The Gene-environment Independence Assumption in the Analysis of Case-control Data

by

Hao Luo

B.Sc., Nanjing University, 2010

M.Sc., The University of British Columbia, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Statistics)

The University of British Columbia

(Vancouver)

December 2016

Abstract

In this thesis, we consider the problem of exploiting the gene-environment independence assumption in a case-control study inferring the joint effect of genotype and environmental exposure on disease risk.

We first take a detour and develop the constrained maximum likelihood estimation theory for parameters arising from a partially identified model, where some parameters of the model may only be identified through constraints imposed by additional assumptions. We show that, under certain conditions, the constrained maximum likelihood estimator exists and locally maximizes the likelihood function subject to constraints. Moreover, we study the asymptotic distribution of the estimator and propose a numerical algorithm for estimating parameters.

Next, we use the frequentist approach to analyze case-control data under the gene-environment independence assumption. By transforming the problem into a constrained maximum likelihood estimation problem, we are able to derive the asymptotic distribution of the estimator in a closed form. We then show that exploiting the gene-environment independence assumption indeed improves estimation efficiency. Also, we propose an easy-to-implement numerical algorithm for finding estimates in practice.

Furthermore, we approach the problem in a Bayesian framework. By introducing a different parameterization of the underlying model for case-control data, we are able to define a prior structure reflecting the gene-environment independence assumption and develop an efficient numerical algorithm for the computation of the posterior distribution. The proposed Bayesian method is further generalized to address the concern about the validity of the gene-environment independence assumption.

Finally, we consider a special variant of the standard case-control design, the case-only design, and study the analysis of case-only data under the gene-environment independence assumption and the rare disease assumption. We show that the Bayesian method for analyzing case-control data is readily applicable for the analysis of case-only data, allowing the flexibility of incorporating different prior beliefs on disease prevalence.

Preface

A version of Chapter 2 has been posted online at arxiv.org (Luo et al. [17]) and will be submitted for publication. The idea was motivated in order to solve the problem in Chapter 3. Under the supervision of Dr. Gustafson, I derived the theoretical results, conducted all computational work, and wrote the majority of the manuscript. Dr. Bouchard-Côté and Dr. Cohen Freue provided thoughtful suggestions and helped with the revisions of the manuscript.

Chapter 3, 4, and 5 are based on manuscripts collaborated with Dr. Gustafson, Dr. Burstyn, Dr. Bouchard-Côté, and Dr. Cohen Freue. Some related work concerning the interaction between a binary genotype and a binary environmental exposure has been published (Luo et al. [18]). Versions of these chapters will be submitted for publication. Dr. Gustafson initiated the research problem and inspired me to the right direction when I had difficulties moving beyond the simple set-up of binary genetic and environmental factors. I conducted all derivations, all computational work and the majority of the writing. Dr. Bouchard-Côté and Dr. Cohen Freue both gave me constructive criticism and helpful suggestions, which prompted more thinking about the proposed methods and greatly improved these manuscripts. Dr. Burstyn gave valuable inputs from the perspective of an epidemiologist to make our research more relevant to the application in epidemiology.

Table of Contents

List of Tables

List of Figures

Acknowledgments

My deepest debt of gratitude goes to my supervisor, Professor Paul Gustafson, for his continuous support of my Ph.D. studies. His guidance helped me build my own skills to ask meaningful questions and find appropriate answers. His patience and encouragement helped me go through the most difficult time when I got stuck in my research. Without any of these, I would never have been able to reach this far.

I would like to thank the members of my supervisory committee, Professor Alexandre Bouchard-Côté, Professor Gabriela Cohen-Freue, and Professor Igor Burstyn, for their inspiring questions and insightful comments. I am grateful to faculty members and staff of the Department of Statistics at UBC for providing such a nice academic environment. I would like to thank my fellow students and friends at UBC for making my study at UBC so enjoyable. Special thanks to Yanling Cai, Xin Geng, Yi Huang, Yang Liu, Ji Lv, Mengzhe Shen, Chumeng Wu, Hui Yang, Hongyang Zhang, and Tingting Zhao.

Last, I leave the warmest part of my heart for my family, my parents (Zhongliang Luo and Liping Wu) and my wife (Shihui Zhong), for their unconditional support and love. Without them, everything is impossible.

Dedication

to my parents and my wife.

Chapter 1

Introduction

Genetic and environmental factors may jointly influence the risk of many complex diseases. Individuals with different genotypes may be affected differently by exposure to the same environmental factors, and have different disease phenotypes as the result of gene-environment interactions. For example, heavy smokers tend to have higher risk of bladder cancer if they also carry NAT2 slow acetylators genotypes [10]. Thus, epidemiologists are interested in inferring gene-environment interaction. The study of gene-environment interactions will lead to better understanding of the biological mechanisms and pathological processes that contribute to the development of complex diseases. Such an understanding can guide the development of more efficient measures for preventing or even curing disease. If an individual carries a genotype that confers susceptibility to a certain disorder in a particular environment, then the disease may be prevented by reducing exposure to the environment.

1.1 Analysis of case-control data

The case-control study design is popular in studies of gene-environment interaction because it allows a better allocation of resources, where data can be collected for more cases. The traditional method for analyzing case-control data is fitting prospective logistic regression models regardless of the nature that data are collected retrospectively. Prentice and Pyke [24] showed that this method will lead to a

consistent estimator of the gene-environment interaction. The traditional estimator does not rely on any *a priori* assumptions about the joint distribution of genotype and environmental exposure. In many settings, however, it is biologically plausible to assert that genotype and environmental exposure are independent of one another in the source population (hereafter the *gene-environment independence assumption*, or GEI for short), since genotype arises from the random assortment of chromosomes carrying genes in meiosis [26]. Therefore, more efficient estimators of the gene-environment interaction from case-control data may be available by exploiting this assumption.

Many authors have investigated the possible benefits of analyzing case-control data under the GEI assumption or its variants ([23], [28], [3], [21], [22], [4]). For instance, Umbach and Weinberg [28] proposed the application of log-linear models to case-control cell counts, using a form of the GEI assumption that asserts gene-environment independence within the population of controls. These authors acknowledged that this form of the assumption is not natural, particularly as acquisition of genotype and environmental exposure are temporally antecedent to the disease. They justified such an assumption by pointing out that if the disease is rare then GEI in the source population will imply near GEI in the control population. Moreover, they showed that the estimator of the gene-environment interaction derived under this form of the GEI assumption differs from the traditional estimator, with a smaller asymptotic variance.

More recently, Chatterjee and Carroll [3] studied the problem of maximum-likelihood estimation for case-control data, assuming GEI in the source population. They argued that, contrary to intuition, the intercept in the prospective relationship can in fact be identified under the GEI assumption. Further, they proposed a profile likelihood technique to obtain the maximum likelihood estimator based on the retrospective likelihood, and presented simulation results to show that their GEI estimator has considerably lower variance than the traditional estimator. However, looking specifically at the situation where both genotype and environmental exposure are binary, Chen and Chen [4] found no efficiency gain for the GEI estimator over the traditional estimator. They claimed that estimating the intercept term in the prospective relationship uses up the additional information inherent in the GEI assumption.

With different frameworks inducing different claims about efficiency gains associated with the GEI assumption, epidemiologists are uncertain about the suitability of analytical strategies relying upon the assumption. This thesis brings clarity to this issue. In Chapter 2, we first take a detour to develop the constrained maximum likelihood estimation theory for estimating parameters arising from a partially identified model with some equality constraints introduced by additional assumptions. This theory is then applied in Chapter 3 for analyzing case-control data under the GEI assumption through a frequentist approach.

Another big hindrance for the use of GEI-based methods is their non-robustness to the violation of the GEI assumption. Albert et al. [2] discussed different scenarios where an environmental factor is associated with a genetic marker. For example, they may be correlated if both are associated with other (uncontrolled confounding) factors. Different methods have been proposed that relaxes the GEI assumption. Mukherjee and Chatterjee [21] developed an empirical Bayes-type shrinkage estimator to trade off between bias and efficiency. Mukherjee et al. [22] proposed a full Bayesian analysis and design strategy to incorporate prior belief on the assumption of gene-environment independence. However, these methods all focused on the assumption that genotype and environmental exposure are independent in the population of controls. In Chapter 4, we develop a Bayesian framework for analyzing case-control data under the GEI assumption (in the source population), and then generalize it to allow uncertainty about the assumption.

1.2 Analysis of case-only data

Piegorsch and Weinberg [23] noticed that, when genotype and environmental exposure are independent of each other in the control population, the gene-environment interaction odds ratio can be estimated using their association odds ratio in cases alone. Thus, the case-only design, which only collects data on diseased subjects, serves as an alternative to the standard case-control design for studying the gene-environment interaction effect, assuming GEI among controls. In a recent systematic review, Dennis et al. [6] showed no substantial difference between the case-only and case-control interaction estimates when the assumption indeed holds. Comparing to the standard case-control design, the case-only design has several

advantages. It not only avoids the difficulty of selecting appropriate controls, but also achieves better precision for estimating interaction effects. More importantly, it greatly saves study resources.

Despite its potential value, there are considerable concerns about the validity of the traditional case-only method because of its susceptibility to bias. Albert et al. [2] pointed out that inference made with the traditional case-only method can be highly sensitive to the assumption of GEI among controls. Even for small amounts of gene-environment association among controls, the type I error for testing the interaction effects can be seriously inflated and/or the case-only estimator of the interaction effect can be greatly biased. Thus, conclusions drawn from the traditional case-only method regarding the existence and/or the magnitude of the gene-environment interaction effect can sometimes be misleading.

The assumption of GEI among controls approximately holds under the GEI assumption and the rare disease assumption. Even though the GEI assumption can be supported in theory by the principle of the random assortment of alleles at the time of gamete formation (Mendel's second law [26]), and in practice by empirical evidence for some genetic variants and environmental factors, the rare disease assumption is not rigorously defined. It is not clear at what disease prevalence the assumption of GEI among controls will approximately hold when the GEI assumption actually holds in the source population. Indeed, Gatto et al. [8] reported that, under the GEI assumption, the gene-environment association in the control population may still not be negligible when the disease is only modestly rare. Therefore, the traditional case-only method may still produce substantial bias for the estimation of the gene-environment interaction effect, when the GEI assumption holds and the disease is rare but not extremely rare. In Chapter 5, we investigate the relationship between the disease prevalence and the performance of the traditional case-only method. We also apply the Bayesian framework developed in Chapter 4 for analyzing case-only data, where the rare disease assumption is clearly quantified through an appropriate prior distribution on disease prevalence.

Chapter 2

The Constrained Maximum Likelihood Estimation with Partially Identified Models

2.1 Introduction

In some scientific studies, due to constraints of logistics and/or resources, data are not collected in the ideal way. Consequently, the available data may only partially identify the statistical model under consideration, i.e., parameters of the statistical model are identified up to a set of possible values instead of just one single value. The set of parameter values that correspond to the same distribution of observables is usually termed the identification region. Manski [19] gives an overview of partial identification and covers many scenarios where partial identification may arise.

Of course, point-identification is preferred as it is fundamental for consistent point estimation and ensures many nice properties of model-based parameter estimators. With a partially identified model, when possible one may impose some reasonable assumptions to achieve point-identification. Under such assumptions, the parameter vector is restricted to a subset of the original parameter space. If this constrained parameter space has only a single point of intersection with the identification region, then the parameter vector is uniquely identified.

