(z_1, z_2) = \{h_*(z_1, z_2) + h_*(z_2, z_1)\}/2,$$

then

$$U_{n*} = U_n = \sum_{i=1}^{n}\sum_{j\neq i}^{n} h(Z_i, Z_j)/\{n(n-1)\}.$$

We recognize U_n as a U-statistic of order 2 with a symmetric kernel $h(\cdot)$. Define

$$h_1(z) = 2E\{h(z, Z_2)\}. \tag{A.1}$$

*Reprinted with permission from "Semiparametric analysis of complex polygenic gene-environment interactions in case-control studies" by Stalder, O., Asher, A., Liang, L., Carroll, R. J., Ma, Y., and Chatterjee, N., 2017. *Biometrika*, 104, 801-812, Copyright 2017 by Oxford University Press.

Then, as in Theorem 12.3 of Van der Vaart (1998),

$$n^{1/2}U_n = n^{-1/2}\sum_{i=1}^{n} h_1(Z_i) + o_p(1). \tag{A.2}$$

Next we consider a special case of two samples, namely the n_0 controls and n_1 cases, denoted as $(U_1, \ldots, U_{n_0})$ and $(V_1, \ldots, V_{n_1})$, respectively, with $n = n_0 + n_1$. The U-statistic of interest is

$$U_n = (n_0 n_1)^{-1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I(D_i = 0)I(D_j = 1)h(U_i, V_j), \tag{A.3}$$

where $0 = E\{h(U_i, V_j) \mid D_i = 0, D_j = 1\}$. Let $n_0/n \to \lambda$ and $n_1/n \to 1 - \lambda$, with $0 < \lambda < 1$. Define

$$\begin{aligned} h_{1,0}(u) &= E\{h(u, V) \mid D = 1\}; \\ h_{0,1}(v) &= E\{h(U, v) \mid D = 0\}. \end{aligned}$$

Then, from Chapter 12.2 of Van der Vaart (1998),

$$n^{1/2}U_n = n^{1/2}n_0^{-1}\sum_{i=1}^{n_0} I(D_i = 0)h_{1,0}(U_i) + n^{1/2}n_1^{-1}\sum_{j=1}^{n_1} I(D_j = 1)h_{0,1}(V_j) + o_p(1). \tag{A.4}$$

Preliminary Lemma

Let the data be $Z_i = (D_i, G_i, X_i)$ for $i = 1, \ldots, n$, ordered so that the first n_0 observations are the controls, and the last n_1 observations are the cases.

Define $n_d = c_d n$.

In the proofs, for generic functions $T(\cdot)$ and $P(\cdot)$, we need to deal with terms

$$\begin{aligned}\mathcal{D}_n(P,T) &= \sum_{d=0}^{1}(\pi_d/n_d)n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{r=0}^{1}I(D_j=d)\\ &\qquad \times P(X_i)\left\{T(r,G_j,X_i)-T_E(r,D_j,X_i)\right\}\\ &= \sum_{t=0}^{1}\sum_{d=0}^{1}(\pi_d/n_d)n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{r=0}^{1}I(D_i=t,D_j=d)\\ &\qquad \times P(X_i)\left\{T(r,G_j,X_i)-T_E(r,d,X_i)\right\},\end{aligned}$$

where

$$T_E(r,d,x)=E\{T(r,G,x)\mid D=d\}.$$

We will use repeatedly the fact that for any constant x,

$$0=E\left[P(x)\left\{T(r,G,x)-T_E(r,d,x)\right\}\mid D=d\right]. \tag{A.5}$$

We will make the following notational convention. We define

$$E\left[P(X)\left\{T(r,g_i,X)-T_E(r,d,X)\right\}\mid D=t\right] \tag{A.6}$$

to mean

$$E\left[P(X)\left\{T(r,g,X)-T_E(r,d,X)\right\}\mid D=t\right]_{g=G_i}.$$

Similarly, $E\left[P(x_i)\left\{T(r,G,x_i)-T_E(r,d,x_i)\right\}\mid D=t\right]$ is

$$E\left[P(x)\left\{T(r,G,x)-T_E(r,d,x)\right\}\mid D=t\right]_{x=X_i}.$$

Below, we will prove the following Lemma, which relies of U-statistics of order 2 for one sample and U-statistics of order 1 for independent samples, namely the cases and the controls. We

use the notation defined at (A.6).

Lemma 1. *Define* $Z_i = (D_i, G_i, X_i)$. *As* $n \to \infty$ *in such a way that* $n_d = c_d n$ *for* $0 < c_0, c_1 < 1$,

$$
\begin{aligned}
& n^{1/2}\mathcal{D}_n(P,T) \\
& \quad = n^{-1/2}\textstyle\sum_{i=1}^{n_0}\sum_{d=0}^{1}\sum_{r=0}^{1}\{c_d\pi_{d_i}/c_{d_i}\}E\{P(X)T(r,g_i,X) \mid D=d\} \\
& \qquad -n^{-1/2}n_0 E\left[P(X)\left\{\pi_0\textstyle\sum_{r=0}^{1}T_E(r,0,X)+\pi_1\sum_{r=0}^{1}T_E(r,1,X)\right\} \mid D=0\right] \\
& \qquad -n^{-1/2}n_1 E\left[P(X)\left\{\pi_0\textstyle\sum_{r=0}^{1}T_E(r,0,X)+\pi_1\sum_{r=0}^{1}T_E(r,1,X)\right\} \mid D=1\right] + o_p(1).
\end{aligned}
$$

Proof of Lemma 1

Now, since there are only n terms with $i = j$, whereas the leading terms before the summations are $O(n^{-2})$, and because $(n-1)^{-1} - n^{-1} = O(n^{-2})$, and because the first n_0 observations are controls, to order $n^{1/2}$, analyzing D_n is equivalent to analyzing

$$\mathcal{D}_n(P,T) = \textstyle\sum_{t=0}^{1}\sum_{d=0}^{1}\mathcal{D}_n(P,T,t,d) + o_p(n^{-1}),$$

where

$$
\begin{aligned}
\mathcal{D}_n(P,T,0,0) &= (\pi_0/n_0)n^{-1}\textstyle\sum_{i=1}^{n_0}\sum_{j=1,j\neq i}^{n_0} I(D_i=0,D_j=0) \\
&\qquad \times P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,0,X_i)\} \\
&= \{n_0(n_0-1)\}^{-1}\textstyle\sum_{i=1}^{n_0}\sum_{j=1,j\neq i}^{n_0} I(D_i=0,D_j=0) \\
&\qquad \times(\pi_0 c_0)P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,0,X_i)\} + O_p(n^{-1});
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{D}_n(P,T,0,1) &= (\pi_1/n_1)n^{-1}\textstyle\sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n} I(D_i=0,D_j=1) \\
&\qquad \times P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,1,X_i)\} \\
&= (n_0 n_1)^{-1}\textstyle\sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n} I(D_i=0,D_j=1) \\
&\qquad \times(\pi_1 c_0)P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,1,X_i)\};
\end{aligned}
$$

$$\begin{aligned}
\mathcal{D}_n(P,T,1,0) &= (\pi_0/n_0)n^{-1}\textstyle\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0} I(D_i=1,D_j=0) \\
&\qquad \times P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,0,X_i)\} \\
&= (n_0n_1)^{-1}\textstyle\sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n} I(D_i=0,D_j=1) \\
&\qquad \times(\pi_0c_1)P(X_j)\textstyle\sum_{r=0}^{1}\{T(r,G_i,X_j)-T_E(r,0,X_j)\}\,;
\end{aligned}$$

$$\begin{aligned}
\mathcal{D}_n(P,T,1,1) &= (\pi_1/n_1)n^{-1}\textstyle\sum_{i=n_0+1}^{n}\sum_{j=n_0+1,j\neq i}^{n} I(D_i=1,D_j=1) \\
&\qquad \times P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,1,X_i)\} \\
&= \{n_1(n_1-1)\}^{-1}\textstyle\sum_{i=n_0+1}^{n}\sum_{j=n_0+1,j\neq i}^{n} I(D_i=1,D_j=1) \\
&\qquad \times(\pi_1c_1)P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,1,X_i)\}+O_p(n^{-1}).
\end{aligned}$$

Now, $\mathcal{D}_n(P,T,1,0)$ and $\mathcal{D}_n(P,T,0,1)$ are U-statistics of order 1 for 2 independent samples, while $\mathcal{D}_n(P,T,0,0)$ and $\mathcal{D}_n(P,T,1,1)$ are U-statistics of order 2 for a single sample, all with asymmetric kernels.

We next analyze $\mathcal{D}_n(P,T,0,1)$. The term $\mathcal{D}_n(P,T,0,1)$ has kernel

$$h(Z_i,Z_j,0,1) = (\pi_1c_0)P(X_i)\textstyle\sum_{r=0}^{1}\{T(r,G_j,X_i)-T_E(r,1,X_i)\}\,.$$

Then

$$\begin{aligned}
h_{1,0}(u,0,1) &= E\{h(u,Z_j,0,1)\mid D_j=1\}=0, \text{ by (A.5)}; \\
h_{0,1}(v,0,1) &= E\{h(Z_i,v)\mid D_i=0\} \\
&= (\pi_1c_0)E\left[P(X)\textstyle\sum_{r=0}^{1}\{T(r,v,X)-T_E(r,1,X)\}\mid D=0\right].
\end{aligned}$$

Thus, from (A.4),

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,0,1) &= (n^{1/2}/n_1)\textstyle\sum_{j=n_0+1}^{n} h_{0,1}(Z_j,0,1) + o_p(1) \\ &= n^{-1/2}\textstyle\sum_{j=n_0+1}^{n} c_1^{-1} h_{0,1}(Z_j,0,1) + o_p(1). \end{aligned} \tag{A.7}$$

In the notation defined at (A.6),

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,0,1) &= n^{-1/2}\textstyle\sum_{j=n_0+1}^{n} (\pi_1 c_0/c_1) I(D_j = 1) \\ &\qquad \times \textstyle\sum_{r=0}^{1} E\left[P(X)\left\{T(r,g_j,X) - T_E(r,1,X)\right\} \mid D=0\right] + o_p(1) \\ &= n^{-1/2}\textstyle\sum_{i=n_0+1}^{n} I(D_i = 1) \\ &\qquad \times (\pi_1 c_0/c_1)\textstyle\sum_{r=0}^{1} E\left[P(X)\left\{T(r,g_i,X) - T_E(r,1,X)\right\} \mid D=0\right] + o_p(1). \end{aligned} \tag{A.8}$$

Now consider he term $\mathcal{D}_n(P,T,1,0)$, which has kernel

$$h(Z_i,Z_j,1,0) = (\pi_0 c_1) P(X_j) \textstyle\sum_{r=0}^{1} \left\{T(r,G_i,X_j) - T_E(r,0,X_j)\right\}.$$

Then

$$\begin{aligned} h_{1,0}(u,1,0) &= E\{h(u,Z_j,1,0) \mid D_j = 1\} \\ &= (\pi_0 c_1) E\left[P(X_j)\textstyle\sum_{r=0}^{1}\left\{T(r,u,X_j) - T_E(r,0,X_j)\right\} \mid D_j = 1\right]; \\ h_{0,1}(v,1,0) &= E\{h(Z_i,v,1,0) \mid D_i = 0\} = 0, \text{ by (A.5).} \end{aligned}$$

Thus, from (A.4),

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,1,0) &= (n^{1/2}/n_0)\textstyle\sum_{i=1}^{n_0} h_{1,0}(Z_i,1,0) + o_p(1) \\ &= n^{-1/2}\textstyle\sum_{i=1}^{n_0} c_0^{-1} h_{1,0}(Z_j,1,0) + o_p(1). \end{aligned} \tag{A.9}$$

In the notation defined at (A.6),

$$n^{1/2}\mathcal{D}_n(P,T,1,0) = n^{-1/2}\textstyle\sum_{i=1}^{n_0} I(D_i=0)(\pi_0 c_1/c_0) \times \sum_{r=0}^{1} E\left[P(X)\left\{T(r,g_i,X)-T_E(r,0,X)\right\} \mid D=1\right] + o_p(1). \quad \text{(A.10)}$$

We next analyze $\mathcal{D}_n(P,T,0,0)$, which is a U-statistic of order 2 but with an asymmetric kernel

$$h_*(Z_i,Z_j,0,0) = I(D_i=D_j=0)(\pi_0 c_0)P(X_i)\textstyle\sum_{r=0}^{1}\left\{T(r,G_j,X_i)-T_E(r,0,X_i)\right\}.$$

To make this a symmetric kernel, we define

$$\begin{aligned} h(Z_i,Z_j,0,0) &= (1/2)I(D_i=D_j=0)(\pi_0 c_0)\big[P(X_i)\textstyle\sum_{r=0}^{1}\left\{T(r,G_j,X_i)-T_E(r,0,X_i)\right\} \\ &\quad + P(X_j)\textstyle\sum_{r=0}^{1}\left\{T(r,G_i,X_j)-T_E(r,0,X_j)\right\}\big]. \end{aligned}$$

We now apply (A.1), so that

$$\begin{aligned} h_1(z,0,0) &= (\pi_0 c_0)E\left[P(x)\textstyle\sum_{r=0}^{1}\left\{T(r,G,x)-T_E(r,0,x)\right\} \mid D=0\right] \\ &\quad + (\pi_0 c_0)E\left[P(X)\textstyle\sum_{r=0}^{1}\left\{T(r,g,X)-T_E(r,0,X)\right\} \mid D=0\right] \\ &= (\pi_0 c_0)E\left[P(X)\textstyle\sum_{r=0}^{1}\left\{T(r,g,X)-T_E(r,0,X)\right\} \mid D=0\right], \text{ by (A.5).} \end{aligned}$$

From (A.2), this means that

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,0,0) &= (n^{1/2}/n_0)\textstyle\sum_{i=1}^{n_0} h_1(Z_i,0,0) + o_p(1) \\ &= (n^{-1/2}/c_0)\textstyle\sum_{i=1}^{n_0} h_1(Z_i,0,0) + o_p(1). \end{aligned}$$

Thus, in the notation defined at (A.6),

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,0,0) &= n^{-1/2}\textstyle\sum_{i=1}^{n_0} I(D_i=0)\pi_0 \\ &\quad \times \textstyle\sum_{r=0}^{1} E\left[P(X)\left\{T(r,g_i,X)-T_E(r,0,X)\right\} \mid D=0\right] + o_p(1). \end{aligned} \tag{A.11}$$

We next analyze $\mathcal{D}_n(P,T,1,1)$, which is a U-statistic of order 2 but with an asymmetric kernel

$$h_*(Z_i,Z_j,1,1) = I(D_i=D_j=1)(\pi_1 c_1)P(X_i)\textstyle\sum_{r=0}^{1}\left\{T(r,G_j,X_i)-T_E(r,1,X_i)\right\}.$$

To make this a symmetric kernel, we define

$$\begin{aligned} h(Z_i,Z_j,1,1) &= (1/2)I(D_i=D_j=1)(\pi_1 c_1)\big[P(X_i)\textstyle\sum_{r=0}^{1}\left\{T(r,G_j,X_i)-T_E(r,1,X_i)\right\} \\ &\quad + P(X_j)\textstyle\sum_{r=0}^{1}\left\{T(r,G_i,X_j)-T_E(r,1,X_j)\right\}\big]. \end{aligned}$$

We now apply (A.1), so that

$$\begin{aligned} h_1(z,1,1) &= (\pi_1 c_1)E\left[P(x)\textstyle\sum_{r=0}^{1}\left\{T(r,G,x)-T_E(r,1,x)\right\} \mid D=1\right] \\ &\quad + (\pi_1 c_1)E\left[P(X)\textstyle\sum_{r=0}^{1}\left\{T(r,g,X)-T_E(r,1,X)\right\} \mid D=1\right] \\ &= (\pi_1 c_1)E\left[P(X)\textstyle\sum_{r=0}^{1}\left\{T(r,g,X)-T_E(r,1,X)\right\} \mid D=1\right], \text{ by (A.5).} \end{aligned}$$

From (A.2),

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,1,1) &= (n^{1/2}/n_1)\textstyle\sum_{i=n_0+1}^{n} h_1(Z_i,1,1) + o_p(1) \\ &= (n^{-1/2}/c_1)\textstyle\sum_{i=n_0+1}^{n} h_1(Z_i,1,1) + o_p(1). \end{aligned}$$

Thus, in the notation at (A.6),

$$\begin{aligned} n^{1/2}\mathcal{D}_n(P,T,1,1) &= n^{-1/2}\textstyle\sum_{i=n_0+1}^{n} I(D_i=1) \\ &\quad \times \pi_1 E\left[P(X)\textstyle\sum_{r=0}^{1}\{T(r,g_i,X)-T_E(r,1,X)\} \mid D=1\right]+o_p(1). \end{aligned} \tag{A.12}$$

Collecting the terms (A.8), (A.10), (A.11) and (A.12), we get that

$$\begin{aligned} & n^{1/2}\mathcal{D}_n(P,T) \\ = \;& n^{-1/2}\textstyle\sum_{i=1}^{n_0} I(D_i=0)\pi_0\sum_{r=0}^{1} E\left[P(X)\{T(r,g_i,X)-T_E(r,0,X)\} \mid D=0\right] \\ & +n^{-1/2}\textstyle\sum_{i=1}^{n_0} I(D_i=0)(\pi_0 c_1/c_0)\sum_{r=0}^{1} E\left[P(X)\{T(r,g_i,X)-T_E(r,0,X)\} \mid D=1\right] \\ & +n^{-1/2}\textstyle\sum_{i=n_0+1}^{n} I(D_i=1)\pi_1 E\left[P(X)\sum_{r=0}^{1}\{T(r,g_i,X)-T_E(r,1,X)\} \mid D=1\right] \\ & +n^{-1/2}\textstyle\sum_{i=n_0+1}^{n} I(D_i=1)(\pi_1 c_0/c_1)\sum_{r=0}^{1} E\left[P(X)\{T(r,g_i,X)-T_E(r,1,X)\} \mid D=0\right] \\ & +o_p(1). \end{aligned}$$

This in turn is seen to be

$$n^{1/2}\mathcal{D}_n(P,T) = \mathcal{G}_1 - \mathcal{G}_2 + o_p(1),$$

where

$$\begin{aligned} \mathcal{G}_1(P,T) &= n^{-1/2}\textstyle\sum_{i=1}^{n}\sum_{d=0}^{1}\sum_{r=0}^{1}(c_d\pi_{d_i}/c_{d_i})E\{P(X)T(r,g_i,X) \mid D=d\}; \\ \mathcal{G}_2(P,T) &= n^{-1/2}\textstyle\sum_{i=n_0+1}^{n}\sum_{r=0}^{1} I(D_i=0)\pi_0 E\{P(X)T_E(r,0,X) \mid D=0\} \\ &\quad +n^{-1/2}\textstyle\sum_{i=1}^{n_0}\sum_{r=0}^{1} I(D_i=0)(\pi_0 c_1/c_0)E\{P(X)T_E(r,0,X) \mid D=1\} \\ &\quad +n^{-1/2}\textstyle\sum_{i=n_0+1}^{n}\sum_{r=0}^{1} I(D_i=1)\pi_1 E\{P(X)T_E(r,1,X) \mid D=1\} \\ &\quad +n^{-1/2}\textstyle\sum_{i=n_0+1}^{n}\sum_{r=0}^{1} I(D_i=1)(\pi_1 c_0/c_1)\sum_{r=0}^{1} E\{P(X)T_E(r,1,X) \mid D=0\}. \end{aligned}$$

It is easily seen that

$$\begin{aligned}
\mathcal{G}_2(P,T) &= n^{-1/2}n_0E\left[P(X)\left\{\pi_0\textstyle\sum_{r=0}^{1}T_E(r,0,X)\right\} \mid D=0\right] \\
&\quad +n^{-1/2}n_1E\left[P(X)\left\{\pi_0\textstyle\sum_{r=0}^{1}T_E(r,0,X)\right\} \mid D=1\right] \\
&\quad +n^{-1/2}n_1E\left[P(X)\left\{\pi_1\textstyle\sum_{r=0}^{1}T_E(r,1,X)\right\} \mid D=1\right] \\
&\quad +n^{-1/2}n_0E\left[P(X)\left\{\pi_1\textstyle\sum_{r=0}^{1}T_E(r,1,X)\right\} \mid D=0\right] \\
&= n^{-1/2}n_0E\left[P(X)\left\{\pi_0\textstyle\sum_{r=0}^{1}T_E(r,0,X)+\pi_1\sum_{r=0}^{1}T_E(r,1,X)\right\} \mid D=0\right] \\
&\quad +n^{-1/2}n_1E\left[P(X)\left\{\pi_0\textstyle\sum_{r=0}^{1}T_E(r,0,X)+\pi_1\sum_{r=0}^{1}T_E(r,1,X)\right\} \mid D=1\right].
\end{aligned}$$

This completes the proof of Lemma 1.

Proof of Theorem 1

With a first-order Taylor series expansion, it is readily seen that

$$n^{-1/2}\textstyle\sum_{i=1}^{n}\left\{\dfrac{S_\Omega(D_i,G_i,X_i,\widehat{\Omega})}{S(D_i,G_i,X_i,\widehat{\Omega})}-\dfrac{S_\Omega(D_i,G_i,X_i,\Omega)}{S(D_i,G_i,X_i,\Omega)}\right\}=\Gamma_1 n^{1/2}(\widehat{\Omega}-\Omega)+o_p(1).$$

Similarly,

$$n^{-1/2}\textstyle\sum_{i=1}^{n}\left\{\dfrac{\widehat{R}_\Omega(X_i,\widehat{\Omega})}{\widehat{R}(X_i,\widehat{\Omega})}-\dfrac{\widehat{R}_\Omega(X_i,\Omega)}{\widehat{R}(X_i,\Omega)}\right\}=\Gamma_2 n^{1/2}(\widehat{\Omega}-\Omega)+o_p(1).$$

In a manner similar to that of Wei et al. (2013), we have that

$$\begin{aligned}
0 &= \widehat{\mathcal{S}}_n(\widehat{\Omega})=\widehat{\mathcal{S}}_n(\Omega)+n^{-1/2}\frac{\partial\widehat{\mathcal{S}}_n(\Omega)}{\partial\Omega^{\mathrm{T}}}n^{1/2}(\widehat{\Omega}-\Omega)+o_p(1) \\
&= \widehat{\mathcal{S}}_n(\Omega)+(\Gamma_1-\Gamma_2)n^{1/2}(\widehat{\Omega}-\Omega)+o_p(1) \\
&= \mathcal{S}_n(\Omega)-n^{-1/2}\textstyle\sum_{i=1}^{n}\left\{\dfrac{\widehat{R}_\Omega(X_i,\Omega)}{\widehat{R}(X_i,\Omega)}-\dfrac{R_\Omega(X_i,\Omega)}{R(X_i,\Omega)}\right\} \\
&\qquad +(\Gamma_1-\Gamma_2)n^{1/2}(\widehat{\Omega}-\Omega)+o_p(1). \qquad \text{(A.13)}
\end{aligned}$$

We now analyze the second term in (A.13), which equals

$$
\begin{aligned}
& n^{-1/2}\textstyle\sum_{i=1}^{n}\left[\dfrac{\widehat{R}_{\Omega}(X_i,\Omega)-R_{\Omega}(X_i,\Omega)}{R(X_i,\Omega)}-\dfrac{R_{\Omega}(X_i,\Omega)\{\widehat{R}(X_i,\Omega)-R(X_i,\Omega)\}}{R^2(X_i,\Omega)}\right]+o_p(1) \\
= \quad & n^{-1/2}\textstyle\sum_{i=1}^{n}P_1(X_i,\Omega)\{\widehat{R}_{\Omega}(X_i,\Omega)-R_{\Omega}(X_i,\Omega)\} \\
& -n^{-1/2}\textstyle\sum_{i=1}^{n}P_2(X_i,\Omega)\{\widehat{R}(X_i,\Omega)-R(X_i,\Omega)\}+o_p(1).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathcal{C}_n &= n^{-1/2}\textstyle\sum_{i=1}^{n}\{\dfrac{\widehat{R}_{\Omega}(X_i,\Omega)}{\widehat{R}(X_i,\Omega)}-\dfrac{R_{\Omega}(X_i,\Omega)}{R(X_i,\Omega)}\} \\
&= n^{-1/2}\textstyle\sum_{i=1}^{n}\dfrac{\widehat{R}_{\Omega}(X_i,\Omega)-R_{\Omega}(X_i,\Omega)}{R(X_i,\Omega)} \\
&\qquad -n^{-1/2}\textstyle\sum_{i=1}^{n}\dfrac{R_{\Omega}(X_i,\Omega)}{R^2(X_i,\Omega)}\{\widehat{R}(X_i,\Omega)-R(X_i,\Omega)\}+o_p(1) \\
&= \mathcal{C}_{n1}-\mathcal{C}_{n2}+o_p(1).
\end{aligned}
$$

First, we calculate that

$$
\begin{aligned}
\widehat{R}(x,\Omega)-R(x,\Omega) &= \textstyle\sum_{j=1}^{J}\sum_{r=0}^{1}\sum_{d=0}^{1}(\pi_d/n_d)I(D_j=d)S(r,G_j,x,\Omega) \\
&\qquad -\textstyle\sum_{r=0}^{1}\sum_{d=0}^{1}\pi_d S_E(r,d,x,\Omega) \\
&= \textstyle\sum_{j=1}^{J}\{\sum_{r=0}^{1}\sum_{d=0}^{1}(\pi_d/n_d)I(D_j=d)S(r,G_j,x,\Omega) \\
&\qquad -\textstyle\sum_{r=0}^{1}\sum_{d=0}^{1}(\pi_d/n_d)I(D_j=d)S_E(r,d,x,\Omega)\} \\
&= \textstyle\sum_{d=0}^{1}n_d^{-1}\sum_{j=1}^{J}\sum_{r=0}^{1}I(D_j=d)\pi_d \\
&\qquad \times\{S(r,G_j,x,\Omega)-S_E(r,d,x,\Omega)\}. \qquad \text{(A.14)}
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\widehat{R}_{\Omega}(x,\Omega)-R_{\Omega}(x,\Omega) &= \textstyle\sum_{d=0}^{1}n_d^{-1}\sum_{j=1}^{J}\sum_{r=0}^{1}I(D_j=d)\pi_d \\
&\qquad \times\{S_{\Omega}(r,G_j,x,\Omega)-S_{E,\Omega}(r,d,x,\Omega)\}. \qquad \text{(A.15)}
\end{aligned}
$$

Then, from (A.14) and (A.15),

$$\begin{aligned}
\mathcal{C}_{n1} &= n^{-1/2}\textstyle\sum_{i=1}^{n} P_1(X_i,\Omega)\{\widehat{R}_\Omega(X_i,\Omega) - R_\Omega(X_i,\Omega)\} \\
&= \textstyle\sum_{d=0}^{1}(\pi_d/n_d)n^{-1/2}\sum_{i=1}^{n} P_1(X_i,\Omega) \\
&\qquad \times\textstyle\sum_{j=1}^{J}\sum_{r=0}^{1} I(D_j = d)\left\{S_\Omega(r,G_j,X_i,\Omega) - S_{E,\Omega}(r,D_j,X_i,\Omega)\right\}; \\
\mathcal{C}_{n2} &= n^{-1/2}\textstyle\sum_{i=1}^{n} P_2(X_i,\Omega)\{\widehat{R}(X_i,\Omega) - R(X_i,\Omega)\} \\
&= \textstyle\sum_{d=0}^{1}(\pi_d/n_d)n^{-1/2}\sum_{i=1}^{n} P_2(X_i,\Omega) \\
&\qquad \times\textstyle\sum_{j=1}^{J}\sum_{r=0}^{1} I(D_j = d)\left\{S(r,G_j,X_i,\Omega) - S_E(r,D_j,X_i,\Omega)\right\}.
\end{aligned}$$

In the notation defined at (A.6),

$$\begin{aligned}
\mathcal{C}_{n1} &= n^{1/2}\mathcal{D}_n(P_1, S_\Omega) + o_p(1); \\
\mathcal{C}_{n2} &= n^{1/2}\mathcal{D}_n(P_2, S) + o_p(1).
\end{aligned}$$

Thus, with Lemma 1,

$$
\begin{aligned}
\mathcal{C}_n &= \mathcal{C}_{n1} - \mathcal{C}_{n2} + o_p(1) \\
&= n^{-1/2}\textstyle\sum_{i=1}^{n}\sum_{d=0}^{1}\sum_{r=0}^{1}\frac{c_d\pi_{d_i}}{c_{d_i}}E\left[\{P_1(X,\Omega)S_\Omega(r,g_i,X) - P_2(X,\Omega)S(r,g_i,X)\} \mid D=d\right] \\
&\quad -n^{-1/2}n_0E\left[P_1(X,\Omega)\left\{\pi_0\textstyle\sum_{r=0}^{1}S_{E,\Omega}(r,0,X) + \pi_1\sum_{r=0}^{1}S_{E,\Omega}(r,1,X)\right\} \mid D=0\right] \\
&\quad -n^{-1/2}n_1E\left[P_1(X,\Omega)\left\{\pi_0\textstyle\sum_{r=0}^{1}S_{E,\Omega}(r,0,X) + \pi_1\sum_{r=0}^{1}S_{E,\Omega}(r,1,X)\right\} \mid D=1\right] \\
&\quad +n^{-1/2}n_0E\left[P_2(X,\Omega)\left\{\pi_0\textstyle\sum_{r=0}^{1}S_{E}(r,0,X) + \pi_1\sum_{r=0}^{1}S_{E}(r,1,X)\right\} \mid D=0\right] \\
&\quad +n^{-1/2}n_1E\left[P_2(X,\Omega)\left\{\pi_0\textstyle\sum_{r=0}^{1}S_{E}(r,0,X) + \pi_1\sum_{r=0}^{1}S_{E}(r,1,X)\right\} \mid D=1\right] + o_p(1) \\
&= n^{-1/2}\textstyle\sum_{i=1}^{n}\sum_{d=0}^{1}\sum_{r=0}^{1}\frac{c_d\pi_{d_i}}{c_{d_i}}E\left[\{P_1(X,\Omega)S_\Omega(r,g_i,X) - P_2(X,\Omega)S(r,g_i,X)\} \mid D=d\right] \\
&\quad -n^{-1/2}n_0E\left\{P_1(X,\Omega)R_\Omega(X,\Omega) \mid D=0\right\} \\
&\quad -n^{-1/2}n_1E\left\{P_1(X,\Omega)R_\Omega(X,\Omega) \mid D=1\right\} \\
&\quad +n^{-1/2}n_0E\left\{P_2(X,\Omega)R(X,\Omega) \mid D=0\right\} \\
&\quad +n^{-1/2}n_1E\left\{P_2(X,\Omega)R(X,\Omega) \mid D=1\right\} + o_p(1).
\end{aligned}
$$

However,

$$
\begin{aligned}
P_1(X,\Omega)R_\Omega(X,\Omega) &= \{R(X,\Omega)\}^{-1}R_\Omega(X,\Omega); \\
P_2(X,\Omega)R(X,\Omega) &= \{R(X,\Omega)\}^{-1}R_\Omega(X,\Omega),
\end{aligned}
$$

so the last 4 terms above cancel, completing the proof of Theorem 1.