In this chapter, we study the maximum-likelihood estimation of parameters arising from a partially identified model with some equality constraints introduced by additional assumptions. In particular, we consider the scenario where there exists a special re-parameterization of all parameters of the model, which is termed a transparent re-parameterization by Gustafson et al. [14], such that the distribution of observables is completely determined by a proper subset of parameters after transformation.

In the situation of adding parameter constraints to a model which is identified even without the constraints, Aitchison and Silvey [1] studied the large-sample behavior of maximum-likelihood estimators via a Lagrange multiplier approach. However, the assumption that the unconstrained version of the model is identified is embedded in their approach. Therefore, our work extends their theory to the situation that identification is only obtained via imposition of the constraints.

The rest of this chapter is organized as follows. We first introduce some general notation and give a mathematical formulation of the problem. We then prove the existence of the constrained maximum likelihood estimate and show that the estimator is asymptotically normally distributed. A numerical algorithm for computing the constrained maximum likelihood estimate is also developed. We then consider an example problem and use a simulation study to compare the performance of the proposed method and the general method, which does not depend on constraints, to investigate the effect of imposing additional assumptions with a partially identified model. Moreover, we comment on a special situation where there is no benefit in terms of estimation efficiency.

2.2 Statistical problem

Suppose our data consist of n *i.i.d.* observations $\mathbf{x} = (x_1, \ldots, x_n)$. The statistical model underlying the data is assumed to be initially parameterized in scientific terms via a vector of s parameters. Let $\omega = (\omega_1, \ldots, \omega_s)$ be a re-parameterization of the original parameters such that the log-likelihood function ℓ for the observed data can be completely determined by its first r elements, say $\phi = (\omega_1, \ldots, \omega_r)$, through

$$\ell(\mathbf{x}, \phi) = \sum_{i=1}^{n} \log f(x_i, \phi),$$

where $f(x,\phi)$ denotes the probability density function for an individual observation x. The remaining $s-r$ parameters of ω are represented by another vector $\psi = (\omega_{r+1},\ldots,\omega_s)$, which cannot be learned from the observed data. Thus, $\omega = (\phi,\psi)$ is partially identified with the identified part ϕ and the unidentified part ψ.

Further, we consider additional assumption that impose t equality constraints on ω:

$$\mathbf{h}(\omega) = \begin{pmatrix} h_1(\omega) \\ \vdots \\ h_t(\omega) \end{pmatrix} = \mathbf{0}.$$

These equality constraints can be used to identify ψ. Since the dimension of ψ is $s-r$, we assume that there are at least $s-r$ equations so that ψ can be fully identified. Also, it is reasonable to assume that the number of constraints does not exceed the number of identified parameters, which is necessary for the development of our method. Thus, we assume that $s-r \le t \le r$. Note that the true, though unknown, parameter value $\omega^* = (\omega_1^*,\ldots,\omega_s^*)$ is presumed to satisfy these constraints itself, i.e., $\mathbf{h}(\omega^*) = \mathbf{0}$.

Our objective is to find the constrained maximum likelihood estimate $\hat{\omega}$ that maximizes the log-likelihood function $\ell(\mathbf{x},\phi)$ subject to the condition $\mathbf{h}(\omega) = \mathbf{0}$, and study the properties of the estimator. Moreover, we will propose a numerical algorithm for computing the constrained maximum likelihood estimate in practice.

2.3 The constrained maximum likelihood estimation

Let $\hat{\phi}^{(u)}$ denote the unconstrained maximum likelihood estimate under general conditional without additional assumptions. If the equation $\mathbf{h}(\hat{\phi}^{(u)},\psi) = \mathbf{0}$ with respect to ψ has a solution, say $\hat{\psi}^{(c)}$, then $\hat{\omega} = (\hat{\phi}^{(u)},\hat{\psi}^{(c)})$ is the constrained maximum likelihood estimate of the problem. This approach may fail, however, since the equation $\mathbf{h}(\phi,\psi) = \mathbf{0}$ with respect to ψ may not necessarily have a solution for some values of ϕ. Alternatively, we propose to estimate ω by finding the stationary point of $(1/n)\ell(\mathbf{x},\phi) + \lambda^T\mathbf{h}(\omega)$, where $\lambda = (\lambda_1,\ldots,\lambda_t)$ is a Lagrange multiplier.

Thus, we consider the following $s+t$ equations:

$$\frac{1}{n}\mathbf{s}(\mathbf{x},\phi)+\mathbf{J}_\omega\lambda=\mathbf{0}, \tag{2.1}$$

$$\mathbf{K}_\omega\lambda=\mathbf{0}, \tag{2.2}$$

$$\mathbf{h}(\omega)=\mathbf{0}, \tag{2.3}$$

where $\mathbf{s}(\mathbf{x},\phi)$ is the score vector of length r whose i-th component is $\partial\ell(\mathbf{x},\phi)/\partial\omega_i$, for $i=1,\ldots,r$, $\mathbf{J}_\omega$ is the $r\times t$ matrix $(\partial h_j(\omega)/\partial\omega_i)$, for $i=1,\ldots,r$, $j=1,\ldots,t$, and $\mathbf{K}_\omega$ is the $(s-r)\times t$ matrix $(\partial h_j(\omega)/\partial\omega_{r+i})$, for $i=1,\ldots,s-r$, $j=1,\ldots,t$.

In this section, we will show that, under some general conditions, if $\mathbf{x}$ belongs to a set whose probability measure tends to 1 as n approaches infinity, then the equations (2.1) - (2.3) have a solution $(\hat{\omega},\hat{\lambda})$ such that $\hat{\omega}$ is within a small neighborhood of the true value ω^*. This solution is then proved to be the constrained maximum likelihood estimate that maximizes $\ell(\mathbf{x},\phi)$ subject to $\mathbf{h}(\omega)=\mathbf{0}$. We then extend the definition of $(\hat{\omega},\hat{\lambda})$ for all $\mathbf{x}\in\mathbb{R}^n$, and show the asymptotic distribution of the random variable thus defined. Finally, we propose an algorithm for numerically computing $(\hat{\omega},\hat{\lambda})$. The development of this section is based on the work by Aitchison and Silvey [1]. However, due to the presence of the unidentified component ψ, our work is more than a simple generalization of their results.

We first impose some conditions on $f(x,\phi)$ and $\mathbf{h}(\omega)$ within some neighborhood of ω^*, say $U_\alpha=\{\omega:||\omega-\omega^*||\le\alpha\}$. We assume that $f(x,\phi)$ satisfies the conditions ($\mathscr{F}1$) - ($\mathscr{F}4$) as defined in [1]. These conditions are quite general and will be satisfied in most practical estimation problems. Here, we just write one important result implied by these conditions for later reference. If the conditions on $f(x,\phi)$ are satisfied, for any given positive numbers $\delta<\alpha$ and $\varepsilon<1$ and for sufficiently large $n\ge n(\delta,\varepsilon)$, there exists a set $\mathbf{X}_n$ with the properties

($\mathscr{X}1$) $\Pr\{\mathbf{X}_n\}>1-\varepsilon$.

($\mathscr{X}2$) $||\mathbf{s}(\mathbf{x},\phi^*)/n||<\delta^2$, if $\mathbf{x}\in\mathbf{X}_n$.

($\mathscr{X}3$) $\left(\mathbf{M}_{\mathbf{x},\phi^*}/n\right)$ can be expressed in the form $-\mathbf{B}_{\phi^*}+\delta\mathbf{m}_{\mathbf{x},\phi^*}$, where $\mathbf{M}_{\mathbf{x},\phi^*}$ is the matrix $(\partial^2\ell(\mathbf{x},\phi^*)/\partial\omega_i\partial\omega_j)$, $i,j=1,\ldots,r$, $\mathbf{B}_{\phi^*}$ is a certain positive definite matrix, and $\mathbf{m}_{\mathbf{x},\phi^*}$ is an $r\times r$ matrix, the moduli of whose elements are

bounded by 1, if $\mathbf{x} \in \mathbf{X}_n$.

($\mathscr{X}4$) There exists a constant, say κ_1, such that for every $\omega \in U_\alpha$ and $i, j, k = 1, 2, \ldots, r$,

$$\left| \frac{1}{n} \frac{\partial^3 \ell(\mathbf{x}, \phi)}{\partial \omega_i \partial \omega_j \partial \omega_k} \right| < 2\kappa_1,$$

if $\mathbf{x} \in \mathbf{X}_n$.

On the other hand, some conditions are assumed for the constraint function $\mathbf{h}(\omega)$ as follows.

($\mathscr{H}1$) For every $\omega \in U_\alpha$, the first order partial derivatives $\partial h_k(\omega)/\partial \omega_i$, $i = 1, \ldots, s$, $k = 1, \ldots, t$, exist and they are continuous function of ω.

($\mathscr{H}2$) For every $\omega \in U_\alpha$, the second order partial derivatives $\partial^2 h_k(\omega)/\partial \omega_i \partial \omega_j$, $i, j = 1, \ldots, s$, $k = 1, \ldots, t$, exist, and they are uniformly bounded by a constant, say κ_2, on U_α.

($\mathscr{H}3$) The $r \times t$ matrix $\mathbf{J}_{\omega^*}$ and the $(s-r) \times t$ matrix $\mathbf{K}_{\omega^*}$ are both of full rank, i.e., $rank(\mathbf{J}_{\omega^*}) = t$ and $rank(\mathbf{K}_{\omega^*}) = s - r$.

2.3.1 The constrained maximum likelihood estimate

We begin by establishing a necessary and sufficient condition for the existence of a solution to the equations (2.1) - (2.3) under some general conditions. It should be noted that the following lemma cannot be directly generalized from Lemma 1 in [1] by simply viewing the log-likelihood function as a function of ω and letting $\mathbf{B}_{\omega^*}$ be the $s \times s$ matrix that naturally extends $\mathbf{B}_{\phi^*}$, due to the singularity of $\mathbf{B}_{\omega^*}$ thus defined. Therefore, some modifications are required.

Lemma 1. *Suppose conditions on f and $\mathbf{h}$ are satisfied, and $\mathbf{X}_n$ is a set with the properties ($\mathscr{X}1$) - ($\mathscr{X}3$) for some given positive numbers $\delta < \alpha$ and $\varepsilon < 1$. When $\mathbf{x} \in \mathbf{X}_n$ and α is sufficiently small, then $(\hat{\omega}, \hat{\lambda})$ is a solution of the equations (2.1) - (2.3) and $\hat{\omega} \in U_\delta$, if and only if $\hat{\omega}$ satisfies a certain equation. This equation takes the form $-\tilde{\mathbf{B}}_{\omega^*}(\omega - \omega^*) + \delta^2 \mathbf{v}(\mathbf{x}, \omega) = \mathbf{0}$, where*

$$\tilde{\mathbf{B}}_{\omega^*} = \begin{pmatrix} \mathbf{B}_{\phi^*} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{s-r} \end{pmatrix},$$

and $\mathbf{v}(\mathbf{x},\omega)$ *is a continuous function of* ω *on* U_δ *and* $||\mathbf{v}(\mathbf{x},\omega)||$ *is bounded on* U_δ *by a positive number* $\kappa^\dagger$*, which does not depend on* δ*.*

Proof. We first prove the necessity of the condition. By expanding the components of $\mathbf{s}(\mathbf{x},\phi)$ at ϕ^* in the equation (2.1), and the components of $\mathbf{h}(\omega)$ at ω^* in the equation (2.3), we find that the solution of the equations (2.1) - (2.3) should also satisfy:

$$\frac{1}{n}\left\{\mathbf{s}(\mathbf{x},\phi^*)+\mathbf{M}_{\mathbf{x},\phi^*}(\phi-\phi^*)+\mathbf{v}^{(1)}(\mathbf{x},\phi)\right\}+\mathbf{J}_\omega\lambda=\mathbf{0}, \tag{2.4}$$

$$\mathbf{J}^T_{\omega^*}(\phi-\phi^*)+\mathbf{K}^T_{\omega^*}(\psi-\psi^*)+\mathbf{v}^{(2)}(\omega)=\mathbf{0}, \tag{2.5}$$

where

(i) $\mathbf{v}^{(1)}(\mathbf{x},\phi)$ is a vector of dimension r whose m-th component is

$$\frac{1}{2}(\phi-\phi^*)^T\mathbf{L}_m(\phi-\phi^*),$$

where $\mathbf{L}_m$ is the matrix $(\partial^3\ell(\mathbf{x},\phi^{(m,1)})/\partial\omega_m\partial\omega_i\partial\omega_j)$, $i,j=1,\ldots,r$, with $\phi^{(m,1)}$ being a point such that $||\phi^{(m,1)}-\phi^*||<||\phi-\phi^*||$, and

(ii) $\mathbf{v}^{(2)}(\omega)$ is a vector of dimension s whose m-th component is

$$\frac{1}{2}(\omega-\omega^*)^T\mathbf{H}_m(\omega-\omega^*),$$

where $\mathbf{H}_m$ is the matrix $(\partial^2 h_m(\omega^{(m,2)})/\partial\omega_i\partial\omega_j)$, $i,j=1,\ldots,s$, with $\omega^{(m,2)}$ being a point such that $||\omega^{(m,2)}-\omega^*||<||\omega-\omega^*||$.