A.2 Alternative Proof Based on a Hypothetical Population

Here we give an alternative argument using the hypothetical population framework of Ma (2010). Define $\mathcal{K}_1(D,G,X,\Omega) = S_\Omega(D,G,X,\Omega)/S(D,G,X,\Omega)$ and $\mathcal{K}_2(X,\Omega) = R_\Omega(X,\Omega)/R(X,\Omega)$. Solving (7) in the main paper leads to the expansion

$$
\begin{aligned}
\mathbf{0} &= n^{-1/2}\sum_{i=1}^{n}\left\{\mathcal{K}_1(D_i,G_i,X_i,\widehat{\Omega}) - \frac{\widehat{R}_\Omega(X_i,\widehat{\Omega})}{\widehat{R}(X_i,\widehat{\Omega})}\right\} \\
&= n^{-1/2}\sum_{i=1}^{n}\left\{\mathcal{K}_1(D_i,G_i,X_i,\Omega) - \frac{\widehat{R}_\Omega(X_i,\Omega)}{\widehat{R}(X_i,\Omega)}\right\} \\
&\quad + \left[n^{-1}\textstyle\sum_{i=1}^{n}\partial\left\{\mathcal{K}_1(D_i,G_i,X_i,\Omega) - \frac{\widehat{R}_\Omega(X_i,\Omega)}{\widehat{R}(X_i,\Omega)}\right\}/\partial\Omega^{\mathrm{T}} + o_p(1)\right]\sqrt{n}(\widehat{\Omega}-\Omega) \\
&= n^{-1/2}\sum_{i=1}^{n}\left[\mathcal{K}_1(D_i,G_i,X_i,\Omega) - \mathcal{K}_2(X_i,\Omega) - \frac{\widehat{R}_\Omega(X_i,\Omega) - R_\Omega(X_i,\Omega)}{R(X_i,\Omega)}\right. \\
&\quad \left. + \frac{R_\Omega(X_i,\Omega)}{R^2(X_i,\Omega)}\left\{\widehat{R}(X_i,\Omega) - R(X_i,\Omega)\right\}\right] + (\Gamma_1-\Gamma_2)\sqrt{n}(\widehat{\Omega}-\Omega) + o_p(1) \\
&= n^{-1/2}\sum_{i=1}^{n}\left\{\mathcal{K}_1(D_i,G_i,X_i,\Omega) - \mathcal{K}_2(X_i,\Omega) - P_1(X_i,\Omega)\widehat{R}_\Omega(X_i,\Omega)\right. \\
&\quad \left. + P_2(X_i,\Omega)\widehat{R}(X_i,\Omega)\right\} + (\Gamma_1-\Gamma_2)\sqrt{n}(\widehat{\Omega}-\Omega) + o_p(1).
\end{aligned}
$$

Now using U-statistics properties,

$$
\begin{aligned}
& n^{-1/2}\textstyle\sum_{i=1}^{n}\left\{P_1(X_i,\Omega)\widehat{R}_\Omega(X_i,\Omega) - P_2(X_i,\Omega)\widehat{R}(X_i,\Omega)\right\} \\
= \;& \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1} n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{J}\frac{\pi_d}{c_d}I(D_j=d)\left\{P_1(X_i,\Omega)S_\Omega(r,G_j,X_i,\Omega)\right. \\
& \left. - P_2(X_i,\Omega)S(r,G_j,X_i,\Omega)\right\} \\
= \;& \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1} n^{-1/2}\sum_{i=1}^{n} E\left[\frac{\pi_d}{c_d}I(D=d)\left\{P_1(x_i,\Omega)S_\Omega(r,G,x_i,\Omega) - P_2(x_i,\Omega)S(r,G,x_i,\Omega)\right\}\right] \\
& + \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1} n^{-1/2}\sum_{j=1}^{J} E\left[\frac{\pi_d}{c_d}I(d_j=d)\left\{P_1(X,\Omega)S_\Omega(r,g_j,X,\Omega) - P_2(X,\Omega)S(r,g_j,X,\Omega)\right\}\right] \\
& - \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1} n^{-1/2}\sum_{j=1}^{J} E\left[\frac{\pi_d}{c_d}I(D_j=d)\left\{P_1(X,\Omega)S_\Omega(r,G_j,X,\Omega) - P_2(X,\Omega)S(r,G_j,X,\Omega)\right\}\right] \\
& + o_p(1).
\end{aligned}
$$

Further, we thus have that

$$
\begin{aligned}
& n^{-1/2}\textstyle\sum_{i=1}^{n}\left\{P_1(X_i,\Omega)\widehat{R}_\Omega(X_i,\Omega)-P_2(X_i,\Omega)\widehat{R}(X_i,\Omega)\right\} \\
= \;& \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1}n^{-1/2}\sum_{i=1}^{n}\pi_d\left\{P_1(X_i,\Omega)S_{E,\Omega}(r,d,X_i,\Omega)-P_2(X_i,\Omega)S_E(r,d,X_i,\Omega)\right\} \\
& +\textstyle\sum_{t=0}^{1}\sum_{r=0}^{1}\sum_{d=0}^{1}n^{-1/2}\sum_{i=1}^{n}\frac{\pi_d c_t}{c_d}I(d_i=d) \\
& \times E\left[\left\{P_1(X,\Omega)S_\Omega(r,g_i,X,\Omega)-P_2(X,\Omega)S(r,g_i,X,\Omega)\right\}\mid D=t\right] \\
& -\textstyle\sum_{t=0}^{1}\sum_{r=0}^{1}\sum_{d=0}^{1}n^{1/2}\pi_d c_t E\left\{P_1(X,\Omega)S_{E,\Omega}(r,d,X,\Omega)-P_2(X,\Omega)S_E(r,d,X,\Omega)\mid D=t\right\} \\
& +o_p(1).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& n^{-1/2}\textstyle\sum_{i=1}^{n}\left\{P_1(X_i,\Omega)\widehat{R}_\Omega(X_i,\Omega)-P_2(X_i,\Omega)\widehat{R}(X_i,\Omega)\right\} \\
= \;& \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1}n^{-1/2}\sum_{i=1}^{n}\pi_d\left\{P_1(X_i,\Omega)S_{E,\Omega}(r,d,X_i,\Omega)-P_2(X_i,\Omega)S_E(r,d,X_i,\Omega)\right\} \\
& +\textstyle\sum_{d=0}^{1}\sum_{r=0}^{1}n^{-1/2}\sum_{i=1}^{n}\frac{\pi_{d_i} c_d}{c_{d_i}}E\left[\left\{P_1(X,\Omega)S_\Omega(r,g_i,X,\Omega)-P_2(X,\Omega)S(r,g_i,X,\Omega)\right\}\mid D=d\right] \\
& -\textstyle\sum_{t=0}^{1}\sum_{r=0}^{1}\sum_{d=0}^{1}n^{1/2}\pi_d c_t E\left\{P_1(X,\Omega)S_{E,\Omega}(r,d,X,\Omega)-P_2(X,\Omega)S_E(r,d,X,\Omega)\mid D=t\right\} \\
& +o_p(1) \\
= \;& \textstyle\sum_{d=0}^{1}\sum_{r=0}^{1}n^{-1/2}\sum_{i=1}^{n}\frac{\pi_{d_i} c_d}{c_{d_i}}E\left[\left\{P_1(X,\Omega)S_\Omega(r,g_i,X,\Omega)-P_2(X,\Omega)S(r,g_i,X,\Omega)\right\}\mid D=d\right] \\
& +o_p(1).
\end{aligned}
$$

Here the last step is because for any X,

$$
\begin{aligned}
& \textstyle\sum_{r=0}^{1}\sum_{d=0}^{1}\pi_d\left\{P_1(X,\Omega)S_{E,\Omega}(r,d,X,\Omega)-P_2(X,\Omega)S_E(r,d,X,\Omega)\right\} \\
= \;& \frac{1}{R(X,\Omega)}\textstyle\sum_{r=0}^{1}\sum_{d=0}^{1}\pi_d S_{E,\Omega}(r,d,X,\Omega)-\frac{R_\Omega(X,\Omega)}{R^2(X,\Omega)}\sum_{r=0}^{1}\sum_{d=0}^{1}\pi_d S_E(r,d,X,\Omega) \\
= \;& \frac{R_\Omega(X,\Omega)}{R(X,\Omega)}-\frac{R(X,\Omega)R_\Omega(X,\Omega)}{R^2(X,\Omega)} \\
= \;& \mathbf{0}.
\end{aligned}
$$

This leads to the result.

A.3 Score and Hessian: Rare Disease Case of §2.2 in the Main Paper

We consider models in which $\kappa + m(g, x, \boldsymbol{\beta}) = Q^{\mathrm{T}}(g, x)\Omega$, and $\Omega = (\kappa, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$. The point of this section is to show that both the log-pseudolikelihood score and its Hessian are very simply calculated, and that the Hessian is negative semidefinite.

In the rare disease case,

$$S(d, g, x, \Omega) = \exp\{dQ^{\mathrm{T}}(g, x)\Omega\}, \tag{A.16}$$

and thus

$$\log\{S(d, g, x, \Omega)\} = dQ^{\mathrm{T}}(g, x)\Omega.$$

This means that

$$\partial \log\{S(d, g, x, \Omega)\}/\partial\Omega = dQ(g, x),$$

and also that

$$\partial^2 \log\{S(d, g, x, \Omega)\}\partial\Omega\partial\Omega^{\mathrm{T}} = 0.$$

Similarly, in the rare disease case,

$$\widehat{R}(X, \Omega) = n_0^{-1}\textstyle\sum_{j=1}^{J}\sum_{r=0}^{1} I(D_j = 0)S(r, G_j, X, \Omega).$$

From (A.16),

$$\begin{aligned}\widehat{R}_{\Omega}(X, \Omega) = \partial\widehat{R}(X, \Omega)/\partial\Omega &= n_0^{-1}\textstyle\sum_{j=1}^{J}\sum_{r=0}^{1} I(D_j = 0)S(r, G_j, X, \Omega)rQ(G_j, X) \\ &= n_0^{-1}\textstyle\sum_{j=1}^{J} I(D_j = 0)S(1, G_j, X, \Omega)Q(G_j, X). \end{aligned} \tag{A.17}$$

Thus,

$$\begin{aligned}\widehat{R}_{\Omega\Omega}(X,\Omega) &= \partial^2\widehat{R}(X,\Omega)/\partial\Omega\partial\Omega^{\mathrm{T}} \\ &= n_0^{-1}\textstyle\sum_{j=1}^{J} I(D_j=0)S(1,G_j,X,\Omega)Q(G_j,X)Q^{\mathrm{T}}(G_j,X). \qquad \text{(A.18)}\end{aligned}$$

This means that the Hessian for the log-pseudolikelihood in equation (6) of the main paper is

$$\begin{aligned}-\frac{\partial\{\widehat{R}_{\Omega}(X,\Omega)/\widehat{R}(X,\Omega)\}}{\partial\Omega^{\mathrm{T}}} &= -\frac{\widehat{R}_{\Omega\Omega}(X,\Omega)}{\widehat{R}(X,\Omega)} + \frac{\widehat{R}_{\Omega}(X,\Omega)\widehat{R}_{\Omega}^{\mathrm{T}}(X,\Omega)}{\widehat{R}^2(X,\Omega)} \\ &= \{\widehat{R}(X,\Omega)\}^{-2}\left\{-\widehat{R}_{\Omega\Omega}(X,\Omega)\widehat{R}(X,\Omega) + \widehat{R}_{\Omega}(X,\Omega)\widehat{R}_{\Omega}^{\mathrm{T}}(X,\Omega)\right\}.\end{aligned}$$

Write $V_j = I(D_j = 0)S(1, G_j, X, \Omega)$. For matrices, define $A \leq B$ to be that $B - A$ is positive semidefinite. By Hölder's inequality

$$\begin{aligned}&\widehat{R}_{\Omega}(X,\Omega)\widehat{R}_{\Omega}^{\mathrm{T}}(X,\Omega) \\ &\quad = n_0^{-1}\textstyle\sum_{j=1}^{J} V_j Q(G_j,X) \times n_0^{-1}\sum_{j=1}^{J} V_j Q^{\mathrm{T}}(G_j,X) \\ &\quad \leq n_0^{-1}\textstyle\sum_{j=1}^{J} V_j Q(G_j,X)Q^{\mathrm{T}}(G_j,X) \times n_0^{-1}\sum_{j=1}^{J} V_j \\ &\quad = \widehat{R}_{\Omega\Omega}(X,\Omega)\widehat{R}(X,\Omega).\end{aligned}$$

Hence, the Hessian is negative semidefinite as claimed.

A.4 Stratification and the Independence Assumption

The assumption of gene-environment independence may not hold when there may exist underlying strata in the population, e.g. defined by ethnicity, across which the distribution of both genetic and environmental factors vary. In this case, as discussed in Section 3.1 of Chatterjee and Carroll (2005), we extend our framework to account for the scenario where the genetic and environmental factors can be assumed to be independent conditional on a discrete stratification $\mathcal{A}$ with $a = 1, ..., A$ levels.

To apply the method in Section 2.1 in the main paper to this case, for stratum a, we replace π_d by π_{da}, the probability that $D = d$ in the a^{th} stratum of the source population, and we replace n, n_0 and n_1 by n_a, n_{0a} and n_{1a}, the number of subjects, controls, and cases in stratum a, respectively. We modify (1) to $\text{pr}(D = 1|G, X, \mathcal{A} = a) = H\{\alpha_{0a} + m(G, X, \beta)\}$: more complex models with possible interactions between (G, X) and the strata can also be considered. We then set $\kappa_a = \alpha_{0a} + \log(n_{1a}/n_{0a}) - \log(\pi_{1a}/\pi_{0a})$. The parameters to be estimated are then $\Omega = (\kappa_1, ..., \kappa_A, \beta^{\mathrm{T}})^{\mathrm{T}}$. We also replace $S(d, g, x, \Omega)$ by

$$S_a(d, g, x, \Omega) = \frac{\exp[d\{\kappa_a + m(g, x, \beta)\}]}{1 + \exp\{\kappa_a + \log(\pi_{1a}/\pi_{0a}) - \log(n_{1a}/n_{0a}) + m(g, x, \beta)\}}.$$

Next, set $n = \sum_{a=1}^{A} n_a$, and replace (5) by

$$\widehat{R}_a(x, \Omega) = \sum_{j=1}^{J}\sum_{r=0}^{1}\sum_{d=0}^{1}(\pi_{da}/n_{da})I(D_j = d, \mathcal{A}_j = a)S_a(r, G_j, x, \Omega),$$

and the estimated loglikelihood (6) becomes

$$\mathcal{L}(\Omega) = \sum_{a=1}^{A} I(\mathcal{A}_i = a)[\sum_{i=1}^{n}\log\{S_a(D_i, G_i, X_i, \Omega)\} - \sum_{i=1}^{n}\log\{\widehat{R}_a(X_i, \Omega)\}],$$

which is then maximized to obtain the estimate $\widehat{\Omega}$. Now replace the score function (7) by

$$\widehat{\mathcal{S}}_n(\Omega) = n^{-1/2}\sum_{a=1}^{A}\sum_{i=1}^{n} I(\mathcal{A}_i = a)\left\{\frac{S_{\Omega,a}(D_i, G_i, X_i, \Omega)}{S_a(D_i, G_i, X_i, \Omega)} - \frac{\widehat{R}_{\Omega,a}(X_i, \Omega)}{\widehat{R}_a(X_i, \Omega)}\right\},$$

using the obvious definitions of $S_{\Omega,a}(\cdot)$, $\widehat{R}_{\Omega,a}(\cdot)$, $P_{1a}(X,\Omega)$, $P_{2a}(X,\Omega)$ and with $Z_i = (D_i, G_i, X_i, \mathcal{A}_i)$.

In terms of the asymptotic theory of Section 2.3 of the main paper, we replace (Γ_1, Γ_2) by

$$\begin{aligned}
\Gamma_1 &= \textstyle\sum_{a=1}^{A}\sum_{d=0}^{1}(n_{da}/n)E\left\{\dfrac{\partial S_{\Omega,a}(D,G,X,\Omega)/S_a(D,G,X,\Omega)}{\partial\Omega^{\mathrm{T}}}\middle|\mathcal{A}=a, D=d\right\};\\
\Gamma_2 &= \textstyle\sum_{a=1}^{A}\sum_{d=0}^{1}(n_{da}/n)E\left\{\dfrac{\partial R_{\Omega,a}(X,\Omega)/R_a(X,\Omega)}{\partial\Omega^{\mathrm{T}}}\middle|\mathcal{A}=a, D=d\right\}.
\end{aligned}$$

Then define

$$\begin{aligned}
\zeta_a(Z_i,\Omega) &= I(\mathcal{A}_i=a)\frac{S_{\Omega,a}(Z_i,\Omega)}{S_a(Z_i,\Omega)} - \frac{R_{\Omega,a}(X_i,\Omega)}{R_a(X_i,\Omega)}\\
&\quad -\sum_{d=0}^{1}\sum_{r=0}^{1}\frac{c_{d,a}\pi_{d_i,a}}{c_{d_i,a}}\\
&\qquad \times E\left[\{P_{1a}(X,\Omega)S_{\Omega,a}(r,g_i,X) - P_{2a}(X,\Omega)S_a(r,g_i,X)\}\mid \mathcal{A}=a, D=d\right],
\end{aligned}$$

and now Σ becomes

$$\Sigma = \textstyle\sum_{a=1}^{A}\sum_{d=0}^{1}(n_{da}/n)\mathrm{cov}\{\zeta_a(D,X,G,\Omega)|D=d,\mathcal{A}=a\}.$$

A.5 Additional Simulations

A.5.1 Comparison with the Method of Chatterjee and Carroll (2005)

Table 2 of this *Supplementary Material* gives results in the same simulation setting as in Section 3 in the main paper, except that to compare with Chatterjee and Carroll (2005), we only use the first SNP for our method and for the Chatterjee-Carroll method. The latter method uses the R package CGEN in Bioconductor, and is based on that package's function *snp.logistic*, which allows for SNP levels 0, 1, 2 and X values 0,1, as in our simulation. The results of that analysis and our method are very similar, indicating that our method is, in this case, almost efficient.

A.5.2 Misspecification of Population Disease Rate

Table 3 of this *Supplementary Material* reports the results of a simulation to evaluate the robustness of our method to misspecification of the population disease rate, using a sample of 1000 cases and 1000 controls. We considered actual disease rates of $\pi_1 = 0.03$, 0.05, 0.085 and 0.12, and compared the results for the rare disease approximation and when the assumed disease rate was $\pi_1 = 0.03$. For the method using the a rare disease approximation, it was only when the rate was $\pi_1 = 0.12$ that there was a deterioration in the coverage probabilities, but even then the lowest coverage rate was 91.8%. When the disease rate was assumed to be $\pi_1 = 0.03$, nominal coverage was seen except when the exact disease rate was $\pi_1 = 0.12$, and even at the worst case the lowest coverage rate was 93.1%, almost nominal. This indicates a surprising robustness to disease rate misspecification.

A.5.3 Violations of the Gene-Environment Independence Assumption

Tables A.4, A.5 and A.6 of this *Supplementary Material* contain simulations to examine the robustness of our method to violations of the gene-environment independence assumption. In these simulations, the genetic variables are generated as described in Section 3 of the main paper, but the environmental variable is normally distributed with mean αG_1, αG_2, or αG_3. We let $\alpha = 0.032$ to introduce a dependence between X and G with $R^2 = 0.001$. Here $\beta_G = \{\log(1.2), \log(1.2), 0, \log(1.2), 0\}$ as in Section 3 of the main paper, but $\beta_X = \log(1.73)$ and

$\beta_{GX} = \{\log(1.42), 0, 0, \log(1.42), 0\}$. In each simulation, the logistic intercept was chosen to give a 3% population disease prevalence. In Table A.4 X is correlated with G_1, which has a nonzero main effect and a nonzero interaction; in Table A.5 X is correlated with G_2, which has a nonzero main effect but no interaction effect; in Table A.6 X is correlated with G_3, which has neither main nor interaction effects.

Similarly to Chatterjee and Carroll (2005), we find that violating the G-E independence assumption induces a bias in the parameter estimates. In Section A.4 of this *Supplementary Material* we describe how to remove this bias when G and E are independent conditional on a discrete stratification variable $\mathcal{A}$. Mukherjee and Chatterjee (2008) and Chen et al. (2009) show how to use empirical-Bayes methods as well to provide additional robustness against violations of the gene-environment independence assumption.

A.6 Properties of $\widehat{R}(x,\Omega)$ in equation (5) of the Main Paper

Equation (5) of the main paper is

$$\widehat{R}(x,\Omega) = \sum_{j=1}^{J}\sum_{r=0}^{1}\sum_{d=0}^{1}(\pi_d/n_d)I(D_j = d)S(r,G_j,x,\Omega).$$

Computing its expectation is facilitated by seeing that

$$E\{I(D_j = d)S(r,G_j,x,\Omega)\} = E\{S(r,G_j,x,\Omega)|D_j = d\} = E\{S(r,G,x,\Omega)|D = d\}$$

Hence, recognizing that there are n_d subjects with $D = d$,

$$\begin{aligned} E\{\widehat{R}(x,\Omega)\} &= \sum_{j=1}^{J}\sum_{r=0}^{1}\sum_{d=0}^{1}(\pi_d/n_d)E\{S(r,G,x,\Omega)|D = d\} \\ &= \sum_{r=0}^{1}\sum_{d=0}^{1}\pi_d/n_d)E\{S(r,G,x,\Omega)|D = d\} \\ &= R(x,\Omega). \end{aligned}$$

Hence, (5) of the main paper is unbiased for $R(x,\Omega)$. Further, we see that

$$\begin{aligned} &\widehat{R}(x,\Omega) - R(x,\Omega) \\ &= \sum_{r=0}^{1}\sum_{d=0}^{1}(\pi_d/n_d)\sum_{j=1}^{J} \\ &\qquad \times \left[I(D_j = d)S(r,G_j,x,\Omega) - E\{I(D_j = d)S(r,G_j,x,\Omega)\}\right], \end{aligned}$$

so that $\widehat{R}(x,\Omega)$ is $n^{1/2}$-consistent for $R(x,\Omega)$, and with proper normalization is asymptotically normally distributed.

A.7 SNPs Involved in Creating the Polygenic Risk Score

Table A.1: SNPs involved in creating the polygenic risk score, and their regression coefficients

Actual RS Number	Variable Name	Coefficient
rs11249433	gene1	-0.02813492
rs1045485	gene2	-0.09307971
rs13387042	gene3	-0.26203658
rs4973768	gene4	0.08013260
rs10069690	gene5	0.06459363
rs10941679	gene6	0.09185539
rs889312	gene7	-0.00565121
rs17530068	gene8	0.09668742
rs2046210	gene9	0.09851217
rs1562430	gene10	-0.14871719
rs1011970	gene11	0.05329783
rs865686	gene12	-0.02913340
rs2380205	gene13	-0.01821032
rs10995190	gene14	-0.04275836
rs2981582	gene15	0.14008397
rs909116	gene16	0.04955235
rs614367	gene17	0.06438418
rs3803662	gene18	0.27080105
rs6504950	gene19	-0.17586244
rs8170	gene20	0.08570773
rs999737_as	gene21	-0.13737833

A.8 Comparison with the Method of Chatterjee and Carroll (2005) in a Special Case

Table A.2: Results of 1000 simulations with 3% disease prevalence as described in Section 3 of the main paper, except that to compare with Chatterjee and Carroll (2005), we only use the first SNP. We compare our semiparametric pseudolikelihood estimator to the method of Chatterjee and Carroll (2005) and to ordinary logistic regression. The simulations were performed with 500 cases and 500 controls

	500 cases & 500 controls			1000 cases & 1000 controls		
	β_{G1}	β_X	β_{G1X}	β_{G1}	β_X	β_{G1X}
True	0.182	0.405	0.262	0.182	0.405	0.262
			Logistic			
Bias	-0.011	0.001	0.015	0.009	0.003	-0.001
CI (%)	93.9	94.1	93.7	95.2	94.2	95.6
			Chatterjee Carroll			
Bias	-0.008	0.005	-0.004	0.013	0.006	-0.016
CI (%)	95.1	94.1	93.6	96.0	94.6	94.4
MSE Eff	1.405	1.108	2.227	1.321	1.118	2.183
			SPMLE, Rare			
Bias	-0.007	0.004	-0.001	0.013	0.006	-0.015
CI (%)	95.1	94.1	94.1	95.8	94.5	94.8
MSE Eff	1.381	1.104	2.166	1.290	1.113	2.141
			SPMLE, π_1 known			
Bias	-0.014	0.001	0.014	0.006	0.003	0.000
CI (%)	95.1	94.2	94.8	95.9	94.7	94.4
MSE Eff	1.359	1.100	2.016	1.292	1.113	2.021

Logistic is ordinary logistic regression; *Chatterjee Carroll* is the method of Chatterjee and Carroll (2005); *SPMLE, Rare* is our estimator using the rare disease approximation with unknown π_1 (Section 2.2 of the main paper); *SPMLE, π_1 known* is our estimator when π_1 is known in the source population (Section 2.1 of the main paper); *Bias* is the mean bias; *CI (%)* is the coverage in percent of a nominal 95% confidence interval (calculated using the asymptotic standard error); *MSE Eff* is the mean squared error efficiency of the method compared to logistic regression.

A.9 Simulation When the Disease Rate is Misspecified

Table A.3: Results of 1000 simulations as described in §3 of the main paper, except that the logistic intercept has been modified to give population disease rates $(0.03, 0.05, 0.085, 0.12)$. We compare ordinary logistic regression, our method using the rare disease approximation, and our method with "known" $\pi_1 = 0.03$, which is misspecified when $\pi_1 > 0.03$. The simulations were performed with 1000 cases and 1000 controls

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
True	0.18	0.18	0.00	0.18	0.00	0.41	0.26	0.00	0.00	0.26	0.00
						Logistic					
Disease Rate $= 0.03$											
Bias	0.00	0.01	0.00	0.01	-0.01	0.01	0.01	-0.01	0.00	0.00	0.01
CI (%)	94.3	95.2	95.7	95.1	94.7	94.6	94.9	94.2	94.5	96.0	94.2
Disease Rate $= 0.05$											
Bias	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.01
CI (%)	95.8	95.2	95.9	94.7	94.4	95.6	95.7	95.5	95.3	94.8	95.3
Disease Rate $= 0.085$											
Bias	-0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01
CI (%)	94.2	94.8	95.6	94.4	93.7	94.4	94.9	94.3	94.9	95.9	94.2
Disease Rate $= 0.12$											
Bias	0.00	0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
CI (%)	94.8	95.5	94.9	95.2	93.8	95.7	94.4	95.9	94.9	95.3	95.0
						SPMLE, Rare					
Disease Rate $= 0.03$											
Bias	0.01	0.00	0.00	0.02	-0.01	0.02	-0.02	-0.01	0.01	-0.02	0.01
CI (%)	95.2	95.4	96.4	95.8	95.3	95.1	95.4	94.8	96.1	95.5	94.9
MSE Eff		All G: 1.28				X: 1.26			All $G*X$: 2.18		
Disease Rate $= 0.05$											
Bias	0.02	0.00	0.00	0.02	-0.01	0.03	-0.04	0.00	0.00	-0.03	0.00
CI (%)	94.4	95.4	96.8	94.4	95.0	95.1	93.8	94.6	96.3	94.5	94.4
MSE Eff		All G: 1.25				X: 1.23			All $G*X$: 1.99		
Disease Rate $= 0.085$											
Bias	0.02	0.01	0.00	0.02	0.00	0.05	-0.05	-0.01	0.00	-0.05	0.00
CI (%)	95.0	94.5	96.1	94.1	93.9	93.5	93.9	94.8	95.8	94.5	95.6
MSE Eff		All G: 1.25				X: 1.14			All $G*X$: 2.02		
Disease Rate $= 0.12$											
Bias	0.03	0.01	-0.01	0.03	0.00	0.06	-0.08	-0.01	0.00	-0.06	0.00
CI (%)	94.2	95.5	94.6	93.3	93.9	93.4	92.0	96.1	94.5	91.8	94.4
MSE Eff		All G: 1.21				X: 1.02			All $G*X$: 1.88		

Continued on next page

Table A.3 – continued from previous page

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
					SPMLE, $\pi_1 = 0.03$						
Disease Rate $= 0.03$											
Bias	0.00	0.00	0.00	0.01	-0.01	0.01	0.00	-0.01	0.01	-0.01	0.01
CI (%)	95.1	95.5	96.4	95.8	95.0	95.5	95.6	94.6	95.9	95.2	94.5
MSE Eff		All G: 1.28				X: 1.28			All $G*X$: 2.07		
Disease Rate $= 0.05$											
Bias	0.01	0.00	0.00	0.01	-0.01	0.01	-0.01	0.00	0.00	-0.01	0.01
CI (%)	94.6	95.4	96.4	94.7	94.7	95.8	94.3	94.6	96.0	94.5	94.1
MSE Eff		All G: 1.25				X: 1.27			All $G*X$: 1.90		
Disease Rate $= 0.085$											
Bias	0.01	0.01	0.00	0.01	0.00	0.03	-0.03	-0.01	0.00	-0.03	0.00
CI (%)	95.1	94.8	96.4	94.4	93.9	94.7	94.9	95.1	95.8	94.9	95.2
MSE Eff		All G: 1.25				X: 1.21			All $G*X$: 1.95		
Disease Rate $= 0.12$											
Bias	0.02	0.01	-0.01	0.03	0.00	0.05	-0.06	-0.01	0.00	-0.05	0.01
CI (%)	94.3	95.6	94.9	93.6	93.8	94.4	93.5	96.3	94.6	93.1	94.6
MSE Eff		All G: 1.22				X: 1.10			All $G*X$: 1.84		

Logistic is ordinary logistic regression; *SPMLE, Rare* is our estimator using the rare disease approximation with unknown π_1 (Section 2.2 of the main paper); *SPMLE,* $\pi_1 = 0.03$ is our estimator calculated as if the disease rate in the source population were known to be 0.03 (Section 2.1 of the main paper); *Bias* is the mean bias; *CI (%)* is the coverage in percent of a nominal 95% confidence interval (calculated using the asymptotic standard error); *MSE Eff* is the mean squared error efficiency of our method compared to logistic regression, averaged over G, over X and over the $G*X$ interactions.

A.10 Simulations When the Gene-Environment Independence Assumption is Violated

Table A.4: Results of 1000 simulations with G as described in Section 3 of the main paper, but $X \sim \mathbf{N}(0.032G_1, 1)$. We compare our semiparametric pseudolikelihood estimator to ordinary logistic regression. Three simulations were performed with sample sizes of (1000, 2000, 3000) cases and controls each

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
True	0.18	0.18	0.00	0.18	0.00	0.55	0.35	0.00	0.00	0.35	0.00
						Logistic: 1000 cases					
Bias	-0.01	0.00	0.01	0.00	-0.01	0.01	0.01	0.01	-0.01	0.01	0.01
CI (%)	94.5	96.2	95.8	94.8	93.7	94.0	95.4	95.7	95.6	95.5	95.3
						Logistic: 2000 cases					
Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CI (%)	95.4	94.7	94.8	95.2	95.0	94.5	95.6	96.1	94.0	94.7	95.9
						Logistic: 3000 cases					
Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CI (%)	94.1	94.1	95.1	95.7	94.6	94.2	94.2	95.4	94.8	95.0	94.7
						SPMLE, π_1 known: 1000 cases					
Bias	-0.01	0.00	0.01	0.00	-0.01	-0.03	0.10	0.00	0.00	0.01	0.00
CI (%)	94.2	95.9	95.0	95.2	93.8	93.3	80.4	94.9	94.9	95.0	94.8
MSE Eff			All G: 1.07			X: 1.31			All $G*X$: 1.75		
						SPMLE, π_1 known: 2000 cases					
Bias	0.00	0.00	0.00	0.01	0.00	-0.03	0.10	0.00	0.00	0.00	0.00
CI (%)	94.2	94.8	95.1	95.5	95.6	90.9	71.4	95.5	94.1	95.0	95.6
MSE Eff			All G: 1.07			X: 1.08			All $G*X$: 1.53		
						SPMLE, π_1 known: 3000 cases					
Bias	-0.01	0.00	0.00	0.00	0.00	-0.03	0.10	0.00	0.00	0.00	0.00
CI (%)	94.7	95.3	95.7	95.2	94.2	88.0	54.8	94.2	95.7	95.0	93.9
MSE Eff			All G: 1.06			X: 0.95			All $G*X$: 1.27		

Logistic is ordinary logistic regression; *SPMLE, π_1 known* is our estimator when π_1 is known in the source population (Section 2.1 of the main paper); *Bias* is the mean bias; *CI (%)* is the coverage in percent of a nominal 95% confidence interval (calculated using the asymptotic standard error); *MSE Eff* is the mean squared error efficiency of our method compared to logistic regression, averaged over G, over X and over the $G*X$ interactions.

Table A.5: Results of 1000 simulations with G as described in Section 3 of the main paper, but $X \sim \mathbf{N}(0.032G_2, 1)$. We compare our semiparametric pseudolikelihood estimator to ordinary logistic regression. Three simulations were performed with sample sizes of (1000, 2000, 3000) cases and controls each

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
True	0.18	0.18	0.00	0.18	0.00	0.55	0.35	0.00	0.00	0.35	0.00
						Logistic: 1000 cases					
Bias	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00
CI (%)	93.4	95.1	94.5	93.0	95.7	94.4	94.4	93.7	94.8	93.4	94.4
						Logistic: 2000 cases					
Bias	0.00	-0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00
CI (%)	95.3	94.0	94.4	94.6	93.2	94.9	94.6	94.8	94.2	95.5	93.8
						Logistic: 3000 cases					
Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CI (%)	94.1	94.5	94.5	95.3	95.2	95.9	94.7	93.9	94.4	95.6	95.3
						SPMLE, π_1 known: 1000 cases					
Bias	0.00	-0.01	0.00	0.01	-0.01	-0.04	0.01	0.06	0.00	0.00	0.00
CI (%)	93.7	95.3	95.4	94.0	95.1	89.4	93.8	86.0	95.0	94.6	94.9
MSE Eff			All G: 1.06			X: 1.12			All $G*X$: 2.19		
						SPMLE, π_1 known: 2000 cases					
Bias	0.00	-0.01	0.00	0.00	0.00	-0.04	0.01	0.06	0.00	0.00	0.00
CI (%)	95.6	94.2	94.9	94.4	93.9	88.1	94.3	78.7	95.1	95.4	95.6
MSE Eff			All G: 1.08			X: 0.91			All $G*X$: 1.91		
						SPMLE, π_1 known: 3000 cases					
Bias	0.00	0.00	0.00	0.00	0.00	-0.04	0.01	0.06	0.00	0.00	0.00
CI (%)	94.8	94.2	94.9	95.9	94.9	84.3	95.4	72.7	95.4	95.3	95.5
MSE Eff			All G: 1.08			X: 0.72			All $G*X$: 1.82		

Logistic is ordinary logistic regression; ***SPMLE, π_1 known*** is our estimator when π_1 is known in the source population (Section 2.1 of the main paper); ***Bias*** is the mean bias; ***CI (%)*** is the coverage in percent of a nominal 95% confidence interval (calculated using the asymptotic standard error); ***MSE Eff*** is the mean squared error efficiency of our method compared to logistic regression, averaged over G, over X and over the $G*X$ interactions.