Further, given property ($\mathscr{X}3$) , we can re-write the equations (2.4) and (2.5) in the following form:

$$-\mathbf{B}_{\phi^*}(\phi-\phi^*)+\mathbf{J}_\omega\lambda+\delta^2\mathbf{v}^{(3)}(\mathbf{x},\phi)=\mathbf{0}, \tag{2.6}$$

$$\mathbf{J}^T_{\omega^*}(\phi-\phi^*)+\mathbf{K}^T_{\omega^*}(\psi-\psi^*)+\delta^2\mathbf{v}^{(4)}(\omega)=\mathbf{0}, \tag{2.7}$$

where

$$\mathbf{v}^{(3)}(\mathbf{x},\phi) = \frac{1}{n\delta^2}\mathbf{s}(\mathbf{x},\phi^*) + \frac{1}{\delta}\mathbf{m}_{\mathbf{x},\phi^*}(\phi-\phi^*) + \frac{1}{n\delta^2}\mathbf{v}^{(1)}(\mathbf{x},\phi), \tag{2.8}$$

$$\mathbf{v}^{(4)}(\omega) = \frac{1}{\delta^2}\mathbf{v}^{(2)}(\omega). \tag{2.9}$$

Moreover, by properties ($\mathscr{X}2$) - ($\mathscr{X}4$), we obtain a bound for $\mathbf{v}^{(3)}(\mathbf{x},\phi)$ as

$$\begin{aligned} ||\mathbf{v}^{(3)}(\mathbf{x},\phi)|| &\leq \frac{1}{n\delta^2}||\mathbf{s}(\mathbf{x},\phi^*)|| + \frac{1}{\delta}||\mathbf{m}_{\mathbf{x},\phi^*}(\phi-\phi^*)|| + \frac{1}{n\delta^2}||\mathbf{v}^{(1)}(\mathbf{x},\phi)|| \\ &< 1 + r^2 + r^3\kappa_1, \end{aligned} \tag{2.10}$$

and, by condition ($\mathscr{H}2$), we have a bound for $\mathbf{v}^{(4)}(\omega)$ as

$$\begin{aligned} ||\mathbf{v}^{(4)}(\omega)|| &< s^3\kappa_2\left(\frac{1}{\delta^2}||\omega-\omega^*||^2\right) \\ &< s^3\kappa_2. \end{aligned} \tag{2.11}$$

Next, since $\mathbf{B}_{\phi^*}$ is positive definite, we can pre-multiply the equation (2.6) by $\mathbf{J}_{\omega^*}^T\mathbf{B}_{\phi^*}^{-1}$ to get an expression for $\mathbf{J}_{\omega^*}^T(\phi-\phi^*)$, which is then plugged into the equation (2.7) to obtain the following equation

$$\mathbf{J}_{\omega^*}^T\mathbf{B}_{\phi^*}^{-1}\mathbf{J}_\omega\lambda + \mathbf{K}_{\omega^*}^T(\psi-\psi^*) + \delta^2\left(\mathbf{J}_{\omega^*}^T\mathbf{B}_{\phi^*}^{-1}\mathbf{v}^{(3)}(\mathbf{x},\phi) + \mathbf{v}^{(4)}(\omega)\right) = \mathbf{0}. \tag{2.12}$$

Now the condition ($\mathscr{H}3$) implies that $\mathbf{J}_{\omega^*}^T\mathbf{B}_{\phi^*}^{-1}\mathbf{J}_{\omega^*}$ is also positive definite. Besides, according to the condition ($\mathscr{H}1$), the elements of $\mathbf{J}_\omega$ are all continuous functions of ω. It then follows that $\mathbf{J}_{\omega^*}^T\mathbf{B}_{\phi^*}^{-1}\mathbf{J}_\omega$ is also non-singular within U_α so long as α is sufficiently small. Thus, we can solve the equation (2.12) with respect to λ and express it in terms of ω

$$\lambda = -\mathbf{A}_\omega\left\{\mathbf{K}_{\omega^*}^T(\psi-\psi^*) + \delta^2\left(\mathbf{J}_{\omega^*}^T\mathbf{B}_{\phi^*}^{-1}\mathbf{v}^{(3)}(\mathbf{x},\phi) + \mathbf{v}^{(4)}(\omega)\right)\right\}, \tag{2.13}$$

where we define the notation $\mathbf{A}_\omega = (\mathbf{J}_{\omega^*}^T\mathbf{B}_{\phi^*}^{-1}\mathbf{J}_\omega)^{-1}$.

So far, we are basically replicating the steps of the proof given by [1]. Now, we need to take some extra steps to find the expression for $(\psi-\psi^*)$. By applying

the equation (2.13) to substitute for λ, the equation (2.2) becomes:

$$\mathbf{K}_\omega \mathbf{A}_\omega \mathbf{K}_{\omega^*}^T (\psi - \psi^*) + \delta^2 \mathbf{K}_\omega \mathbf{A}_\omega \left(\mathbf{J}_{\omega^*}^T \mathbf{B}_{\phi^*}^{-1} \mathbf{v}^{(3)}(\mathbf{x}, \phi) + \mathbf{v}^{(4)}(\omega) \right) = \mathbf{0}. \quad (2.14)$$

Following the same argument for $\mathbf{J}_{\omega^*}^T \mathbf{B}_{\phi^*}^{-1} \mathbf{J}_\omega$, the condition ($\mathscr{H}4$) ensures that the matrix $\mathbf{K}_\omega \mathbf{A}_\omega \mathbf{K}_{\omega^*}^T$ is not singular within U_α provided that α is sufficiently small. Thus, we can solve the equation (2.14) with respect to ψ and get

$$\psi - \psi^* = -\delta^2 \mathbf{v}^{(5)}(\mathbf{x}, \omega), \quad (2.15)$$

where

$$\mathbf{v}^{(5)}(\mathbf{x}, \omega) = \left(\mathbf{K}_\omega \mathbf{A}_\omega \mathbf{K}_{\omega^*}^T \right)^{-1} (\mathbf{K}_\omega \mathbf{A}_\omega) \left(\mathbf{J}_{\omega^*}^T \mathbf{B}_{\phi^*}^{-1} \mathbf{v}^{(3)}(\mathbf{x}, \phi) + \mathbf{v}^{(4)}(\omega) \right). \quad (2.16)$$

We then plug the equation (2.15) into the equation (2.13) and derive an updated expression for λ:

$$\lambda = -\delta^2 \mathbf{v}^{(6)}(\mathbf{x}, \omega), \quad (2.17)$$

where

$$\mathbf{v}^{(6)}(\mathbf{x}, \omega) = \mathbf{A}_\omega \left\{ -\mathbf{K}_{\omega^*}^T \mathbf{v}^{(5)}(\mathbf{x}, \omega) + \left(\mathbf{J}_{\omega^*}^T \mathbf{B}_{\phi^*}^{-1} \mathbf{v}^{(3)}(\mathbf{x}, \phi) + \mathbf{v}^{(4)}(\omega) \right) \right\}. \quad (2.18)$$

By combining the equations (2.6) and (2.15), with λ substituted using the equation (2.17), we find that the solution of the equations (2.1) - (2.3) should also satisfy

$$-\tilde{\mathbf{B}}_{\omega^*}(\omega - \omega^*) + \delta^2 \mathbf{v}(\mathbf{x}, \omega) = \mathbf{0}, \quad (2.19)$$

where

$$\tilde{\mathbf{B}}_{\omega^*} = \begin{pmatrix} \mathbf{B}_{\phi^*} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{s-r} \end{pmatrix},$$

and

$$\mathbf{v}(\mathbf{x}, \omega) = \begin{pmatrix} \mathbf{v}^{(3)}(\mathbf{x}, \phi) - \mathbf{J}_\omega \mathbf{v}^{(6)}(\mathbf{x}, \omega) \\ -\mathbf{v}^{(5)}(\mathbf{x}, \omega) \end{pmatrix}.$$

Finally, we have shown in the inequalities (2.10) and (2.11) that $||\mathbf{v}^{(3)}(\mathbf{x}, \phi)||$

and $||\mathbf{v}^{(4)}(\omega)||$ are uniformly bounded within U_α. Also, given that $\mathbf{A}_\omega$ and $\mathbf{K}_\omega \mathbf{A}_\omega \mathbf{K}_{\omega^*}^T$ are positive definite within the closed set U_α, their determinants are both positive within U_α. Therefore, the continuity of the elements of these two matrices ensures that their determinants are uniformly bounded within U_α. Then it follows that $\mathbf{v}(\mathbf{x},\omega)$ is a continuous function on U_δ and $||\mathbf{v}(\mathbf{x},\omega)||$ is bounded on U_δ by a positive number, say $\kappa^\dagger$, which does not depend on δ.