Table A.6: Results of 1000 simulations with G as described in Section 3 of the main paper, but $X \sim \mathbf{N}(0.032G_3, 1)$. We compare our semiparametric pseudolikelihood estimator to ordinary logistic regression. Three simulations were performed with sample sizes of (1000, 2000, 3000) cases and controls each

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
True	0.18	0.18	0.00	0.18	0.00	0.55	0.35	0.00	0.00	0.35	0.00
						Logistic: 1000 cases					
Bias	-0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01
CI (%)	95.5	94.4	95.2	96.2	95.3	94.7	94.9	94.0	94.9	95.5	94.9
						Logistic: 2000 cases					
Bias	0.00	0.00	-0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
CI (%)	94.0	94.1	94.4	94.6	94.9	95.2	95.5	95.1	95.5	94.0	94.7
						Logistic: 3000 cases					
Bias	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
CI (%)	95.9	94.2	94.1	94.8	94.3	95.1	95.4	95.9	95.8	92.9	94.4
						SPMLE, π_1 known: 1000 cases					
Bias	0.00	0.00	0.00	0.00	0.00	-0.04	0.01	0.00	0.06	0.01	0.00
CI (%)	95.6	94.8	95.5	96.3	95.3	92.0	94.8	95.5	88.3	95.7	96.2
MSE Eff			All G: 1.07			X: 1.20			All $G*X$: 2.12		
						SPMLE, π_1 known: 2000 cases					
Bias	0.00	0.00	-0.01	0.00	0.00	-0.04	0.01	0.00	0.06	0.00	0.00
CI (%)	95.2	94.4	94.5	94.0	94.8	89.4	95.0	94.8	82.3	94.9	94.6
MSE Eff			All G: 1.06			X: 0.95			All $G*X$: 1.95		
						SPMLE, π_1 known: 3000 cases					
Bias	0.00	0.00	0.00	0.00	-0.01	-0.04	0.00	0.00	0.06	0.00	0.00
CI (%)	95.3	94.7	94.0	95.3	94.2	84.5	94.4	94.9	75.7	95.0	94.8
MSE Eff			All G: 1.06			X: 0.76			All $G*X$: 1.82		

Logistic is ordinary logistic regression; *SPMLE, π_1 known* is our estimator when π_1 is known in the source population (Section 2.1 of the main paper); *Bias* is the mean bias; *CI (%)* is the coverage in percent of a nominal 95% confidence interval (calculated using the asymptotic standard error); *MSE Eff* is the mean squared error efficiency of our method compared to logistic regression, averaged over G, over X and over the $G*X$ interactions.

A.11 The Simulation in Table 1 of the Main Paper With Componentwise Mean Squared Error Efficiencies

Table A.7: Results of 1000 simulations as described in Section 3 of the main paper, with mean bias, coverage probabilities of a 95% nominal confidence interval, and mean squared error efficiency of our semiparametric pseudolikelihood estimator compared to ordinary logistic regression. The sample sizes were performed with 500 cases and 500 controls, and again with 1000 cases and 1000 controls

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
True	0.18	0.18	0.00	0.18	0.00	0.41	0.26	0.00	0.00	0.26	0.00
						Logistic, 500 cases					
Bias	0.02	-0.02	0.02	-0.01	0.01	0.00	0.00	0.02	-0.02	0.02	-0.01
CI (%)	94.7	94.9	94.8	94.5	95.2	96.4	94.3	93.6	94.3	94.9	95.4
						Logistic, 1000 cases					
Bias	0.00	0.01	0.00	0.01	-0.01	0.01	0.01	-0.01	0.00	0.00	0.01
CI (%)	94.3	95.2	95.7	95.1	94.7	94.6	94.9	94.2	94.5	96.0	94.2
						SPMLE, Rare, 500 cases					
Bias	0.02	-0.01	0.02	0.00	0.00	0.01	-0.01	0.01	-0.02	-0.01	0.00
CI (%)	95.0	95.8	94.2	94.5	95.5	95.6	95.8	95.3	94.3	95.0	95.9
MSE Eff	1.37	1.34	1.23	1.27	1.27	1.29	2.44	2.13	1.87	1.91	2.22
						SPMLE, Rare, 1000 cases					
Bias	0.01	0.00	0.00	0.02	-0.01	0.02	-0.02	-0.01	0.01	-0.02	0.01
CI (%)	95.2	95.4	96.4	95.8	95.3	95.1	95.4	94.8	96.1	95.5	94.9
MSE Eff	1.35	1.25	1.29	1.25	1.24	1.26	2.36	2.00	2.19	2.02	2.21
						SPMLE, π_1 known: 500 cases					
Bias	0.01	-0.01	0.02	-0.01	0.00	0.00	0.01	0.01	-0.02	0.01	0.00
CI (%)	95.0	95.7	94.3	94.4	95.5	95.7	95.4	95.1	94.3	94.9	95.7
MSE Eff	1.39	1.34	1.22	1.26	1.28	1.28	2.31	2.01	1.78	1.81	2.09
						SPMLE, π_1 known: 1000 cases					
Bias	0.00	0.00	0.00	0.01	-0.01	0.01	0.00	-0.01	0.01	-0.01	0.01
CI (%)	95.1	95.5	96.4	95.8	95.0	95.5	95.6	94.6	95.9	95.2	94.5
MSE Eff	1.36	1.25	1.28	1.27	1.24	1.28	2.25	1.91	2.06	1.96	2.08

Logistic is ordinary logistic regression; *SPMLE, Rare* is our estimator using the rare disease approximation with unknown π_1, Section 2.2; *SPMLE, π_1 known* is our estimator when π_1 is known in the source population, Section 2.1; *CI (%)* is the coverage in percent of a nominal 95% confidence interval, calculated using the asymptotic standard error; *MSE Eff* is the mean squared error efficiency of our method compared to logistic regression.

A.12 Skewness, Kurtosis and qq-Plots for the Simulation in Table 1 of the Main Paper

Table A.8 gives skewness and kurtosis for the simulation in Table 1 of the main paper with 1000 cases and controls. Figure A.1 presents q–q plots for the main effects for $(G_1, \ldots, G_5, X)$ in the same simulation. Figure A.2 presents q–q plots for the interaction effects for X and $(G_1, \ldots, G_5)$ in the same simulation.

Table A.8: Skewness and kurtosis for the simulation in Table 1 of the main paper with 1000 cases and controls. Kurtosis = 0 for the normal distribution

Skewness	Kurtosis
-0.02	-0.08
-0.05	0.12
-0.13	0.07
-0.02	-0.15
0.06	-0.04
-0.21	0.15
-0.01	-0.20
-0.03	-0.10
0.04	0.11
0.09	-0.13
0.01	-0.06
0.14	0.25

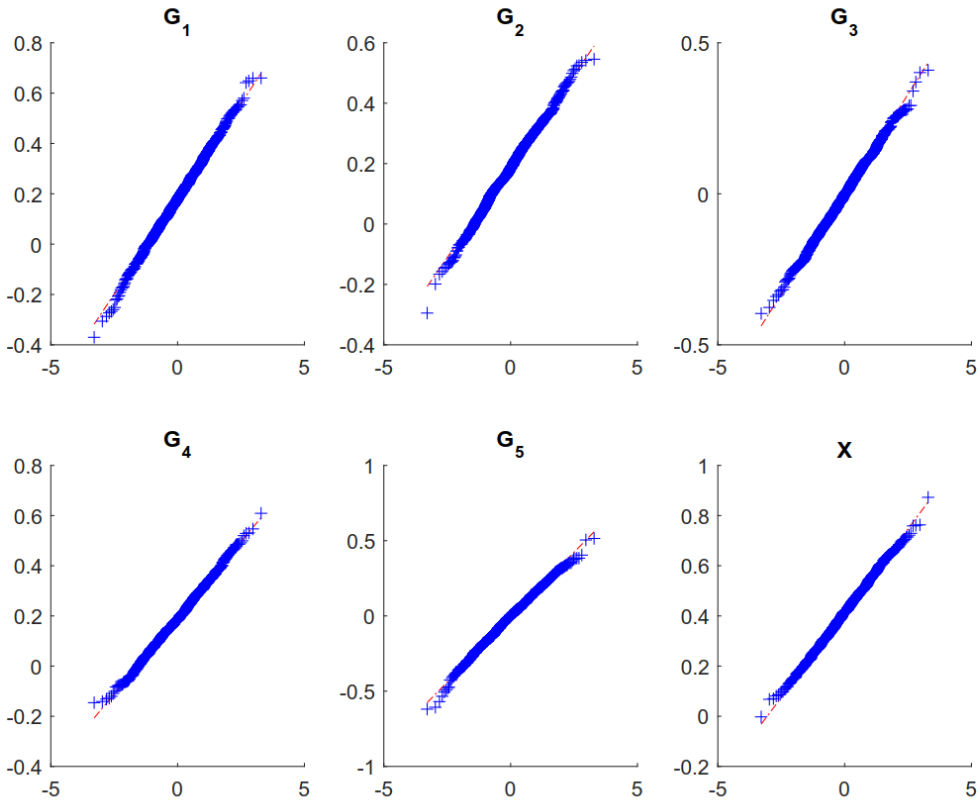

Figure A.1: The qq-plots for the main effects for $(G_1, \ldots, G_5, X)$ in the simulation in Table 1 of the main paper with 1000 cases and controls.

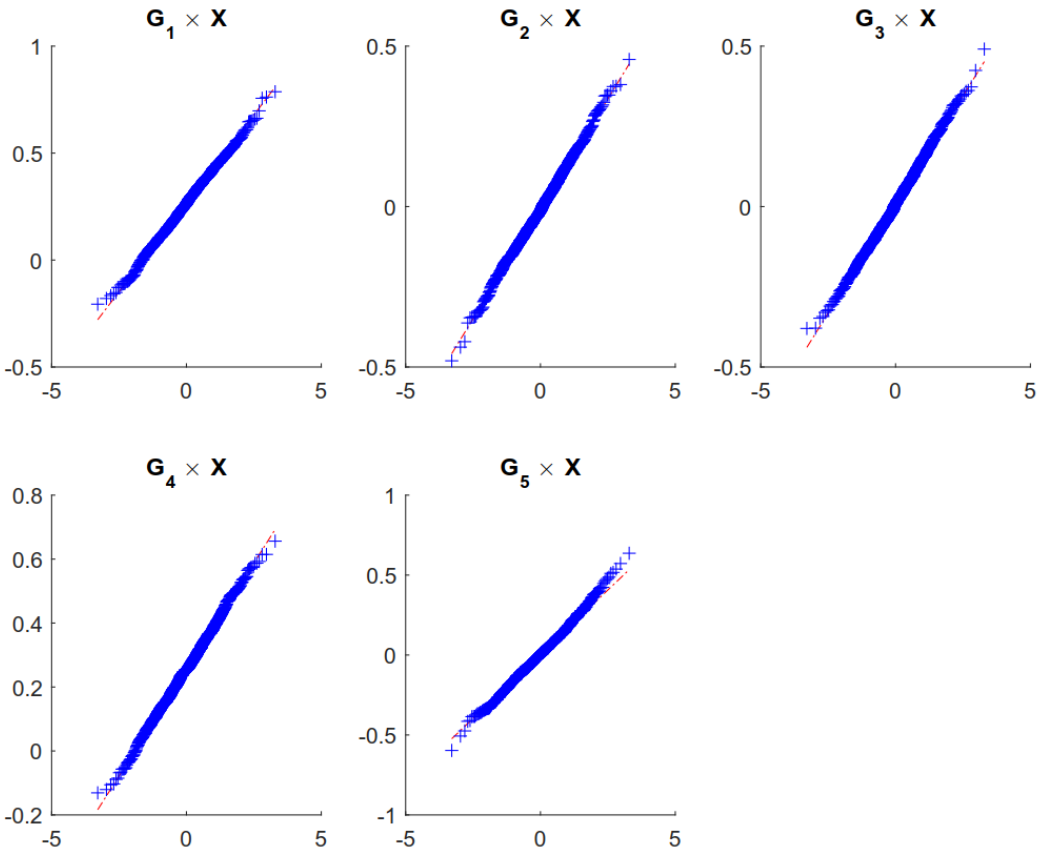

Figure A.2: The qq-plots for the interaction effects for X and $(G_1, \ldots, G_5)$ in the simulation in Table 1 of the main paper with 1000 cases and controls.

A.13 The Simulation in Table 1 of the Main Paper With 500 Cases and Controls

Table A.9: Results of 1000 simulations as described in §3 of the main paper, with mean bias, coverage probabilities of a 95% nominal confidence interval, and mean squared error efficiency of our semiparametric pseudolikelihood estimator compared to ordinary logistic regression. The simulations were performed with 500 cases and 500 controls

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{G1X}	β_{G2X}	β_{G3X}	β_{G4X}	β_{G5X}
True	0.18	0.18	0.00	0.18	0.00	0.41	0.26	0.00	0.00	0.26	0.00
						Logistic: 500 cases					
Bias	0.02	-0.02	0.02	-0.01	0.01	0.00	0.00	0.02	-0.02	0.02	-0.01
CI (%)	94.7	94.9	94.8	94.5	95.2	96.4	94.3	93.6	94.3	94.9	95.4
						SPMLE, Rare: 500 cases					
Bias	0.02	-0.01	0.02	0.00	0.00	0.01	-0.01	0.01	-0.02	-0.01	0.00
CI (%)	95.0	95.8	94.2	94.5	95.5	95.6	95.8	95.3	94.3	95.0	95.9
Avg MSE Eff		All G: 1.30				All X: 1.29			All $G*X$: 2.13		
						SPMLE, π_1 known: 500 cases					
Bias	0.01	-0.01	0.02	-0.01	0.00	0.00	0.01	0.01	-0.02	0.01	0.00
CI (%)	95.0	95.7	94.3	94.4	95.5	95.7	95.4	95.1	94.3	94.9	95.7
Avg MSE Eff		All G: 1.30				All X: 1.28			All $G*X$: 2.02		

Logistic is ordinary logistic regression; *SPMLE, Rare* is our estimator using the rare disease approximation with unknown π_1 (§2.2); *SPMLE, π_1 known* is our estimator when π_1 is known in the source population (§2.1); *CI (%)* is the coverage in percent of a nominal 95% confidence interval (calculated using the asymptotic standard error); *Avg MSE Eff* is the mean squared error efficiency of our method compared to logistic regression averaged over G, over X and over the $G*X$ interactions, respectively.

APPENDIX B

APPENDIX TO SECTION 3

B.1 Composite Likelihood Estimator

The composite profile likelihood is just the average of the two symmetric profile likelihoods

$$\begin{aligned}\mathcal{L}_{\mathrm{CL}}(\Omega) &= (\mathcal{L}_X(\Omega) + \mathcal{L}_G(\Omega))/2 \\ &= \textstyle\sum_{i=1}^n \log\{S(D_i, G_i, X_i, \Omega)\} - 0.5\sum_{i=1}^n \log\{\widehat{R}_X(G_i, \Omega)\} - 0.5\sum_{i=1}^n \log\{\widehat{R}_G(X_i, \Omega)\}.\end{aligned}$$

The estimated score function is thus the average of the two symmetric score functions

$$\begin{aligned}\widehat{\mathcal{S}}_{\mathrm{CL}}(\Omega) &= (\widehat{\mathcal{S}}_X(\Omega) + \widehat{\mathcal{S}}_{\mathrm{G}}(\Omega))/2 \\ &= n^{-1/2}\sum_{i=1}^n \left\{ \frac{S_\Omega(D_i, G_i, X_i, \Omega)}{S(D_i, G_i, X_i, \Omega)} - \frac{1}{2}\frac{\widehat{R}_{X\Omega}(X_i, \Omega)}{\widehat{R}_X(X_i, \Omega)} - \frac{1}{2}\frac{\widehat{R}_{G\Omega}(G_i, \Omega)}{\widehat{R}_G(G_i, \Omega)} \right\}.\end{aligned}$$

Estimate $\widehat{\Omega}_{\mathrm{CL}}$ is calculated by solving $\widehat{\mathcal{S}}_{\mathrm{CL}}(\Omega) = 0$, or equivalently, maximizing $\mathcal{L}_{\mathrm{CL}}(\Omega)$.

Following the notation defined previously, we sum eqs. (3.7) and (3.8) together instead of stacking them as in Theorem 2.

Theorem 3. *Suppose that* $0 < \lim_{n\to\infty} n_d/n < 1$*, and* π_1 *is known. Then*

$$n^{1/2}(\widehat{\Omega}_{\mathrm{CL}} - \Omega) = -(\Gamma_X + \Gamma_G)^{-1} n^{-1/2} \textstyle\sum_{i=1}^n \{\zeta_{X*}(Z_i, \Omega) + \zeta_{G*}(Z_i, \Omega)\} + o_p(1).$$

To calculate the asymptotic variance, write

$$\begin{aligned}
\Sigma_{\text{all}} &= \begin{bmatrix} \Sigma_{XX} & \Sigma_{XG} \\ \Sigma_{GX} & \Sigma_{GG} \end{bmatrix} = \text{cov}\left\{ \begin{matrix} \zeta_{X*}(Z_i, \Omega) \\ \zeta_{G*}(Z_i, \Omega) \end{matrix} \right\} = \text{cov}\left\{ \begin{matrix} \zeta_{X}(Z_i, \Omega) \\ \zeta_{G}(Z_i, \Omega) \end{matrix} \right\}; \\
\Sigma_{XX} &= \textstyle\sum_{d=0}^{1}(n_d/n)\text{cov}\{\zeta_{X*}(Z, \Omega)|D = d\} = \sum_{d=0}^{1}(n_d/n)\text{cov}\{\zeta_{X}(Z, \Omega)|D = d\}; \\
\Sigma_{GG} &= \textstyle\sum_{d=0}^{1}(n_d/n)\text{cov}\{\zeta_{G*}(Z, \Omega)|D = d\} = \sum_{d=0}^{1}(n_d/n)\text{cov}\{\zeta_{G}(Z, \Omega)|D = d\}; \\
\Sigma_{XG} &= \textstyle\sum_{d=0}^{1}(n_d/n)\text{cov}\{\zeta_{X*}(Z, \Omega), \zeta_{G*}(Z, \Omega)|D = d\} \\
&= \textstyle\sum_{d=0}^{1}(n_d/n)\text{cov}\{\zeta_{X}(Z, \Omega), \zeta_{G}(Z, \Omega)|D = d\} = \Sigma_{XG}^{\text{T}}.
\end{aligned}$$

Since the Z_i are independent and $E\{\zeta_{X*}(Z_i, \Omega)|D_i\} = E\{\zeta_{G*}(Z_i, \Omega)|D_i\} = 0$, then

$$\begin{aligned}
n^{1/2}(\widehat{\Omega}_{\text{CL}} - \Omega) &\rightarrow \textit{Normal}(0, \Lambda_{\text{CL}}); \\
\Sigma_{\text{CL}} &= \textstyle\sum_{d=0}^{1}(n_d/n)\text{cov}\{\zeta_{X*}(Z_i, \Omega) + \zeta_{G*}(Z_i, \Omega)\} \\
&= \Sigma_{XX} + \Sigma_{GG} + \Sigma_{XG} + \Sigma_{GX}; \\
\Lambda_{\text{CL}} &= (\Gamma_X + \Gamma_G)^{-1}\Sigma_{\text{CL}}\{(\Gamma_X + \Gamma_G)^{-1}\}^{\text{T}}.
\end{aligned}$$

The proof of Theorem 3 follows directly from the proofs of eqs. (3.7) and (3.8) and the properties of M-estimators $\widehat{\Omega}_X$ and $\widehat{\Omega}_G$.

B.2 Additional Simulations

B.2.1 Unabridged version of Table 3.1 from Section 3.3

Table 3.1 in Section 3.3 of the main paper reports the results of four estimators: logistic regression, the SPMLE with known π_1, our Symmetric Combination Estimator with known π_1, and our Symmetric Combination Estimator using the rare disease approximation. Table B.1 presents the results of *all* estimators in 1000 simulations under the simulation settings of Section 3.3.1. In addition to logistic regression, four retrospective methods are presented: the SPMLE ($\widehat{\Omega}_X$), the SPMLE_G ($\widehat{\Omega}_G$), the Composite Likelihood Estimator ($\widehat{\Omega}_{\text{CL}}$), and the Symmetric Combination Es-

timator ($\widehat{\Omega}_{\mathrm{Symm}}$). Each retrospective estimator was calculated under two conditions: with known π_1, and with unknown π_1 using the rare disease approximation.

We see that the rare disease approximation of each retrospective estimator closely matches the version calculated with known π_1. The efficiency of the Composite Likelihood Estimator is equivalent to that of the SPMLE and its symmetric counterpart, the SPMLE_G. The Symmetric Combination Estimator stands out as markedly more efficient than the other estimators.

Table B.1: Results of 1000 simulations as described in Section 3.3.1, comparing the bias, coverage, and efficiency of all estimators

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{XG1}	β_{XG2}	β_{XG3}	β_{XG4}	β_{XG5}
True	0.18	0.18	0.00	0.18	0.00	0.41	0.26	0.00	0.00	0.26	0.00
Logistic Regression											
Bias	0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
CI(%)	95.2	95.5	94.4	94.7	95.3	95.8	94.5	95.9	94.7	94.6	95.3
SPMLE, known π_1											
Bias	0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.01
CI(%)	95.4	95.8	94.8	96.1	96.3	95.6	94.6	96.0	94.3	95.6	95.1
MSE Eff	1.32	1.25	1.26	1.32	1.30	1.27	2.08	1.78	1.88	1.95	2.12
SPMLE, rare											
Bias	0.02	0.00	0.00	0.01	-0.01	0.02	-0.02	0.00	0.00	-0.01	0.01
CI(%)	95.0	95.7	94.8	95.9	96.4	95.5	94.4	96.0	94.7	95.4	95.7
MSE Eff	1.31	1.26	1.27	1.32	1.30	1.26	2.18	1.89	2.01	2.06	2.27
SPMLE_G, known π_1											
Bias	0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.01
CI(%)	95.0	95.8	94.8	96.1	96.3	95.2	94.8	95.5	94.1	95.6	95.6
MSE Eff	1.35	1.27	1.29	1.34	1.33	1.28	2.12	1.82	1.90	1.98	2.14
SPMLE_G, rare											
Bias	0.02	0.00	0.00	0.01	-0.01	0.02	-0.02	0.00	0.00	-0.01	0.01
CI(%)	94.9	95.7	94.9	95.9	96.3	94.8	94.1	95.5	94.3	95.0	95.7
MSE Eff	1.35	1.29	1.31	1.35	1.35	1.27	2.25	1.94	2.04	2.10	2.32
Composite Likelihood Estimator, known π_1											
Bias	0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.01
CI(%)	94.9	95.8	94.9	96.1	96.5	95.4	94.7	95.7	94.4	95.7	95.5
MSE Eff	1.34	1.26	1.28	1.33	1.32	1.28	2.11	1.81	1.90	1.98	2.14
Composite Likelihood Estimator, rare											
Bias	0.02	0.00	0.00	0.01	-0.01	0.02	-0.02	0.00	0.00	-0.01	0.01
CI(%)	94.9	95.6	94.9	95.9	96.4	95.2	94.1	95.8	94.8	95.3	95.7
MSE Eff	1.32	1.27	1.29	1.33	1.32	1.27	2.23	1.92	2.03	2.09	2.31
Symmetric Combination Estimator, known π_1											
Bias	0.00	-0.03	0.00	0.00	-0.01	0.01	-0.03	0.02	0.00	-0.02	0.01
CI*(%)	96.7	95.7	96.7	96.5	97.8	95.4	94.8	96.7	96.2	96.6	97.2
MSE Eff	1.92	1.71	2.00	1.83	2.05	1.31	2.84	2.51	2.99	2.68	3.34
Symmetric Combination Estimator, rare											
Bias	0.01	-0.02	0.00	0.01	-0.01	0.02	-0.05	0.02	0.00	-0.03	0.00
CI*(%)	96.4	95.7	95.7	96.3	98.1	94.9	94.0	97.0	96.4	95.5	97.5
MSE Eff	1.86	1.71	1.92	1.78	1.95	1.27	2.75	2.66	3.08	2.69	3.58

CI: coverage of a 95% nominal confidence interval, calculated using asymptotic standard error. *CI**: coverage of a 95% nominal confidence interval, calculated using 200 bootstrap samples. *MSE Eff*: mean squared error efficiency when compared to logistic regression.

B.2.2 Simulation when the disease rate is misspecified

Table B.2 presents the results of a simulation to evaluate the robustness of our method to misspecification of the population disease rate. A sample of 1000 cases and 1000 controls was simulated using the same scenario as described in Section 3.3.1 except the logistic intercept was modified to yield true population disease rates of 0.05, 0.085, and 0.12. In each instance, 1000 data sets were simulated and the Symmetric Combination Estimator was calculated with misspecified "known $\pi_1 = 0.03$" and again using the rare disease approximation.

When using the rare disease approximation, coverage remains near nominal until the true disease rate reached 0.085, and even then the lowest coverage rate was 91.3% (for interaction parameter β_{XG1}, which still demonstrated a mean squared error efficiency of 2.51 compared to logistic regression). When the disease rate was assumed "known $\pi_1 = 0.03$", nominal coverage was seen except when the population disease rate was 0.12. This indicates the Symmetric Combination Estimator is fairly robust to disease rate misspecification, and even an imprecise estimate of π_1 is likely to be sufficient to conduct a valid analysis.

Table B.2: Results of simulations as described in Section 3.3.1, but with population disease rates (0.05, 0.085, 0.12). For each disease rate, we simulated 1000 data sets and compared logistic regression, our method with misspecified "known $\pi_1 = 0.03$", and our method using the rare disease approximation.

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{XG1}	β_{XG2}	β_{XG3}	β_{XG4}	β_{XG5}
True	0.18	0.18	0.00	0.18	0.00	0.41	0.26	0.00	0.00	0.26	0.00
Disease Rate = 0.05				Logistic Regression							
Bias	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.01
CI(%)	95.8	95.2	95.9	94.7	94.4	95.6	95.7	95.5	95.3	94.8	95.3
							Symmetric Combination Estimator, "known $\pi_1 = 0.03$"				
Bias	0.00	-0.03	0.00	0.00	-0.01	0.03	-0.05	0.02	0.00	-0.04	0.01
CI*(%)	97.6	94.1	97.0	94.8	95.8	95.1	93.7	95.6	96.8	94.8	96.8
MSE Eff	1.84	1.74	2.07	1.69	1.97	1.30	2.61	2.65	3.14	2.37	2.97
							Symmetric Combination Estimator, rare				
Bias	0.01	-0.02	0.00	0.01	-0.01	0.04	-0.06	0.02	0.00	-0.05	0.00
CI*(%)	96.9	94.3	97.4	94.4	96.1	94.7	92.2	95.4	96.7	93.4	96.6
MSE Eff	1.75	1.73	2.03	1.60	1.89	1.22	2.48	2.76	3.22	2.25	3.11
Disease Rate = 0.085				Logistic Regression							
Bias	-0.01	0.01	0.00	0.00	0.00	0.00	0.01	-0.01	0.00	0.00	0.01
CI(%)	94.3	94.9	95.4	94.4	93.5	94.5	94.8	94.0	95.1	95.8	94.5
							Symmetric Combination Estimator, "known $\pi_1 = 0.03$"				
Bias	0.00	-0.02	0.00	0.01	-0.01	0.05	-0.07	0.01	0.00	-0.06	0.00
CI*(%)	96.4	95.2	96.9	95.7	95.7	93.6	92.7	96.0	97.6	92.9	97.1
MSE Eff	1.84	1.81	1.99	1.61	1.90	1.18	2.65	2.81	3.18	2.15	3.20
							Symmetric Combination Estimator, rare				
Bias	0.01	-0.01	0.00	0.02	-0.01	0.06	-0.08	0.01	0.00	-0.06	0.00
CI*(%)	96.5	95.8	96.5	95.7	95.3	92.5	91.3	96.1	97.3	91.8	97.1
MSE Eff	1.77	1.81	1.98	1.59	1.86	1.12	2.51	2.91	3.21	2.07	3.30
Disease Rate = 0.12				Logistic Regression							
Bias	0.00	0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CI(%)	94.6	95.4	94.9	94.8	93.7	95.7	94.4	95.9	94.8	94.8	94.8
							Symmetric Combination Estimator, "known $\pi_1 = 0.03$"				
Bias	0.00	-0.02	0.00	0.02	-0.01	0.06	-0.08	0.01	0.00	-0.07	0.01
CI*(%)	96.4	95.6	96.1	94.5	95.6	93.5	89.2	96.4	97.2	89.0	96.9
MSE Eff	1.83	1.71	1.95	1.59	1.86	1.08	2.33	2.80	3.07	1.90	3.02
							Symmetric Combination Estimator, rare				
Bias	0.02	-0.01	0.00	0.03	-0.01	0.07	-0.10	0.01	0.00	-0.08	0.00
CI*(%)	95.6	96.1	96.4	94.5	95.1	91.8	86.0	96.6	97.0	87.4	96.0
MSE Eff	1.72	1.72	1.91	1.53	1.79	0.99	2.11	2.95	3.14	1.78	3.11

CI: coverage of a 95% nominal confidence interval, calculated using asymptotic standard error. *CI**: coverage of a 95% nominal confidence interval, calculated using 200 bootstrap samples.
MSE Eff: mean squared error efficiency when compared to logistic regression.

B.2.3 Violations of the Gene-Environment Independence Assumption

Table B.3 presents the results of simulations to examine the robustness of our methods to violations of the gene-environment independence assumption. In these simulations, a sample of 1000 cases and 1000 controls is simulated with genetic variables as described in Section 3.3.1, but the environmental variable is normally distributed with mean αG_1, αG_2, or αG_3. We set $\alpha = 0.032$ to induce dependence between X and G_j with $R^2 = 0.001$. Here $\beta_G = \{\log(1.2), \log(1.2), 0, \log(1.2), 0\}$ as in Section 3.3.1, but $\beta_X = \log(1.35)$, and $\beta_{GX} = \{\log(1.21), 0, 0, \log(1.21), 0\}$. In each simulation, the logistic intercept was selected to give a population disease rate of 0.03. In the first simulation, X is correlated with G_1, which has a nonzero main effect and a nonzero interaction; in the second simulation, X is correlated with G_2, which has a nonzero main effect but no interaction effect; in the third simulation, X is correlated with G_3, which has neither main nor interaction effects.

We find that violating the gene-environment independence assumption induces bias in the estimate of the interaction parameter of the environmental variable and the specific SNP that is in violation of the gene-environment independence assumption, while the estimated interaction parameters of the other SNPs are unaffected. When π_1 is known, estimates of the main effects of the SNP that is in violation of the gene-environment independence assumption are uncompromised.

Table B.3: Results of simulations violating the gene-environment independence assumption with $X \sim N(0,\ 0.032G_j)$ for SNPs (G_1, G_2, G_3). In each instance, we simulated 1000 data sets and compared our method, both with known π_1 and using the rare disease approximation, to logistic regression.