Now, we prove the sufficiency of the condition. Suppose the equation (2.19) has a solution $\hat{\omega}$. That is, $\hat{\omega}$ satisfies

$$\begin{pmatrix} \mathbf{B}_{\phi^*} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{s-r} \end{pmatrix} \begin{pmatrix} \hat{\phi} - \phi^* \\ \hat{\psi} - \psi^* \end{pmatrix} = \delta^2 \begin{pmatrix} \mathbf{v}^{(3)}(\mathbf{x},\hat{\phi}) - \mathbf{J}_{\hat{\omega}} \mathbf{v}^{(6)}(\mathbf{x},\hat{\omega}) \\ -\mathbf{v}^{(5)}(\mathbf{x},\hat{\omega}) \end{pmatrix}. \quad (2.20)$$

By pre-multiplying the equation (2.20) by the $t \times s$ matrix $(\mathbf{J}_{\omega^*}^T \mathbf{B}_{\phi^*}^{-1},\ \mathbf{K}_{\omega^*}^T)$, we have

$$\mathbf{J}_{\omega^*}^T(\hat{\phi} - \phi^*) + \mathbf{K}_{\omega^*}^T(\hat{\psi} - \psi^*) + \delta^2 \mathbf{v}^{(4)}(\hat{\omega}) = \mathbf{0}. \quad (2.21)$$

We first write $\mathbf{v}^{(1)}(\mathbf{x},\phi)$ and $\mathbf{v}^{(2)}(\omega)$ as the remainders after expanding $\mathbf{s}(\mathbf{x},\phi)$ and $\mathbf{h}(\omega)$, respectively,

$$\mathbf{v}^{(1)}(\mathbf{x},\phi) = \mathbf{s}(\mathbf{x},\phi) - \mathbf{s}(\mathbf{x},\phi^*) - \mathbf{M}_{\mathbf{x},\phi^*}(\phi - \phi^*), \quad (2.22)$$

$$\mathbf{v}^{(2)}(\omega) = \mathbf{h}(\omega) - \mathbf{J}_{\omega^*}^T(\phi - \phi^*) - \mathbf{K}_{\omega^*}^T(\psi - \psi^*). \quad (2.23)$$

Applying the equations (2.22) and (2.23) to substitute for $\mathbf{v}^{(1)}(\mathbf{x},\phi)$ and $\mathbf{v}^{(2)}(\omega)$ in the equations (2.8) and (2.9), respectively, we get

$$\mathbf{v}^{(3)}(\mathbf{x},\phi) = \frac{1}{\delta^2}\left\{\frac{1}{n}\mathbf{s}(\mathbf{x},\phi) + \mathbf{B}_{\phi^*}(\phi - \phi^*)\right\}, \quad (2.24)$$

$$\mathbf{v}^{(4)}(\omega) = \frac{1}{\delta^2}\left\{\mathbf{h}(\omega) - \mathbf{J}_{\omega^*}^T(\phi - \phi^*) - \mathbf{K}_{\omega^*}^T(\psi - \psi^*)\right\}. \quad (2.25)$$

Finally, we substitute for $\mathbf{v}^{(4)}(\hat{\omega})$ in the equation (2.21) using the equation (2.25). It immediately follows that $\mathbf{h}(\hat{\omega}) = \mathbf{0}$.

Next, we apply the equations (2.24) and (2.25) to substitute for $\mathbf{v}^{(3)}(\mathbf{x},\phi)$ and $\mathbf{v}^{(4)}(\omega)$ in the equations (2.16) and (2.18), and end with the following expressions

for $\mathbf{v}^{(5)}(\mathbf{x},\omega)$ and $\mathbf{v}^{(6)}(\mathbf{x},\omega)$:

$$\mathbf{v}^{(5)}(\mathbf{x},\omega) = -(\psi-\psi^*) + \left(\mathbf{K}_{\hat{\omega}}\mathbf{A}_{\omega}\mathbf{K}_{\omega^*}^T\right)^{-1}\mathbf{K}_{\omega}\mathbf{Y}_{\omega}\left(\frac{1}{n}\mathbf{s}(\mathbf{x},\phi)\right), \tag{2.26}$$

$$\mathbf{v}^{(6)}(\mathbf{x},\omega) = \mathbf{Y}_{\omega}\left(\frac{1}{n}\mathbf{s}(\mathbf{x},\phi)\right) - \mathbf{K}_{\omega^*}^T\left(\mathbf{K}_{\omega}\mathbf{A}_{\omega}\mathbf{K}_{\omega^*}^T\right)^{-1}\mathbf{K}_{\omega}\mathbf{Y}_{\omega}\left(\frac{1}{n}\mathbf{s}(\mathbf{x},\phi)\right), \tag{2.27}$$

where $\mathbf{Y}_{\omega}$ is defined as $\mathbf{Y}_{\omega} = \mathbf{A}_{\omega}\mathbf{J}_{\omega^*}^T\mathbf{B}_{\phi^*}^{-1}$. Now, by using the equations (2.24), (2.26) and (2.27) to substitue for $\mathbf{v}^{(3)}(\mathbf{x},\hat{\phi})$, $\mathbf{v}^{(5)}(\mathbf{x},\hat{\omega})$ and $\mathbf{v}^{(6)}(\mathbf{x},\hat{\omega})$ in the equation (2.20), respectively, we can see that $\hat{\omega}$ satisfies

$$\frac{1}{n}\mathbf{s}(\mathbf{x},\hat{\phi}) - \mathbf{J}_{\hat{\omega}}\mathbf{Y}_{\hat{\omega}}\left(\frac{1}{n}\mathbf{s}(\mathbf{x},\hat{\phi})\right) = \mathbf{0},$$

$$-\mathbf{K}_{\hat{\omega}}\mathbf{Y}_{\hat{\omega}}\left(\frac{1}{n}\mathbf{s}(\mathbf{x},\hat{\phi})\right) = \mathbf{0}.$$

As we have shown earlier that $\mathbf{h}(\hat{\omega}) = \mathbf{0}$, it is easy to see that $\hat{\omega}$, jointly with $\hat{\lambda} = -\mathbf{Y}_{\hat{\omega}}\mathbf{s}(\mathbf{x},\hat{\phi})/n$, solves the equations (2.1) - (2.3). □

We now give the following theorem to show the existence of a solution of the equations (2.1) - (2.3).

Theorem 1. *Subject to conditions on f and $\mathbf{h}$, if $\mathbf{x} \in \mathbf{X}_n$ for a sufficiently small given positive number δ and another given positive number $\varepsilon < 1$, then the equations (2.1) - (2.3) have a solution $(\hat{\omega},\hat{\lambda})$ such that $\hat{\omega} \in U_{\delta}$.*

Proof. The proof of Theorem 1 in [1] works here, provided the modified version of Lemma 1 given above is used. Also, it is important to notice that the matrix $\tilde{\mathbf{B}}_{\omega^*}$ defined in Lemma 1 is positive definite provided that $\mathbf{B}_{\phi^*}$ is positive definite, and its minimum latent root is $\min\{\mu_0, 1\}$, where μ_0 is the latent minimum root of $\mathbf{B}_{\phi^*}$. Details are omitted. □

In the remainder of this section, we are going to show that the solution of the equations (2.1) - (2.3) as stated in Theorem 1 locally maximizes the log-likelihood subject to the constraints. This result was proved in [1] for the identified model. However, we are not able to prove this result for the partially identified model with

a direct extension of their proof. Alternatively, we take another route and use the result given by Spring [27].

To match with the set-up in [27], we change the order of variables and let $\eta = (\lambda, \omega)$. Let $\mathbf{HT}$ denote the second order partial derivatives of the Lagrangian function $\ell(\mathbf{x}, \phi)/n + \lambda^T \mathbf{h}(\omega)$ evaluated at the critical point $\hat{\eta} = (\hat{\lambda}, \hat{\omega})$

$$\mathbf{HT}^{(n)} = \begin{pmatrix} \mathbf{0} & \mathbf{J}_{\hat{\omega}}^T & \mathbf{K}_{\hat{\omega}}^T \\ \mathbf{J}_{\hat{\omega}} & \frac{1}{n}\mathbf{M}_{\hat{\phi}} + \mathbf{X}_{\hat{\lambda},\hat{\omega}} & \mathbf{Y}_{\hat{\lambda},\hat{\omega}}^T \\ \mathbf{K}_{\hat{\omega}} & \mathbf{Y}_{\hat{\lambda},\hat{\omega}} & \mathbf{Z}_{\hat{\lambda},\hat{\omega}} \end{pmatrix},$$

where

$$\begin{pmatrix} \mathbf{X}_{\hat{\lambda},\hat{\omega}} & \mathbf{Y}_{\hat{\lambda},\hat{\omega}}^T \\ \mathbf{Y}_{\hat{\lambda},\hat{\omega}} & \mathbf{Z}_{\hat{\lambda},\hat{\omega}} \end{pmatrix} = \sum_{k=1}^{t} \hat{\lambda}_k \begin{pmatrix} \frac{\partial^2 h_k}{\partial \omega_1 \partial \omega_1} & \cdots & \frac{\partial^2 h_k}{\partial \omega_1 \partial \omega_s} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 h_k}{\partial \omega_s \partial \omega_1} & \cdots & \frac{\partial^2 h_k}{\partial \omega_1 s \partial \omega_s} \end{pmatrix},$$

with $\mathbf{X}_{\hat{\lambda},\hat{\omega}}$ being the upper-left $r \times r$ block matrix, $\mathbf{Y}_{\hat{\lambda},\hat{\omega}}$ being the bottom-left $r \times (s-r)$ block matrix, and $\mathbf{Z}_{\hat{\lambda},\hat{\omega}}$ being the bottom-right $(s-r) \times (s-r)$ block matrix. Let $\Lambda_k^{(n)}$ denote the principal upper left k-th order minor of the Hessian Matrix $\mathbf{HT}^{(n)}$. According to Theorem 1 of [27], $\hat{\omega}$ locally maximizes the log-likelihood function subject to the constraints, so long as $(-1)^{t+p}\Lambda_{2t+p}^{(n)}$, $p = 1, \ldots, s-t$, are all positive.

Note that $\hat{\lambda}$ was defined as $\hat{\lambda} = -\mathbf{Y}_{\hat{\omega}} \left(\mathbf{s}(\mathbf{x}, \hat{\phi})/n \right)$. For any small number δ, by the equation (2.24) and the inequality (2.10), if n is sufficiently large, we have

$$\begin{aligned} \frac{1}{n}||\mathbf{s}(\mathbf{x}, \hat{\phi})|| &= ||-\mathbf{B}_{\phi^*}(\hat{\phi} - \phi^*) + \delta^2 \mathbf{v}^{(3)}(\mathbf{x}, \hat{\phi})|| \\ &< \kappa_3 \delta + (1 + r^2 + r^3 \kappa_1)\delta^2, \end{aligned}$$

where κ_3 is a positive number that depends only on the elements of $\mathbf{B}_{\phi^*}$. Also, the elements of $\mathbf{Y}_{\hat{\omega}}$ are bounded by a number independent of δ for $\hat{\omega} \in U_\delta$. Therefore,

we have

$$||\hat{\lambda}|| = \frac{1}{n}||\mathbf{Y}_{\hat{\omega}}\mathbf{s}(\mathbf{x},\hat{\phi})||$$
$$< \kappa_4\delta + \kappa_5\delta^2,$$

where κ_4 and κ_5 are positive numbers independent of δ. That is, $\hat{\lambda}$ converges to $\mathbf{0}$ as n goes to infinity ($\delta \to 0$). By condition ($\mathscr{H}2$), the second partial derivatives $\partial^2 h_k(\omega)/\partial\omega_i\partial\omega_j$, $i,j = 1,\ldots,s$, $k = 1,\ldots,k$, are all bounded by a constant $2\kappa_2$. Thus, it follows that $\mathbf{X}_{\hat{\lambda},\hat{\omega}} \to \mathbf{0}_{r\times r}$, $\mathbf{Y}_{\hat{\lambda},\hat{\omega}} \to \mathbf{0}_{r\times(s-r)}$, and $\mathbf{Z}_{\hat{\lambda},\hat{\omega}} \to \mathbf{0}_{(s-r)\times(s-r)}$. Also, it is easy to see from Theorem 1 that, for $\hat{\omega} \in U_\delta$ with sufficiently small value of δ, $\hat{\omega}$ converges to ω^* as n goes to infinity. By condition ($\mathscr{H}1$), the elements of $\mathbf{J}_\omega$ and $\mathbf{K}_\omega$ are all continuous functions of ω. Thus, as n goes to infinity, $\mathbf{J}_{\hat{\omega}}$, $\mathbf{K}_{\hat{\omega}}$, and $\mathbf{M}_{\hat{\phi}}/n$ approach $\mathbf{J}_{\omega^*}$, $\mathbf{K}_{\omega^*}$ and $\mathbf{M}_{\phi^*}/n$, respectively. Furthermore, by property ($\mathscr{X}2$), we have $\mathbf{M}_{\phi^*}/n$ approaches $-\mathbf{B}_{\phi^*}$ as n goes to infinity. Finally, we have $\mathbf{HT}^{(n)}$ converges to $\mathbf{HT}^{(\infty)}$ as n goes to infinity, where

$$\mathbf{HT}^{(\infty)} = \begin{pmatrix} \mathbf{0} & \mathbf{J}_{\omega^*}^T & \mathbf{K}_{\omega^*}^T \\ \mathbf{J}_{\omega^*} & -\mathbf{B}_{\phi^*} & \mathbf{0} \\ \mathbf{K}_{\omega^*} & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Then, for sufficiently large n, the signs of the leading principal minors of $\mathbf{HT}^{(n)}$ are the same as those of their corresponding minors of $\mathbf{HT}^{(\infty)}$. Therefore, we can instead study the signs of the leading principal minors of $\mathbf{HT}^{(\infty)}$.

For brevity, we suppress the subscripts ω^* and ϕ^*. First, given that $\mathbf{B}$ is positive definite, by Sylvester's criterion the upper left $d \times d$ corner matrix of $\mathbf{B}$, denoted by $\mathbf{B}_d$, is also positive definite, for $d = 1,\ldots,r$. Next, since $rank(\mathbf{J}) = t$, with some re-ordering of the rows if necessary, the first d rows of $\mathbf{J}$, denoted by $\mathbf{J}_d$, is a $d \times t$ matrix of full column rank t, and thus the matrix $\mathbf{J}_d^T\mathbf{B}_d^{-1}\mathbf{J}_d$ is positive definite, for $d = t+1,\ldots,r$. Similarly, as $rank(\mathbf{K}) = s-r$, the first d rows of $\mathbf{K}$, denoted by $\mathbf{K}_d$, is a $d \times t$ matrix of full row rank d, and thus the matrix $\mathbf{K}_d\left(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J}\right)^{-1}\mathbf{K}_d^T$ is again positive definite, for $d = 1,\ldots,s-r$. Now we are ready to study the sign of

$(-1)^{t+p}\Lambda^{(\infty)}_{2t+p}$, for $p = 1,\ldots,s-t$. On one hand, for $p = 1,\ldots,r-t$, we have

$$\begin{aligned}
(-1)^{t+p}\Lambda^{(\infty)}_{2t+p} &= (-1)^{t+p} \times \det\left(\begin{pmatrix} \mathbf{0} & \mathbf{J}^T_{t+p} \\ \mathbf{J}_{t+p} & -\mathbf{B}_{t+p} \end{pmatrix}\right) \\
&= (-1)^{t+p} \times \det(-\mathbf{B}_{t+p}) \times \det\left(-\mathbf{J}^T_{t+p}(-\mathbf{B}_{t+p})^{-1}\mathbf{J}_{t+p}\right) \\
&= (-1)^{2t+2p} \times \det(\mathbf{B}_{t+p}) \times \det\left(\mathbf{J}^T_{t+p}\mathbf{B}^{-1}_{t+p}\mathbf{J}_{t+p}\right) \\
&> 0.
\end{aligned}$$

On the other hand, for $p = r-t+1,\ldots,s-t$, we have

$$\begin{aligned}
(-1)^{t+p}\Lambda^{(\infty)}_{2t+p} &= (-1)^{t+p} \times \det\left(\begin{pmatrix} \mathbf{0} & \mathbf{J}^T & \mathbf{K}^T_{t+p-r} \\ \mathbf{J} & -\mathbf{B} & \mathbf{0} \\ \mathbf{K}_{t+p-r} & \mathbf{0} & \mathbf{0} \end{pmatrix}\right) \\
&= (-1)^{t+p} \times \det\left(\begin{pmatrix} \mathbf{0} & \mathbf{J}^T \\ \mathbf{J} & -\mathbf{B} \end{pmatrix}\right) \times \\
&\quad \det\left(-\begin{pmatrix} \mathbf{K}^T_{t+p-r} \\ \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{0} & \mathbf{J}^T \\ \mathbf{J} & -\mathbf{B} \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{K}_{t+p-r} & \mathbf{0} \end{pmatrix}\right) \\
&= (-1)^{t+p} \times \det(-\mathbf{B}) \times \det\left(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J}\right) \times \\
&\quad \det\left(-\mathbf{K}_{t+p-r}\left(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J}\right)^{-1}\mathbf{K}^T_{t+p-r}\right) \\
&= (-1)^{2t+2p} \times \det(\mathbf{B}) \times \det\left(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J}\right) \times \\
&\quad \det\left(\mathbf{K}_{t+p-r}\left(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J}\right)^{-1}\mathbf{K}^T_{t+p-r}\right) \\
&> 0.
\end{aligned}$$

Therefore, we have shown that $(-1)^{t+p}\Lambda^{(\infty)}_{2t+p}$, $p = 1,\ldots,s-t$, is always positive, and so is $(-1)^{t+p}\Lambda^{(n)}_{2t+p}$ for sufficiently large n. Thus, it follows that $\hat{\boldsymbol{\omega}}$ is the constrained maximum likelihood estimate of the problem.

2.3.2 Asymptotic distributions

In this section, we define sequences $\left\{(\hat{\omega}_n, \hat{\lambda}_n)\right\}$ that extends $(\hat{\omega}, \hat{\lambda})$, as stated in the Theorem 1, for all $\mathbf{x} \in \mathbb{R}^n$, and develop the asymptotic distribution for $(\hat{\omega}_n, \hat{\lambda}_n)$. Note that this section differs from the Section 5 of [1] in that the covariance matrix here becomes a partitioned matrix of 3×3 blocks.

Lemma 2. *The following partitioned matrix is non-singular.*

$$\begin{pmatrix} \mathbf{B}_{\phi^*} & \mathbf{0} & -\mathbf{J}_{\omega^*} \\ \mathbf{0} & \mathbf{0} & -\mathbf{K}_{\omega^*} \\ -\mathbf{J}_{\omega^*}^T & -\mathbf{K}_{\omega^*}^T & \mathbf{0} \end{pmatrix}$$

Proof. For brevity, we omit the suffix ϕ^* and ω^*. Then we wish to find a matrix

$$\begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{P}_{13} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \mathbf{P}_{23} \\ \mathbf{P}_{31} & \mathbf{P}_{32} & \mathbf{P}_{33} \end{pmatrix}$$

such that

$$\begin{pmatrix} \mathbf{B} & \mathbf{0} & -\mathbf{J} \\ \mathbf{0} & \mathbf{0} & -\mathbf{K} \\ -\mathbf{J}^T & -\mathbf{K}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{P}_{13} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \mathbf{P}_{23} \\ \mathbf{P}_{31} & \mathbf{P}_{32} & \mathbf{P}_{33} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{s-r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_t \end{pmatrix}.$$

Since $\mathbf{B}$ is positive definite, and $\mathbf{J}$ and $\mathbf{K}$ are of full rank, it can be solved that

$$\begin{aligned}
\mathbf{P}_{11} &= \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{J}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{J}^T\mathbf{B}^{-1} + \\
&\qquad \mathbf{B}^{-1}\mathbf{J}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{K}^T\left\{\mathbf{K}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{K}^T\right\}^{-1}\mathbf{K}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{J}^T\mathbf{B}^{-1}, \\
\mathbf{P}_{12} &= -\mathbf{B}^{-1}\mathbf{J}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{K}^T\left\{\mathbf{K}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{K}^T\right\}^{-1}, \\
\mathbf{P}_{13} &= -\mathbf{B}^{-1}\mathbf{J}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1} + \\
&\qquad \mathbf{B}^{-1}\mathbf{J}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{K}^T\left\{\mathbf{K}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{K}^T\right\}^{-1}\mathbf{K}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}, \\
\mathbf{P}_{22} &= \left\{\mathbf{K}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{K}^T\right\}^{-1}, \\
\mathbf{P}_{23} &= -\left\{\mathbf{K}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{K}^T\right\}^{-1}\mathbf{K}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}, \\
\mathbf{P}_{33} &= -(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1} + \\
&\qquad (\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{K}^T\left\{\mathbf{K}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{K}^T\right\}^{-1}\mathbf{K}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1},
\end{aligned}$$

and $\mathbf{P}_{21}$, $\mathbf{P}_{31}$, and $\mathbf{P}_{32}$ are the transposes of $\mathbf{P}_{12}$, $\mathbf{P}_{13}$, and $\mathbf{P}_{23}$, respectively, as it is easy to see that the matrix is symmetric. □

Suppose $\mathbf{x} \in \mathbf{X}_n$, δ is small enough for Theorem 1 to apply, and $(\hat{\omega}, \hat{\lambda})$ is a solution of equations (2.1) - (2.3) such that $\hat{\omega} \in U_\delta$. We now write the equations (2.1) - (2.3) in a different form:

$$\begin{pmatrix}
\mathbf{B}_{\phi^*} + \hat{\mathbf{b}}(\mathbf{x}) & \mathbf{0} & -\mathbf{J}_{\omega^*} - \hat{\mathbf{j}}(\mathbf{x}) \\
\mathbf{0} & \mathbf{0} & -\mathbf{K}_{\omega^*} - \hat{\mathbf{k}}(\mathbf{x}) \\
-\mathbf{J}_{\omega^*}^T - \hat{\mathbf{j}}'(\mathbf{x}) & -\mathbf{K}_{\omega^*}^T - \hat{\mathbf{k}}'(\mathbf{x}) & \mathbf{0}
\end{pmatrix}
\begin{pmatrix}
\hat{\phi} - \phi^* \\ \hat{\psi} - \psi^* \\ \hat{\lambda}
\end{pmatrix} =
\begin{pmatrix}
\frac{1}{n}\mathbf{s}(\mathbf{x}, \phi^*) \\ \mathbf{0} \\ \mathbf{0}
\end{pmatrix}, \tag{2.28}$$

where $\hat{\mathbf{b}}(\mathbf{x})$, $\hat{\mathbf{j}}(\mathbf{x})$, $\hat{\mathbf{j}}'(\mathbf{x})$, $\hat{\mathbf{k}}(\mathbf{x})$, and $\hat{\mathbf{k}}'(\mathbf{x})$ are matrices whose elements tend to 0 as δ goes to 0. Thus, by Lemma 2, if δ is sufficiently small, then the matrix

$$\begin{pmatrix}
\mathbf{B}_{\phi^*} + \hat{\mathbf{b}}(\mathbf{x}) & \mathbf{0} & -\mathbf{J}_{\omega^*} - \hat{\mathbf{j}}(\mathbf{x}) \\
\mathbf{0} & \mathbf{0} & -\mathbf{K}_{\omega^*} - \hat{\mathbf{k}}(\mathbf{x}) \\
-\mathbf{J}_{\omega^*}^T - \hat{\mathbf{j}}'(\mathbf{x}) & -\mathbf{K}_{\omega^*}^T - \hat{\mathbf{k}}'(\mathbf{x}) & \mathbf{0}
\end{pmatrix}$$

is also non-singular and we write its inverse as

$$\begin{pmatrix} \hat{\mathbf{P}}_{11}(\mathbf{x}) & \hat{\mathbf{P}}_{12}(\mathbf{x}) & \hat{\mathbf{P}}_{13}(\mathbf{x}) \\ \hat{\mathbf{P}}_{21}(\mathbf{x}) & \hat{\mathbf{P}}_{22}(\mathbf{x}) & \hat{\mathbf{P}}_{23}(\mathbf{x}) \\ \hat{\mathbf{P}}_{31}(\mathbf{x}) & \hat{\mathbf{P}}_{32}(\mathbf{x}) & \hat{\mathbf{P}}_{33}(\mathbf{x}) \end{pmatrix}.$$