	β_{G1}	β_{G2}	β_{G3}	β_{G4}	β_{G5}	β_X	β_{XG1}	β_{XG2}	β_{XG3}	β_{XG4}	β_{XG5}
True	0.18	0.18	0.00	0.18	0.00	0.30	0.19	0.00	0.00	0.19	0.00
X **correlated with** G_1				Logistic Regression							
Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01
CI(%)	95.3	94.9	95.3	94.3	93.6	95.4	95.5	94.0	94.9	94.1	95.6
						Symmetric Combination Estimator, known π_1					
Bias	0.00	-0.03	0.00	0.00	0.00	-0.01	0.05	0.01	0.00	-0.02	0.00
CI*(%)	95.8	93.3	95.6	94.3	95.4	94.2	92.5	92.7	95.6	93.5	95.4
MSE Eff	1.31	1.13	1.37	1.26	1.44	1.34	1.91	2.40	2.71	2.41	2.91
						Symmetric Combination Estimator, rare					
Bias	0.00	-0.03	0.00	0.00	0.00	0.01	0.01	0.01	0.00	-0.03	0.00
CI*(%)	96.4	92.5	96.3	94.8	95.7	95.5	95.3	93.5	95.8	90.3	95.3
MSE Eff	1.35	1.14	1.47	1.34	1.51	1.43	3.24	2.67	2.96	2.27	3.59
X **correlated with** G_2				Logistic Regression							
Bias	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
CI(%)	94.8	95.1	94.5	94.5	95.3	96.2	94.3	93.4	94.7	95.3	95.2
						Symmetric Combination Estimator, known π_1					
Bias	-0.01	-0.02	0.00	0.00	0.00	-0.03	-0.02	0.06	0.00	-0.02	0.00
CI*(%)	95.6	95.2	95.8	95.5	96.9	93.1	94.2	83.5	95.3	94.1	96.1
MSE Eff	1.33	1.27	1.35	1.32	1.42	1.11	2.79	1.54	2.84	2.63	3.21
						Symmetric Combination Estimator, rare					
Bias	-0.01	-0.02	0.00	0.00	0.00	-0.01	-0.05	0.05	0.00	-0.03	0.00
CI*(%)	95.8	94.9	95.9	95.1	96.3	96.0	87.7	84.7	95.3	90.3	97.0
MSE Eff	1.34	1.27	1.44	1.35	1.45	1.37	2.41	1.79	3.21	2.43	3.95
X **correlated with** G_3				Logistic Regression							
Bias	0.00	0.00	0.00	0.00	-0.01	0.00	0.01	0.00	0.00	0.00	0.01
CI(%)	94.9	94.8	96.0	94.9	95.2	95.0	95.9	95.0	95.6	94.7	94.3
						Symmetric Combination Estimator, known π_1					
Bias	-0.01	-0.02	0.01	0.00	-0.01	-0.03	-0.01	0.01	0.05	-0.02	0.00
CI*(%)	96.0	93.8	96.3	96.4	96.5	92.2	93.5	95.6	89.9	94.0	94.8
MSE Eff	1.33	1.16	1.34	1.25	1.34	1.15	2.56	2.62	1.63	2.33	2.89
						Symmetric Combination Estimator, rare					
Bias	-0.01	-0.03	0.01	0.00	-0.01	-0.01	-0.04	0.01	0.04	-0.03	0.00
CI*(%)	95.6	93.5	96.3	96.3	96.5	94.4	88.2	96.5	89.1	90.8	95.5
MSE Eff	1.41	1.16	1.35	1.28	1.37	1.39	2.37	3.01	1.78	2.17	3.62

CI: coverage of a 95% nominal confidence interval, calculated using asymptotic standard error. *CI**: coverage of a 95% nominal confidence interval, calculated using 100 bootstrap samples.
MSE Eff: mean squared error efficiency when compared to logistic regression.

B.2.4 Simulations with alternative distributions for G and X

Table B.4 presents the results of a simulation in which X and G are both multivariate with a combination of discrete and continuous components. G_1 and G_2 are correlated SNPs in Hardy-Weinberg equilibrium with minor allele frequencies (0.2, 0.3), and G_3 has a gamma distribution with shape = 20 and scale = 20 (to simulate a skewed polygenic risk score). X_1 is binary with frequency 0.5 and X_2 has a standard normal distribution. Here $\beta_G = \{\log(1.2), 0, \log(1.38)\}$, $\beta_X = \{\log(1.5), \log(1.14)\}$, $\beta_{GX} = \{\log(1.1), 0, 0, 0, 0, 0\}$, and the logistic intercept was selected to give a population disease rate of 0.05. Using these settings, 1000 data sets were simulated with 1000 cases and 1000 controls each.

Table B.4: Results of 1000 simulations with multivariate G and X, comparing the bias, coverage, and efficiency of standard logistic regression to our Symmetric Combination Estimator, both with known π_1 and using the rare disease approximation.

	β_{G1}	β_{G2}	β_{G3}	β_{X1}	β_{X2}	β_{X1G1}	β_{X1G2}	β_{X1G3}	β_{X2G1}	β_{X2G2}	β_{X2G3}
True	0.18	0.00	0.32	0.41	0.14	0.10	0.00	0.00	0.00	0.00	0.00
Logistic Regression											
Bias	0.01	-0.01	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00
CI(%)	94.4	94.3	95.2	94.6	94.0	95.7	94.7	94.7	94.5	95.4	94.6
Symmetric Combination Estimator, known π_1											
Bias	-0.02	0.00	-0.06	-0.01	-0.02	-0.02	0.00	0.01	0.00	0.00	0.00
CI*(%)	93.4	94.7	94.9	94.8	96.0	94.1	95.4	96.1	95.8	96.9	96.5
MSE Eff	1.48	1.66	1.61	2.44	2.75	2.22	2.67	2.84	2.90	2.80	3.03
Symmetric Combination Estimator, rare											
Bias	-0.01	0.00	-0.06	0.00	-0.02	-0.03	0.00	0.01	0.00	0.00	0.00
CI*(%)	93.3	94.5	95.3	94.8	95.8	94.1	95.6	96.3	95.6	96.9	96.2
MSE Eff	1.54	1.73	1.66	2.60	2.98	2.37	2.98	3.08	3.25	3.07	3.32

CI: coverage of a 95% nominal confidence interval, calculated using asymptotic standard error.
*CI**: coverage of a 95% nominal confidence interval, calculated using 100 bootstrap samples.
MSE Eff: mean squared error efficiency when compared to logistic regression.

B.2.5 Creating the polygenic risk score for the PLCO data analysis

Table B.5 displays the SNPs used in the calculation of the polygenic risk score for the analysis of the Prostate, Lung, Colorectal and Ovarian cancer screening trial data described in Section 3.4.1.

Table B.5: SNPs involved in creating the polygenic risk score, and their regression coefficients

RS Number	Coefficient
rs11249433	-0.02813492
rs1045485	-0.09307971
rs13387042	-0.26203658
rs4973768	0.08013260
rs10069690	0.06459363
rs10941679	0.09185539
rs889312	-0.00565121
rs17530068	0.09668742
rs2046210	0.09851217
rs1562430	-0.14871719
rs1011970	0.05329783
rs865686	-0.02913340
rs2380205	-0.01821032
rs10995190	-0.04275836
rs2981582	0.14008397
rs909116	0.04955235
rs614367	0.06438418
rs3803662	0.27080105
rs6504950	-0.17586244
rs8170	0.08570773
rs999737	-0.13737833

APPENDIX C

APPENDIX TO SECTION 4

Package 'caseControlGE'

Type Package

Title Semiparametric Gene-Environment Interactions in Case-Control Studies

Version 0.2

Author Alex Asher

Maintainer Alex Asher <alexasher@stat.tamu.edu>

Description

Implements the methods of Stalder et. al. (https://doi.org/10.1093/biomet/asx045) and Wang et. al. (forthcoming). These are retrospective estimators that assume Gene-Environment independence in the source population, but place no assumptions on the marginal distributions of genetic or environmental variables. G and E can be multivariate, and both continuous and discrete variables may be used.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Imports Rcpp (>= 0.12.16), ucminf (>= 1.1-3)

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 6.0.1

NeedsCompilation yes

Suggests knitr, rmarkdown, pander

VignetteBuilder knitr

R topics documented:

`caseControlGE-package`

Semiparametric Gene-Environment Interactions in Case-Control Studies

---

Description

An R package for analysis of gene-environment interactions in case-control studies, using distribution-free retrospective methodology. The method of Stalder et. al. (2017) and the improvement suggested by Wang et. al (2018) use a retrospective likelihood framework under the assumption of gene-environment independence (in the population) to gain efficiency when estimating the interaction effects of genetic and environmental variables.

Details

Both methods treat the genetic and environmental variables nonparametrically, facilitating the analysis of polygenic risk factors for which distributional assumptions are difficult to justify.

Keywords

case-control study; gene-environment interaction; genetic epidemiology; retrospective method; semiparametric analysis; pseudolikelihood; polygenic analysis

Contents

`spmle` implements method of Stalder et. al. (2017). Given binary response `D` (disease status), a vector or matrix of genetic risk factors `G`, a vector or matrix of environmental risks `E`, and the population disease rate `pi1`, `spmle` fits a model of the form `D ~ G * E` by maximizing the retrospective pseudolikelihood.

`spmleCombo` implements the method of of Wang et. al (2018) under the same set of assumptions as `spmle`. This function takes the same imput as `spmle` with the addition of `nboot` (number of bootstrap samples) and `ncores` (number of CPU cores to use simultaneously). `spmleCombo` produces estimates that, on average, have significantly smaller mean squared error than `spmle`, at the cost of increased computation to calculate the bootstrap standard error.

`simulateCC` simulates case-control data with a wide range of possible genetic and environmental variables.

methods for class `"spmle"` both `spmleCombo` and `spmle` return objects of class `"spmle"`. A range of S3 methods are provided: `anova.spmle`, `confint.spmle`, `logLik.spmle`, `model.matrix.spmle`, `plot.spmle`, `predict.spmle`, `print.spmle`, `print.summary.spmle`, `summary.spmle`, `vcov.spmle`.

References

Stalder, O., Asher, A., Liang, L., Carroll, R. J., Ma, Y., and Chatterjee, N. (2017). *Semiparametric analysis of complex polygenic gene-environment interactions in case-control studies.* Biometrika, 104, 801–812.

Wang, T., Asher, A., Carroll, R. J. (2018). *Improved Semiparametric Analysis of Polygenic Gene-Environment Interactions in Case-Control Studies* Unpublished.

`predict.spmle` *Predict method for spmle objects*

Description

Obtains predictions from a fitted spmle object.

Usage

```
## S3 method for class 'spmle'
predict(object, newdata, se.fit = FALSE,
  interval = c("none", "confidence"), level = 0.95,
  type = c("link", "response"), na.action = na.pass, ...)
```

Arguments

`object`	of class inheriting from "spmle"
`newdata`	an optional list or data frame in which to look for variables to use when making predictions. If omitted, the fitted values are used.
`se.fit`	a switch indicating if standard errors are required.
`interval`	Type of interval calculation. Can be abbreviated. Prediction intervals are not meaningful for binary responses and are not allowed.
`level`	confidence level.
`type`	the type of prediction required. The default is `"link"` which uses the logit scale of the linear predictors, giving the log odds. The alternative `"response"` uses the probability scale, giving `Pr(D=1|G,E)`.
`na.action`	function determining what should be done with missing values in newdata. The default is to predict NA.
`...`	further arguments passed to or from other methods.

Details

`predict.spmle` produces predicted values, obtained by evaluating the `spmle` function in the frame newdata (which defaults to `model.frame(object)`). If the logical `se.fit` is `TRUE`, standard errors of the predictions are calculated (only in the link scale). Setting `interval="confidence"` specifies computation of confidence intervals at the specified level.

Value

`predict.spmle` produces a vector of predictions or a matrix of predictions and bounds with column names fit, lwr, and upr if interval is set.

If `se.fit` is `TRUE`, a list with the following components is returned:

`fit` vector or matrix as above

`se.fit` vector with standard error of predicted means, in the link scale

`residual.scale` residual standard deviation

`df` degrees of freedom for residual

`simulateCC` *Simulate case-control data with multivariate, possibly dependent genetic and environmental components.*

Description

`simulateCC` simulates case-control data to be analyzed by `spmle`, `spmleCombo`, logistic regression, or other methods.

Usage

```
simulateCC(ncase, ncontrol, beta0, betaG_SNP, betaG_normPRS,
  betaG_gammaPRS, betaG_bimodalPRS, betaE_bin, betaE_norm,
  betaGE_SNP_bin, betaGE_normPRS_bin, betaGE_gammaPRS_bin,
  betaGE_bimodalPRS_bin, betaGE_SNP_norm, betaGE_normPRS_norm,
  betaGE_gammaPRS_norm, betaGE_bimodalPRS_norm, MAF,
  SNP_cor = 0, G_normPRS_cor = 0, E_norm_cor = 0, E_bin_freq,
  regress_E_bin_on_G_SNP, regress_E_bin_on_G_normPRS,
  regress_E_bin_on_G_gammaPRS, regress_E_bin_on_G_bimodalPRS,
```

```
  regress_E_norm_on_G_SNP, regress_E_norm_on_G_normPRS,
  regress_E_norm_on_G_gammaPRS, regress_E_norm_on_G_bimodalPRS,
  control = list())
```

Arguments

`ncase, ncontrol`

number of cases and controls, both must be positive integers.

`beta0` logistic intercept, required. Can be manipulated to change the population disease rate.

`betaG_SNP, betaG_normPRS, betaG_gammaPRS, betaG_bimodalPRS`

optional coefficients for genetic main effects (at least one must be specified). Genetic variables can include SNPs and polygenic risk scores with standard normal, gamma(20, 20), and bimodal distributions. Vector valued coefficients produce multivariate genetic data. When simulating SNPs, you must provide `MAF` with the same length as `betaG_SNP`.

`betaE_bin, betaE_norm`

optional coefficients for environmental variable main effects (at least one must be specified). Environmental variables can include binary and standard normal random variables. Vector valued coefficients produce multivariate environmental data.

`betaGE_SNP_bin, betaGE_normPRS_bin, betaGE_gammaPRS_bin,`

`betaGE_bimodalPRS_bin, betaGE_SNP_norm, betaGE_normPRS_norm,`

`betaGE_gammaPRS_norm, betaGE_bimodalPRS_norm`

coefficients for multiplicative G*E interaction effects. The length of the coefficient of any given G*E interaction must equal the product of the lengths of the coefficients of the corresponding G and E main effects.

`MAF` Minor Allele Frequency of SNPs. This vector is the same length as `beta_G_SNP` and has values between 0 and 1. The MAF is used to generate SNP data that

is in Hardy-Weinberg Equilibrium. This specifies `Pr[G=(0, 1, 2)] = [(1-MAF)^2, 2*MAF(1-MAF), MAF^2]`.

`SNP_cor` scalar specifying the correlation between adjacent SNPs. SNPs are simulated by generating multivariate normal random draws with an AR1(`SNP_cor`) covariance matrix. These normal draws are then trichotomized according to HWE to simulate SNPs. Default `SNP_cor = 0`.

`G_normPRS_cor`

correlation matrix for multivariate normal polygenic risk scores. In the bivariate case, a 2x2 matrix or a scalar (for correlation) are accepted. Default `G_normPRS_cor = 0` generates independent normal polygenic risk scores.

`E_norm_cor` correlation matrix for multivariate normal environmental variable. In the bivariate case, a 2x2 matrix or a scalar (for correlation) are accepted. Default `E_norm_cor = 0` generates independent normal environmental variables.

`E_bin_freq` marginal probability that E_bin = 1. Must have length equal to `length(betaE_bin)` and values between 0 and 1.

`regress_E_bin_on_G_SNP, regress_E_bin_on_G_normPRS,`

`regress_E_bin_on_G_gammaPRS, regress_E_bin_on_G_bimodalPRS`

allow the simulation of case-control data that violates the G-E independence assumption. If specified, binary environmental variables will be generated with `Pr(E=1|G) =plogis[regress_E_bin_on_G * G + qlogis(E_bin_freq)]`. If these arguments are missing, `NULL`, or all `0`s, the binary environmental variables will be independent of the genetic variables. If specified, the length of the regression argument must equal the product of the lengths of the coefficients of the corresponding G and E main effects.

`regress_E_norm_on_G_SNP, regress_E_norm_on_G_normPRS,`

`regress_E_norm_on_G_gammaPRS, regress_E_norm_on_G_bimodalPRS`

allow the simulation of case-control data that violates the G-E independence

assumption. If specified, normal environmental variables will be generated from a Normal(`regress_E_norm_on_G * G`, 1) distribution. If these arguments are missing, `NULL`, or all `0`s, the normal environmental variables will be independent of the genetic variables. If specified, the length of the regression argument must equal the product of the lengths of the coefficients of the corresponding G and E main effects.

`control` a list of control parameters, all of which are ignored except `trace`, a scalar. If `trace > -1`, information about the simulation (e.g. population disease rate, correlations between SNPs, etc.) is produced. Default `trace=0`.

Details

The user can specify up to four types of genetic variables, each of which can be multivariate: SNPs with additive effects under Hardy-Weinberg Equilibrium and polygenic risk scores with standard normal, gamma(20, 20), and bimodal distributions. Two types of environmental variables (binary and normal) can also be potentially multivariate.

SNPs may be generated in linkage disequilibrium, yielding correlated SNPs. Multivariate normal polygenic risk scores may have a user-specified correlation matrix, as may multivariate normal environmental variables. Correlation may also be specified between genetic and environmental variables to simulate data in violation of the gene-environment independence assumption.