Thus, if δ is sufficiently small and if $\mathbf{x} \in \mathbf{X}_n$, we can solve from the equation (2.28) that

$$\begin{pmatrix} \hat{\phi} - \phi^* \\ \hat{\psi} - \phi^* \\ \hat{\lambda} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{P}}_{11}(\mathbf{x}) & \hat{\mathbf{P}}_{12}(\mathbf{x}) & \hat{\mathbf{P}}_{13}(\mathbf{x}) \\ \hat{\mathbf{P}}_{21}(\mathbf{x}) & \hat{\mathbf{P}}_{22}(\mathbf{x}) & \hat{\mathbf{P}}_{23}(\mathbf{x}) \\ \hat{\mathbf{P}}_{31}(\mathbf{x}) & \hat{\mathbf{P}}_{32}(\mathbf{x}) & \hat{\mathbf{P}}_{33}(\mathbf{x}) \end{pmatrix} \begin{pmatrix} \frac{1}{n}\mathbf{s}(\mathbf{x}, \phi^*) \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}. \qquad (2.29)$$

Since the asymptotic distribution of $\mathbf{s}(\mathbf{x}, \phi^*)/n$ is known, we can use the above relationship to induce the asymptotic distribution of $(\hat{\omega}, \hat{\lambda})$. However, this may only be valid for $\mathbf{x} \in \mathbf{X}_n$, and we need to extend it to also account for $\mathbf{x} \notin \mathbf{X}_n$.

Let (δ_m), (ε_m) be two decreasing sequences of positive real numbers, such that $\delta_1 < \mu_1/\kappa_3$, $\varepsilon_1 < 1$, and δ_m and ε_m both tend to 0 as m goes to infinity. Define an increasing sequence (n_m) of integers such that, if $n \geq n_m$, there exists a set $\mathbf{X}_n$ with properties $(\mathscr{X}1)$ - $(\mathscr{X}4)$ for $\varepsilon = \varepsilon_m$ and $\delta = \delta_m$. For $m = 1, 2, \ldots,$, if $n_m \leq n < n_{m+1}$, we choose a set $\mathbf{X}_n$ with properties $(\mathscr{X}1)$ - $(\mathscr{X}4)$ for $\varepsilon = \varepsilon_m$ and $\delta = \delta_m$. When $\mathbf{x} \in \mathbf{X}_n$, the equations (2.1) - (2.3) have a solution $(\hat{\omega}_n, \hat{\lambda}_n)$ such that $||\hat{\omega}_n - \omega^*|| < \delta_m$, with $\hat{\omega}_n$ being the constrained maximum likelihood estimate for ω. Thus, $\hat{\omega}_n$ and $\hat{\lambda}_n$ satisfy the equation (2.29). When $\mathbf{x} \notin \mathbf{X}_n$, we define

$$\begin{pmatrix} \hat{\phi}_n - \phi^* \\ \hat{\psi}_n - \psi^* \\ \hat{\lambda}_n \end{pmatrix} = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{P}_{13} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \mathbf{P}_{23} \\ \mathbf{P}_{31} & \mathbf{P}_{32} & \mathbf{P}_{33} \end{pmatrix} \begin{pmatrix} \frac{1}{n}\mathbf{s}(\mathbf{x}, \phi^*) \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{P}_{ij}$, $i, j = 1, 2, 3$, are defined in the proof of Lemma 2. Note that the probability of $\mathbf{x} \notin \mathbf{X}_n$ goes to zero as n goes to infinity. Thus, we have defined two sequences of random variables, $(\hat{\omega}_n)$ and $(\hat{\lambda}_n)$, $n = n_m, n_{m+1}, \ldots$, which have the property that $\hat{\omega}_n$ converges in probability to ω^* as n goes to infinity. Moreover, $\hat{\omega}_n$

and $\hat{\lambda}_n$ jointly satisfy the equations (2.1) - (2.3).

Theorem 2.

$$\sqrt{n}\begin{pmatrix} \hat{\phi}_n - \phi^* \\ \hat{\psi}_n - \psi^* \\ \hat{\lambda}_n \end{pmatrix} \xrightarrow{d} \mathcal{N}_{s+t}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{0} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{P}_{33} \end{pmatrix}\right).$$

Proof. If $\mathbf{x} \notin \mathbf{X}_n$, we define $\hat{\mathbf{P}}_{ij}(\mathbf{x}) = \mathbf{P}_{ij}$, $i, j = 1, 2, 3$. Then, for sufficiently large n, we have

$$\sqrt{n}\begin{pmatrix} \hat{\phi}_n - \phi^* \\ \hat{\psi}_n - \psi^* \\ \hat{\lambda}_n \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{P}}_{11}(\mathbf{x}) & \hat{\mathbf{P}}_{12}(\mathbf{x}) & \hat{\mathbf{P}}_{13}(\mathbf{x}) \\ \hat{\mathbf{P}}_{21}(\mathbf{x}) & \hat{\mathbf{P}}_{22}(\mathbf{x}) & \hat{\mathbf{P}}_{23}(\mathbf{x}) \\ \hat{\mathbf{P}}_{31}(\mathbf{x}) & \hat{\mathbf{P}}_{32}(\mathbf{x}) & \hat{\mathbf{P}}_{33}(\mathbf{x}) \end{pmatrix}\begin{pmatrix} \sqrt{n}\begin{pmatrix} \frac{1}{n}\mathbf{s}(\mathbf{x}, \phi^*) \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \end{pmatrix}.$$

Since $\hat{\mathbf{b}}(\mathbf{x})$, $\hat{\mathbf{j}}(\mathbf{x})$, $\hat{\mathbf{j}}^*(\mathbf{x})$, $\hat{\mathbf{k}}(\mathbf{x})$, and $\hat{\mathbf{k}}^*(\mathbf{x})$ all tend to $\mathbf{0}$ as $\delta \to 0$, it follows that the elements of

$$\begin{pmatrix} \hat{\mathbf{P}}_{11}(\mathbf{x}) & \hat{\mathbf{P}}_{12}(\mathbf{x}) & \hat{\mathbf{P}}_{13}(\mathbf{x}) \\ \hat{\mathbf{P}}_{21}(\mathbf{x}) & \hat{\mathbf{P}}_{22}(\mathbf{x}) & \hat{\mathbf{P}}_{23}(\mathbf{x}) \\ \hat{\mathbf{P}}_{31}(\mathbf{x}) & \hat{\mathbf{P}}_{32}(\mathbf{x}) & \hat{\mathbf{P}}_{33}(\mathbf{x}) \end{pmatrix}$$

converge in probability to the elements of

$$\begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{P}_{13} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \mathbf{P}_{23} \\ \mathbf{P}_{31} & \mathbf{P}_{32} & \mathbf{P}_{33} \end{pmatrix}.$$

Moreover, it is known that the asymptotic distribution of $(\mathbf{s}(\mathbf{x}, \phi^*)/n)$ is normal with mean zero and asymptotic variance $\mathbf{B}_{\phi^*}$. Thus, we have

$$\sqrt{n}\begin{pmatrix} \frac{1}{n}\mathbf{s}(\mathbf{x}, \phi^*) \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \xrightarrow{d} \mathcal{N}_{s+t}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{B}_{\phi^*} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}\right).$$

It then follows that the asymptotic distribution of $\sqrt{n}\left(\hat{\phi}_n - \phi^*, \hat{\psi}_n - \psi^*, \hat{\lambda}_n\right)$ is

$$\mathcal{N}_{s+t}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{P}_{13}\\ \mathbf{P}_{21} & \mathbf{P}_{22} & \mathbf{P}_{23}\\ \mathbf{P}_{31} & \mathbf{P}_{32} & \mathbf{P}_{33}\end{pmatrix}\begin{pmatrix}\mathbf{B}_{\phi^*} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{P}_{13}\\ \mathbf{P}_{21} & \mathbf{P}_{22} & \mathbf{P}_{23}\\ \mathbf{P}_{31} & \mathbf{P}_{32} & \mathbf{P}_{33}\end{pmatrix}^T\right).$$

Finally, using the expressions for $\mathbf{P}_{ij}$, $i, j = 1,2,3$, that were derived in the proof of Lemma 2, it can be verified that the asymptotic variance simplifies to

$$\begin{pmatrix}\mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{0}\\ \mathbf{P}_{21} & \mathbf{P}_{22} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & -\mathbf{P}_{33}\end{pmatrix}.$$

The result then follows. □

2.3.3 Numerical algorithm

The solution of the equations (2.1) - (2.3), say $\hat{\omega} = (\hat{\phi}, \hat{\psi})$, usually does not have a closed form, and thus must be computed numerically. We may immediately consider the Newton-Raphson method to solve the problem. However, that method requires the form of the Hessian matrix of $\mathbf{h}(\omega)$, which is an $s \times s$ matrix and may be very complicated, especially when s is large. Thus, we follow the approach proposed by [1] and develop an algorithm that is easier to implement.

Suppose $\omega^{(0)} = (\phi^{(0)}, \psi^{(0)})$ is an initial guess for $\hat{\omega}$ such that $||\omega^{(0)} - \hat{\omega}||$ is small. Then we consider a first order of approximation to $\mathbf{s}(\mathbf{x}, \hat{\phi})$ and $\mathbf{h}(\hat{\omega})$:

$$\begin{aligned}\mathbf{s}(\mathbf{x}, \hat{\phi}) &\approx \mathbf{s}(\mathbf{x}, \phi^{(0)}) + \mathbf{M}_{\mathbf{x}, \phi^{(0)}}(\hat{\phi} - \phi^{(0)}),\\ \mathbf{h}(\hat{\omega}) &\approx \mathbf{h}(\omega^{(0)}) + \mathbf{J}^T_{\omega^{(0)}}(\hat{\phi} - \phi^{(0)}) + \mathbf{K}^T_{\omega^{(0)}}(\hat{\psi} - \psi^{(0)}).\end{aligned}$$

Also, we assume that $\hat{\lambda}$ is close to $\mathbf{0}$ when n is large. Then to a first order of

approximation, we have

$$\mathbf{J}_{\hat{\omega}}\hat{\lambda} \approx \mathbf{J}_{\omega^{(0)}}\hat{\lambda},$$
$$\mathbf{K}_{\hat{\omega}}\hat{\lambda} \approx \mathbf{K}_{\omega^{(0)}}\hat{\lambda}.$$

Since $\hat{\omega}$ and $\hat{\lambda}$ jointly satisfy the equations (2.1) - (2.3), they should also approximately satisfy

$$\begin{pmatrix} -\frac{1}{n}\mathbf{M}_{\mathbf{x},\phi^{(0)}} & \mathbf{0} & -\mathbf{J}_{\omega^{(0)}} \\ \mathbf{0} & \mathbf{0} & -\mathbf{K}_{\omega^{(0)}} \\ -\mathbf{J}^T_{\omega^{(0)}} & -\mathbf{K}^T_{\omega^{(0)}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\phi} - \phi^{(0)} \\ \hat{\psi} - \psi^{(0)} \\ \hat{\lambda} \end{pmatrix} \approx \begin{pmatrix} \frac{1}{n}\mathbf{s}(\mathbf{x}, \phi^{(0)}) \\ \mathbf{0} \\ \mathbf{h}(\omega^{(0)}) \end{pmatrix}.$$

When n is large, $-\mathbf{M}_{\mathbf{x},\phi^{(0)}}/n$ should be close to $\mathbf{B}_{\phi^{(0)}}$. Thus, we use $\mathbf{B}_{\phi^{(0)}}$ to approximate $-\mathbf{M}_{\mathbf{x},\phi^{(0)}}/n$. Finally, we have the formula for updating $\omega^{(0)}$, and in general for updating $\omega^{(r-1)}$ in the r-th iteration,

$$\begin{pmatrix} \phi^{(r)} \\ \lambda^{(r)} \end{pmatrix} = \begin{pmatrix} \phi^{(r-1)} \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{B}_{\phi^{(r-1)}} & -\mathbf{H}_{\phi^{(r-1)}} \\ -\mathbf{H}^T_{\phi^{(r-1)}} & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{n}\mathbf{s}(\mathbf{x}, \phi^{(r-1)}) \\ \mathbf{h}(\phi^{(r-1)}) \end{pmatrix}.$$

If the sequence $\left\{(\omega^{(r)}, \lambda^{(r)})\right\}$ converges, then it converges to a solution of the equation (2.1) - (2.3). The convergence of these sequences may depend on the initial value used. For our problem, we can use the unconstrained maximum likelihood estimate of ϕ plus a reasonable guess on the value of ψ as the initial value for ω, which we believe provides a good starting point. Thus, we expect these sequences to converge in most practical situations. Finally, it should be noted that $\lambda^{(r-1)}$ is actually missing from the right hand side of the above equation. Thus, the updating procedure only needs to store the current value of $\omega = (\phi, \psi)$ for the next iteration.

2.4 Example problem and simulation study

In this section, we use the proposed method to solve a missing data problem, where parameters associated with the missing mechanism may only be identified with additional assumptions. This sort of problem might otherwise be tackled with an expectation-maximization algorithm. More specifically, consider a binary response variable Y and two binary explanatory variables X_1 and X_2. The probability of having $Y = 1$ given (X_1, X_2) is assumed to be determined through a logistic model:

$$\text{logit}Pr(Y = 1|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2,$$

where $\text{logit}(p) = \log(p) - \log(1-p)$. Suppose we can observe X_1 and X_2 for everyone sampled, but the status of Y is missing for some people. Let R indicate missingness. The data structure is displayed in Table 1, where n_{ijk} is the number of subjects with complete data of $(Y = i, X_1 = j, X_2 = k, R = 1)$, and m_{jk} is the number of subjects with incomplete data of $(X_1 = j, X_2 = k, R = 0)$, $i, j, k = 0, 1$. The corresponding cell probabilities, as enclosed in parentheses in the Table 2.1, are

$$r_{ijk} = Pr(Y = i, X_1 = j, X_2 = k, R = 1),$$

$$s_{jk} = Pr(X_1 = j, X_2 = k, R = 0),$$

for $i, j, k = 0, 1$. Based on Table 2.1, the log-likelihood of data is:

$$\ell = \sum_{i,j,k} n_{ijk} \log r_{ijk} + \sum_{j,k} m_{jk} \log s_{jk}.$$

In order to understand the relationship between Y and (X_1, X_2), we need to infer the proportions of subjects with $Y = 1$ among the groups of incomplete data

$$t_{jk} = Pr(Y = 1|X_1 = j, X_2 = k, R = 0),$$

for $j, k = 0, 1$. However, these quantities are not identifiable from data without additional assumptions.

Now, we make two assumptions. First, we assume that the status of Y is miss-

	$Y=0,\ R=1$	$Y=1,\ R=1$	$Y=?,\ R=0$
$X_1=0,\ X_2=0,$	$n_{000}\ (r_{000})$	$n_{100}\ (r_{100})$	$m_{00}\ (s_{00})$
$X_1=1,\ X_2=0,$	$n_{010}\ (r_{010})$	$n_{110}\ (r_{110})$	$m_{10}\ (s_{10})$
$X_1=0,\ X_2=1,$	$n_{001}\ (r_{001})$	$n_{101}\ (r_{101})$	$m_{01}\ (s_{01})$
$X_1=1,\ X_2=1,$	$n_{011}\ (r_{011})$	$n_{111}\ (r_{111})$	$m_{11}\ (s_{11})$

Table 2.1: Data structure for the example problem considered in Section 2.4.

ing at random, i.e., R and Y are conditionally independent given (X_1, X_2). This assumption imposes four constraints on parameters, and implies

$$\log t_{jk} - \log(1-t_{jk}) = \log r_{1jk} - \log r_{0jk},$$

for $j,k = 0,1$. Secondly, we assume that the effects of X_1 and X_2 on Y are additive on the logit scale, which means that the interaction effect β_3 is zero. This assumption introduces one more constraint on parameters as

$$\log \frac{(r_{100}+s_{00}t_{00})(r_{111}+s_{11}t_{11})}{(r_{101}+s_{01}t_{01})(r_{110}+s_{10}t_{10})} = \log \frac{(r_{000}+s_{00}(1-t_{00}))(r_{011}+s_{11}(1-t_{11}))}{(r_{001}+s_{01}(1-t_{01}))(r_{010}+s_{10}(1-t_{10}))}.$$

Under these two assumptions, we can apply the proposed method to obtained the maximum likelihood estimates $\hat{r}_{ijk}$, $\hat{s}_{jk}$, and $\hat{t}_{jk}$, $i,j,k = 0,1$, subject to the above five constraints. Next, the constrained maximum likelihood estimates for the main effects of X_1 and X_2 can be deduced through

$$\hat{\beta}_1 = \log \frac{\hat{r}_{110}+\hat{s}_{10}\hat{t}_{10}}{\hat{r}_{100}+\hat{s}_{00}\hat{t}_{00}} - \log \frac{\hat{r}_{010}+\hat{s}_{10}(1-\hat{t}_{10})}{\hat{r}_{000}+\hat{s}_{00}(1-\hat{t}_{00})},$$
$$\hat{\beta}_2 = \log \frac{\hat{r}_{101}+\hat{s}_{01}\hat{t}_{01}}{\hat{r}_{100}+\hat{s}_{00}\hat{t}_{00}} - \log \frac{\hat{r}_{001}+\hat{s}_{01}(1-\hat{t}_{01})}{\hat{r}_{000}+\hat{s}_{00}(1-\hat{t}_{00})},$$

and the corresponding estimated variances can be obtained by the delta method.

Finally, based on the above problem, we conduct a simulation study to illustrate the performance of the proposed method. In particular, we randomly generate 10000 datasets of size 1000 under the parameter setting $\beta_0 = \text{logit}\ 0.1$, $\beta_1 = \log 2$,

$\beta_2 = \log 3$, $\beta_3 = 0$, and

$$
\begin{aligned}
&Pr(X_1 = 0, X_2 = 0) = 0.4, \quad Pr(R = 0 | X_1 = 0, X_2 = 0) = 0.2, \\
&Pr(X_1 = 1, X_2 = 0) = 0.3, \quad Pr(R = 0 | X_1 = 1, X_2 = 0) = 0.1, \\
&Pr(X_1 = 0, X_2 = 1) = 0.2, \quad Pr(R = 0 | X_1 = 0, X_2 = 1) = 0.05, \\
&Pr(X_1 = 1, X_2 = 1) = 0.1, \quad Pr(R = 0 | X_1 = 1, X_2 = 1) = 0.05.
\end{aligned}
$$

For each dataset, we apply the proposed method to obtain the constrained maximum likelihood estimates for $\hat{\beta}_1$ and $\hat{\beta}_2$, and the associated 95% confidence intervals. Our simulation results show that the empirical biases for the estimators of β_1 and β_2 are 0.0033 and 0.0029, respectively. Correspondingly, the coverage probabilities of the 95% confidence intervals are 95.1% and 95.2%, which match well with the nominal level. We can see that the proposed method performs very well.

2.5 Just- and over-identified situations

In the previous section, we have considered a partially identified model with four non-identifiable parameters and made additional assumptions that impose five constraints on parameters. Consequently, the constrained maximum likelihood estimators for the identifiable parameters, r_{ijk}'s and s_{jk}'s, $i, j, k = 0, 1$, differ from their unconstrained estimators. More importantly, compared to the unconstrained estimators, the constrained estimators are associated with smaller variances. For example, under the parameter setting considered in the previous section, the variance of the asymptotic distribution of the the unconstrained estimator for $(r_{000}, \ldots, r_{111})$

is

$$\begin{pmatrix}
\mathbf{0.205} & -0.06 & -0.04 & -0.02 & -0.01 & -0.01 & -0.01 & -0.01 \\
-0.06 & \mathbf{0.172} & -0.03 & -0.01 & -0.01 & -0.01 & -0.01 & -0.01 \\
-0.04 & -0.03 & \mathbf{0.122} & -0.01 & -0.01 & -0.01 & -0.01 & -0.01 \\
-0.02 & -0.01 & -0.01 & \mathbf{0.054} & -0.00 & -0.00 & -0.00 & -0.00 \\
-0.01 & -0.01 & -0.01 & -0.00 & \mathbf{0.031} & -0.00 & -0.00 & -0.00 \\
-0.01 & -0.01 & -0.01 & -0.00 & -0.00 & \mathbf{0.047} & -0.00 & -0.00 \\
-0.01 & -0.01 & -0.01 & -0.00 & -0.00 & -0.00 & \mathbf{0.045} & -0.00 \\
-0.01 & -0.01 & -0.01 & -0.00 & -0.00 & -0.00 & -0.00 & \mathbf{0.037}
\end{pmatrix},$$

and the asymptotic variance of the corresponding constrained estimator is

$$\begin{pmatrix}
\mathbf{0.197} & -0.06 & -0.03 & -0.02 & -0.00 & -0.02 & -0.02 & -0.00 \\
-0.06 & \mathbf{0.165} & -0.04 & -0.01 & -0.02 & -0.00 & -0.00 & -0.02 \\
-0.03 & -0.04 & \mathbf{0.115} & -0.00 & -0.01 & 0.00 & 0.00 & -0.01 \\
-0.02 & -0.01 & -0.00 & \mathbf{0.046} & 0.01 & -0.01 & -0.01 & 0.01 \\
-0.00 & -0.02 & -0.01 & 0.01 & \mathbf{0.023} & 0.01 & 0.01 & -0.01 \\
-0.02 & -0.00 & 0.00 & -0.01 & 0.01 & \mathbf{0.039} & -0.01 & 0.01 \\
-0.02 & -0.00 & 0.00 & -0.01 & 0.01 & -0.01 & \mathbf{0.038} & 0.01 \\
-0.00 & -0.02 & -0.01 & 0.01 & -0.01 & 0.01 & 0.01 & \mathbf{0.029}
\end{pmatrix}.$$

By comparing the elements along the diagonal of these two matrices, it is clear that the constrained estimator is more efficient than the unconstrained estimator for the problem considered in the previous section.

However, if we only make the missing at random assumption and allow the model for $Y|X_1,X_2$ to be saturated, then we have only four constraints for four non-identifiable parameters. In this case, we find that the constrained and unconstrained maximum likelihood estimators for the identifiable parameters always coincide and have the same asymptotic distribution. Thus, making the missing at random assumption alone leads to no efficiency gain.