The number of variables generated is determined by the length of the betas given. If you specify `betaG_normPRS = c(log(1.1), log(1.2))` and `betaE_bin = c(log(1.2), log(1.2), log(1.2))`, you will get two `G` variables (normally distributed polygenic risk scores), and three `E` variables (with binary distributions). In this example, you would supply a vector of length 6 for `betaGE_normPRS_bin`.

If both G and E are multivariate, `beta_GE_` and `regress_E_` arguments iterate G quickly and E slowly. In the example above, `betaGE_normPRS_bin` is ordered (G1*E1, G2*E1,

G1*E2, G2*E2, G1*E3, G2*E3).

`simulateCC` works by simulating a population using user-specified parameters, then selecting `ncase` cases and `ncontrol` controls as the case-control sample. The entire population is used to calculate disease prevalance (reported when `control$trace > -1`). Case-control studies deliberately oversample cases, so the distribution of G and E in the sample may be quite different from the distribution of G and E in the population (especially for variables that are strongly correlated with disease status).

Value

`simulateCC` produces a list with three elements:

`D` a binary vector with `ncontrol` zeros and `ncase` ones.

`G` a matrix with `ncontrol + ncase` rows and a column for each genetic variable. Genetic variables are ordered: SNPs, normal PRSs, gamma PRSs, bimodal PRSs.

`E` a matrix with `ncontrol + ncase` rows and a column for each environmental variable. Environmental variables are ordered: binary, normal.

See Also

`spmleCombo`, `spmle`

Examples

```
set.seed(2018)
# Simulation from Table 1 in Stalder et. al. (2017)
dat = simulateCC(ncase=1000,
                 ncontrol=1000,
                 beta0=-4.165,
                 betaG_SNP=c(log(1.2), log(1.2), 0, log(1.2), 0),
                 betaE_bin=log(1.5),
```

```
                betaGE_SNP_bin=c(log(1.3), 0, 0, log(1.3), 0),
                MAF=c(0.1, 0.3, 0.3, 0.3, 0.1),
                SNP_cor=0.7,
                E_bin_freq=0.5)

# Simulation with 5 SNPs and a single normal environmental variable
# that is dependent on G1 with an R^2 of 0.001.
# True population disease rate in this simulation is 0.03.
# This simulation scenario was used in the Supplementary Material
# of Stalder et. al. (2017)
dat2 = simulateCC(ncase=1000,
                ncontrol=1000,
                beta0=-3.89,
                betaG_SNP=c(log(1.2), log(1.2), 0, log(1.2), 0),
                betaE_norm=log(1.5)/(qnorm(0.75)-qnorm(0.25)),
                betaGE_SNP_norm=c(log(1.3), 0, 0, log(1.3), 0) /
                                (qnorm(0.75)-qnorm(0.25)),
                MAF=c(0.1, 0.3, 0.3, 0.3, 0.1),
                SNP_cor=0.7,
                regress_E_norm_on_G_SNP=c(sqrt(0.001),rep(0,4)),
                control=list(trace=1))
```

`spmle` *Semiparametric Maximum Pseudolieklihood Estimator for Case-Control Studies Under G-E Independence.*

Description

`spmle` maximizes the retrospective pseudolikelihood of case-control data under the assumption of G-E independence in the underlying population. The marginal distributions of G and E

are treated nonparametrically.

Usage

```
spmle(D, G, E, pi1, data, control=list(), swap=FALSE, startvals)
```

Arguments

`D`	a binary vector of disease status (1=case, 0=control).
`G`	a vector or matrix (if multivariate) containing genetic data. Can be continuous, discrete, or a combination.
`E`	a vector or matrix (if multivariate) containing environmental data. Can be continuous, discrete, or a combination.
`pi1`	the population disease rate, a scalar in [0, 1) or the string "rare". Using `pi1=0` is the rare disease approximation.
`data`	an optional list or environment containing the variables in the model. If not found in `data`, the variables are taken from the environment from which `spmle` is called.
`control`	a list of control parameters that allow the user to control the optimization algorithm. See 'Details'.
`swap`	a logical scalar rarely of interest to the end user. Dependence on the distributions of G and E are removed using different methods; this switch swaps them to produce a symmetric estimator with identical properties to the SPMLE. Default `FALSE`.
`startvals`	an optional numeric vector of coefficient starting values for optimization. Usually left blank, in which case logistic regression estimates are used as starting values.

Details

This function applies the method of Stalder et. al. (2017) to maximize the retrospective pseudolikelihood of case-control data under the assumption of G-E independence. It currently supports the model with G and E main effects and a multiplicative G*E interaction.

The `control` argument is a list that controls the behavior of the optimization algorithm `ucminf` from the **ucminf** package. When `ucminf` works, it works brilliantly (typically more than twice as fast as the next-fastest algorithm). But it has a nasty habit of declaring convergence before actually converging. To address this, `spmle` checks the maximum gradient at "convergence", and can rerun the optimization using different starting values. The `control` argument can supply any of the following components:

`max_grad_tol` maximum allowable gradient at convergence. `spmle` does not consider the optimization to have converged if the maximum gradient `>` `max_grad_tol` when `ucminf` stops. Default `max_grad_tol = 0.001`.

`num_retries` number of times to retry optimization. An error is produced if the optimization has not converged after `num_retries`. Different starting values are used for each retry. Default `num_retries = 2`.

`use_hess` a logical value instructing `spmle` to use the analytic hessian to precondition the optimization. This brings significant speed benefits, and is one reason `ucminf` is so fast. For unknown reasons, preconditioning causes computers with certain Intel CPUs to prematurely terminate iterating. By default, `use_hess = TRUE`, but if you notice that `ucminf` never converges during the first attempt, try setting `use_hess = FALSE`.

`trace` a scalar or logical value that is used by both `spmle` and `ucminf` to control the printing of detailed tracing information. If TRUE or > 0, details of each `ucminf` iteration are printed. If FALSE or 0, `ucminf` iteration details are suppressed but `spmle` still prints optimization retries. If `trace < 0` nothing is printed. Default `trace = 0`.

additional control parameters not used by `spmle`, but are passed to `ucminf`. Note that

the ucminf algorithm has four stopping criteria, and ucminf will declare convergence if any one of them has been met. The ucminf control parameter "grtol" controls ucminf's gradient stopping criterion, which defaults to 1e-6. grtol should not be set larger than the spmle control parameter max_grad_tol.

Value

an object of class "spmle". The function summary (i.e., summary.spmle) can be used to obtain or print a summary of the results.

The function anova (i.e., anova.spmle) will conduct likelihood-ratio tests comparing one spmle object to another. These are valid tests because the loglikelihood reported by logLik.spmle is accurate up to an additive constant. However anova should not be used to compare an spmle object to a model fit by a different method.

predict.spmle, the predict method for S3 class "spmle", can predict the expected response (on logistic or probability scales), compute confidence intervals for the expected response, and provide standard errors.

The generic accessor functions coefficients, fitted.values and residuals can be used to extract various useful features of the value returned by spmle.

An object of class "spmle" is a list containing at least the following components:

coefficients a named vector of coefficients

pi1 the value of pi1 used during the analysis

SE standard error estimate of coefficients

cov estimated covariance matrix of coefficients

glm_fit a logistic regression model fit using the same model as spmle

call the matched call

formula the formula used

`data` the `data argument`

`model` the model frame

`terms` the `terms` object used

`linear.predictors` the linear fit on the logistic link scale

`fitted.values` the fitted values on the probability scale

`residuals` the Pearson residuals

`null.deviance` the deviance for the null model. Deviance = `-2*logLik`.

`df.residual` the residual degrees of freedom

`df.null` the residual degrees of freedom for the null model

`rank` the numeric rank of the fitted linear model (i.e. the number of parameters estimated)

`nobs` number of observations

`ncase` number of cases

`ncontrol` number of controls

`spmle` objects created by `spmle()` additionally have components `logLik` (log pseudolikelihood), `deviance` (-2 * log pseudolikelihood), `aic`, `bic`, `ucminf` (optimization output), and matrices `H_inv`, `Sigma`, `zeta0`, and `zeta1`, which are used in calculating the asymptotic estimate of standard error.

References

Stalder, O., Asher, A., Liang, L., Carroll, R. J., Ma, Y., and Chatterjee, N. (2017). *Semi-parametric analysis of complex polygenic gene-environment interactions in case-control studies.* Biometrika, 104, 801–812.

See Also

`spmleCombo` for a slower but more precise estimator, `simulateCC` to simulate data

Examples

```
# Simulation from Table 1 in Stalder et. al. (2017)
set.seed(2018)
dat = simulateCC(ncase=500, ncontrol=500, beta0=-4.165,
                 betaG_SNP=c(log(1.2), log(1.2), 0, log(1.2), 0),
                 betaE_bin=log(1.5),
                 betaGE_SNP_bin=c(log(1.3), 0, 0, log(1.3), 0),
                 MAF=c(0.1, 0.3, 0.3, 0.3, 0.1),
                 SNP_cor=0.7, E_bin_freq=0.5)

# SPMLE with known population disease rate of 0.03
spmle(D=D, G=G, E=E, pi1=0.03, data=dat)

# Simulation with a single SNP and a single binary environmental
# variable.
# True population disease rate in this simulation is 0.03.
# This simulation scenario was used in the Supplementary Material
# of Stalder et. al. (2017) to compare performance against the
# less flexible method of Chatterjee and Carroll (2005), which is
# available as the function as snp.logistic in the Bioconductor
# package CGEN.
dat2 = simulateCC(ncase=100, ncontrol=100, beta0=-3.77,
                  betaG_SNP=log(1.2), betaE_bin=log(1.5),
                  betaGE_SNP_bin=log(1.3), MAF=0.1,
                  E_bin_freq=0.5)

# SPMLE using the rare disease assumption, optimization tracing,
# and no hessian preconditioning.
```

```
spmle(D=D, G=G, E=E, pi1=0, data=dat2,
      control=list(trace=0, use_hess=FALSE))
```

`spmleCombo` *Improved Semiparametric Estimator for Case-Control Studies Under G-E Independence.*

Description

`spmleCombo` estimates Gene-Environment interactions in case-control data under the assumption of G-E independence in the underlying population. This is an improved version of `spmle` that, on average, has significantly smaller mean squared error than `spmle`, at the cost of increased computing time.

Usage

```
spmleCombo(D, G, E, pi1, data, nboot = 50, ncores = 1,
  control = list(), startvals)
```

Arguments

`D`	a binary vector of disease status (1=case, 0=control).
`G`	a vector or matrix (if multivariate) containing genetic data. Can be continuous, discrete, or a combination.
`E`	a vector or matrix (if multivariate) containing environmental data. Can be continuous, discrete, or a combination.
`pi1`	the population disease rate, a scalar in [0, 1) or the string "rare". Using `pi1=0` is the rare disease approximation.
`data`	an optional list or environment containing the variables in the model. If not found in `data`, the variables are taken from the environment from which `spmleCombo` is called.

`nboot` the number of bootstraps to use when estimating the standard error, an integer. Setting `nboot=0` disables the bootstrap and uses the asymptotic standard error estimate (not recommended because the asymptotic SE often has poor coverage: setting `nboot=0` will trigger a warning). Default `nboot = 50`.

`ncores` the number of cpu cores to use when parallelizing bootstraps, an integer. Default `ncores = 1` executes the bootstrap sequentially.

`control` a list of control parameters that allow the user to control the optimization algorithm. See 'Details'.

`startvals` an optional numeric vector of coefficient starting values for optimization. Usually left blank, in which case logistic regression estimates are used as starting values.

Details

This function calculates the Symmetric Combination Estimator of Wang et. al. (2018), which improves estimation efficiency in case-control studies of gene-environment interactions while treating the marginal distributions of G and E nonparametrically.

This is done by calling `spmle` twice (once with `swap=TRUE`) to generate two symmetric estimates, which are then combined using a GLS approach. It currently supports the model with G and E main effects and a multiplicative G*E interaction.

The `control` argument is a list that controls the behavior of the optimization algorithm `ucminf` from the **ucminf** package. When `ucminf` works, it works brilliantly (typically more than twice as fast as the next-fastest algorithm). But it has a nasty habit of declaring convergence before actually converging. To address this, `spmleCombo` checks the maximum gradient at "convergence", and can rerun the optimization using different starting values. The `control` argument can supply any of the following components:

`max_grad_tol` maximum allowable gradient at convergence. `spmleCombo` does not consider the optimization to have converged if the maximum gradient `> max_grad_tol`

when ucminf stops. Default max_grad_tol = 0.001.

num_retries number of times to retry optimization. An error is produced if the optimization has not converged after num_retries. Different starting values are used for each retry. Default num_retries = 2.

use_hess a logical value instructing spmleCombo to use the analytic hessian to precondition the optimization. This brings significant speed benefits, and is one reason ucminf is so fast. For unknown reasons, preconditioning causes computers with certain Intel CPUs to prematurely terminate iterating. By default, use_hess = TRUE, but if you notice that ucminf never converges during the first attempt, try setting use_hess = FALSE.

trace a scalar or logical value that is used by both spmleCombo and ucminf to control the printing of detailed tracing information. If TRUE or > 0, details of each ucminf iteration are printed. If FALSE or 0, ucminf iteration details are suppressed but spmleCombo still prints optimization retries. If trace < 0 nothing is printed. Default trace = 0.

additional control parameters not used by spmleCombo, but are passed to ucminf. Note that the ucminf algorithm has four stopping criteria, and ucminf will declare convergence if any one of them has been met. The ucminf control parameter "grtol" controls ucminf's gradient stopping criterion, which defaults to 1e-6. grtol should not be set larger than the spmleCombo control parameter max_grad_tol.

Value

an object of class "spmle". The function summary (i.e., summary.spmle) can be used to obtain or print a summary of the results. The Symmetric Combination Estimator is not a maximum (pseudo)likelihood estimator like spmle; it is the optimal combination of two such estimators. As such, it has no associated loglikelihood and the function anova.spmle cannot be used to compare models fit using spmleCombo.

predict.spmle, the predict method for S3 class "spmle", can predict the expected response (on logistic or probability scales), compute confidence intervals for the expected re-

sponse, and provide standard errors.

The generic accessor functions `coefficients`, `fitted.values` and `residuals` can be used to extract various useful features of the value returned by `spmleCombo`.

An object of class `"spmle"` is a list containing at least the following components:

`coefficients` a named vector of coefficients

`pi1` the value of pi1 used during the analysis

`SE` standard error estimate of coefficients

`cov` estimated covariance matrix of coefficients

`glm_fit` a logistic regression model fit using the same model as `spmleCombo`

`call` the matched call

`formula` the formula used

`data` the `data argument`

`model` the model frame

`terms` the `terms` object used

`linear.predictors` the linear fit on the logistic link scale

`fitted.values` the fitted values on the probability scale

`residuals` the Pearson residuals

`null.deviance` the deviance for the null model

`df.residual` the residual degrees of freedom

`df.null` the residual degrees of freedom for the null model

`rank` the numeric rank of the fitted linear model (i.e. the number of parameters estimated)

`nobs` number of observations

`ncase` number of cases

`ncontrol` number of controls

`spmle` objects created by `spmleCombo()` additionally have components `spmle_E` (model from `spmle` that profiled out the distribution of E), `spmle_G` (model from `spmle` that profiled out the distribution of G with `swap=TRUE`), and `bootstraps` (matrix of bootstrapped parameter estimates, if `nboot > 0`).

References

Stalder, O., Asher, A., Liang, L., Carroll, R. J., Ma, Y., and Chatterjee, N. (2017). *Semiparametric analysis of complex polygenic gene-environment interactions in case-control studies.* Biometrika, 104, 801–812.

Wang, T., Asher, A., Carroll, R. J. (2018). *Improved Semiparametric Analysis of Polygenic Gene-Environment Interactions in Case-Control Studies* Unpublished.

See Also

`spmle`, `simulateCC` to simulate data

Examples

```
# Simulation from Table 1 in Stalder et. al. (2017)
set.seed(2018)
dat = simulateCC(ncase=500, ncontrol=500, beta0=-4.165,
                 betaG_SNP=c(log(1.2), log(1.2), 0, log(1.2), 0),
                 betaE_bin=log(1.5),
                 betaGE_SNP_bin=c(log(1.3), 0, 0, log(1.3), 0),
                 MAF=c(0.1, 0.3, 0.3, 0.3, 0.1),
                 SNP_cor=0.7, E_bin_freq=0.5)

# SPMLE with known population disease rate of 0.03
# and asymptotic SE estimates
spmleCombo(D=D, G=G, E=E, pi1=0.03, data=dat, nboot=0)
```