Generally, we say that the parameters are over-identified when the number of constraints is greater than the number of unidentified parameters. In this case, the constrained maximum likelihood estimator differs from the unconstrained estimator and achieves better efficiency. On the other hand, we say that the parameters are just-identified when the number of constraints is equal to the number of unidentified parameters. If that is the case, then the constrained estimator will coincide with the unconstrained estimator, at least asymptotically. Moreover, identifying the unidentified parameters uses up the information provided by the additional constraints and thus a more efficient estimator is not available. This phenomena was also observed by Chen and Chen [4] in the context of a gene-environment independence problem.

2.6 Conclusion

Parameters arising from a partially identified model can be estimated when we have enough equality constraints enforced by additional assumptions. Moreover, the constrained maximum likelihood estimator for the identified part may or may not coincide with its unconstrained counterpart, and achieves higher efficiency when they do not coincide.

Another possibility for estimating parameters of a partially identified model subject to constraints is to exploit a reduced-form parameterization that is free of constraints. However, the capability of such approach is limited, as a closed form for a reduced-form parameterization is often very complicated or even sometimes not available. In contrast, the method presented in this paper is applicable in more general settings. Moreover, since the log-likelihood function is usually expressed in its simplest form with a transparent re-parameterization, taking the second partial derivatives of the log-likelihood function becomes much more straightforward. Thus, the proposed method is also advantageous in terms of calculation.

Chapter 3

The Benefit of Exploiting the GEI Assumption for Analyzing Case-Control Data

3.1 Introduction

In this chapter, we apply the constrained maximum likelihood estimation theory developed in Chapter 2 to study the benefit of exploiting the GEI assumption for analyzing case-control data concerning a binary genotype and an environmental exposure that has two or more categories. This chapter is organized as follows. We first describe the underlying model for case-control data and formulate the problem under consideration in more detail. We then develop a reparameterization of the model and transform the problem into a constrained maximum likelihood estimation problem. Next, we propose methods for analyzing case-control data exploiting the GEI assumption in different scenarios. We then investigate the efficiency gain of exploiting the GEI assumption, and conduct simulation studies to compare the performance of the proposed GEI-based methods with the traditional method. We also consider a real dataset for the application of the proposed method. Finally, some concluding thoughts are given at the end of this chapter.

3.2 Formulation of the problem

Let D be the binary disease status, with $D = 1$ for presence and $D = 0$ for absence. Suppose the risk of having the disease is affected by a subject's binary genetic factor G, coded as $\{0,1\}$, and a categorical environmental exposure E with $K+1$ levels, coded as $\{0,1,\ldots,K\}$. In a case-control study, genotype and environmental exposure status are collected for $n = n_0 + n_1$ subjects, including n_0 controls and n_1 cases, where the case-to-control ratio, n_1/n_0, is pre-specified. Let n_{ijk} denote the number of subjects in the $D = i$ study arm with genotype $G = j$ and environmental exposure $E = k$. Thus, we can summarize case-control data in a $2 \times 2 \times (K+1)$ table.

Let $\iota = (\iota_{00},\ldots,\iota_{0K},\iota_{10},\ldots,\iota_{1K})$ denote the vector of probability masses for the joint distribution of genotype and exposure, with $\iota_{jk} = Pr(G = j, E = k)$ for $j = 0,1$ and $k = 0,\ldots,K$. It has $2K+1$ degrees of freedom since the sum of all elements is equal to one. We assume that the disease risk, given a subject's status of genotype and environmental exposure, is parameterized by a saturated logistic regression model:

$$Pr(D=1|E,G) = \text{expit}\left\{\beta_0 + \beta_G G + \sum_{k=1}^{K}\left(\beta_E^{(k)}\mathbf{1}_{\{k\}}(E) + \beta_{EG}^{(k)}\mathbf{1}_{\{k\}}(E)G\right)\right\},$$

where $\text{expit}(x) = 1/\{1+\exp(-x)\}$ is the inverse of the logit function. For brevity, let $\beta_E = (\beta_E^{(1)},\ldots,\beta_E^{(K)})$ represent the vector of all main environmental effects, and $\beta_{EG} = (\beta_{EG}^{(1)},\ldots,\beta_{EG}^{(K)})$ represent the vector of all gene-environment interaction effects. Then, the model can be parameterized by $\phi = (\iota,\beta_0,\beta_G,\beta_E,\beta_{EG})$, which has $4K+3$ degrees of freedom.

Under the GEI assumption, the joint distribution of genotype and environmental exposure can be determined by their marginal distributions. Let $\kappa = (\kappa_0,\kappa_1)$ and $\delta = (\delta_0,\ldots,\delta_K)$ denote two vectors of probability masses for the marginal distributions of genotype and environmental exposure, respectively. The numbers of free parameters in these two vectors are 1 and K, respectively, as the sum of all elements in each vector is equal to one. Then the joint probability takes the product form $\iota_{jk} = \kappa_j\delta_k$ for $j = 0,1$ and $k = 0,\ldots,K$. Therefore, we get a reduced form parameterization $\phi_r = (\kappa,\delta,\beta_0,\beta_G,\beta_E,\beta_{EG})$, which has $3K+3$ degrees of

freedom. We can then estimate ϕ_r by maximizing the retrospective log-likelihood for case-control data

$$\ell(\phi_r) = \sum_{i,j,k} n_{ijk} \log \frac{\kappa_j \delta_k p_{ijk}(\beta_0, \beta_G, \beta_E, \beta_{EG})}{\sum_{j',k'} \kappa_{j'} \delta_{k'} p_{ij'k'}(\beta_0, \beta_G, \beta_E, \beta_{EG})},$$

where $p_{ijk}(\beta_0, \beta_G, \beta_E, \beta_{EG}) = \Pr(D = i | G = j, E = k),\ i, j = 0, 1,\ k = 0, \ldots, K.$ However, due to the complicated form with the denominator including summation over all control or case cell probabilities, direct maximization of the log-likelihood with respect to ϕ_r can be numerically challenging or even infeasible, especially when K is large.

Alternatively, Chatterjee and Carroll [3] proposed to obtain the estimate using a profile-likelihood technique. The likelihood is first maximized with respect to δ for fixed values of (κ, β) to derive the profile likelihood of the data, where β generically represents $(\beta_0, \beta_G, \beta_E, \beta_{EG})$. The profile likelihood is then maximized with respect to (κ, β) to obtain the maximum likelihood estimator of (κ, β). If $\hat{\delta}(\kappa, \beta)$ denotes the value of δ that maximizes the likelihood for fixed (κ, β), the profile likelihood is then $\ell(\kappa, \beta, \hat{\delta}(\kappa, \beta))$. Chatterjee and Carroll [3] have shown that the profile likelihood $\ell(\kappa, \beta, \hat{\delta}(\kappa, \beta))$ can be computed without having to maximize the log-likelihood numerically with respect to the potentially high-dimensional parameter δ. Instead, it can be obtained in a closed form up to only one additional parameter. More specifically, let θ denote the disease prevalence $Pr(D = 1)$, and S denote the indicator of whether or not a subject has been selected in the case-control sample. We consider the joint probability distribution for D and G given E in the case-control sample and let

$$Pr(D = i, G = j | E = k, S = 1) = \frac{n_i \left\{ \theta^{1-i} (1-\theta)^i \right\} \kappa_j p_{ijk}(\beta_0, \beta_G, \beta_E, \beta_{EG})}{\sum_{i',j'} n_{i'} \left\{ \theta^{1-i'} (1-\theta)^{i'} \right\} \kappa_{j'} p_{i'j'k}(\beta_0, \beta_G, \beta_E, \beta_{EG})},$$

which only concerns parameters κ, β, and θ. Let n_{++k} be the marginal frequency of the kth category of E for $k = 0, \ldots, K$. Then the profile likelihood $\ell(\kappa, \beta, \hat{\delta}(\kappa, \beta))$ can be computed as $\ell^*(\kappa, \beta, \hat{\theta}(\kappa, \beta))$, where

$$\ell^*(\kappa, \beta, \theta(\kappa, \beta)) = \sum_{i,j,k} n_{ijk} \log Pr(D = i, G = j | E = k, S = 1),$$

and $\hat{\theta}(\kappa,\beta)$ is defined by the solution of the equation:

$$n_1 = \sum_k \sum_j n_{++k} Pr(D=1, G=j|E=k, S=1).$$

Thus, the semiparametric maximum likelihood estimate of (κ,β) can be obtained by solving score equations $\partial \ell^*(\kappa,\beta,\theta)/\partial(\kappa,\beta,\theta) = 0$ jointly with respect to κ, β, and θ. However, given the complex form of $Pr(D,G|E,S=1)$, numerically solving these estimating equations using standard methods, such as the Newton-Raphson method, is still challenging. Moreover, although it has been shown in [3] that, under suitable regularity conditions, the semiparametric maximum likelihood estimator is consistent and asymptotically follows a normal distribution, the asymptotic variance of the estimator was given in the form that includes a double expectation, first with respect to $Pr(D,G|E,S=1)$ and then with respect to $Pr(G,E|D)$. Thus, it is not trivial to obtain the standard error of the estimate. Therefore, we are seeking a simpler solution for analyzing case-control data under the GEI assumption.

3.3 A reparameterization of the model

We now consider a different parameterization of the model such that the form of the log-likelihood function can be greatly simplified. First, as already defined in the previous section, we use θ to denote the disease prevalence $Pr(D=1)$. Also, we define $\gamma = (\gamma_{001},\ldots,\gamma_{01K},\gamma_{101},\ldots,\gamma_{11K})$, where $\gamma_{ijk} = Pr(G=j,E=k|D=i)$, for $i,j=0,1$ and $k=0,\ldots,K$, are sampling probabilities actually underlying case-control data. Note that γ_{000} and γ_{100} are excluded from γ as their values can be uniquely determined through the constraints $\sum_{j,k}\gamma_{ijk}=1$ for $i=0,1$. Now consider the parameterization $\xi=(\gamma,\theta)$. It can be easily verified that the mapping between ξ and ϕ is invertible. Particularly, by comparing subjects with environmental exposure $E=k$ to those with baseline level $E=0$, we are able to express $\beta_{EG}^{(k)}$ by ξ through

$$\beta_{EG}^{(k)}(\gamma) = \log \frac{\gamma_{11k}\gamma_{100}\gamma_{00k}\gamma_{010}}{\gamma_{10k}\gamma_{110}\gamma_{01k}\gamma_{000}}, \tag{3.1}$$

for $k=1,\ldots,K$.

Suppose $\hat{\xi} = (\hat{\gamma}, \hat{\theta})$ is an estimator of ξ and its asymptotic distribution is

$$\sqrt{n}\begin{pmatrix} \hat{\gamma} - \gamma^* \\ \hat{\theta} - \theta^* \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_\gamma & \Sigma_{\gamma\cdot\theta} \\ \Sigma_{\gamma\cdot\theta}^T & \Sigma_\theta \end{pmatrix}\right),$$

where $\xi^* = (\gamma^*, \theta^*)$ is the true value of ξ. We can then easily deduce an estimator of ϕ and apply the delta method to derive the corresponding asymptotic distribution. In particular, the estimator of the interaction effect $\beta_{EG}^{(k)}$ can be derived from the equation (3.1) as $\hat{\beta}_{EG}^{(k)} = \beta_{EG}^{(k)}(\hat{\gamma})$, and its asymptotic distribution is

$$\sqrt{n}\left(\hat{\beta}_{EG}^{(k)} - \beta_{EG}^{(k)*}\right) \xrightarrow{d} \mathcal{N}\left(0,\ \nabla_k^{*T}\Sigma_{\gamma^*}\nabla_k^*\right), \tag{3.2}$$

where ∇_k^* is the gradient of the function $\beta_{EG}^{(k)}(\gamma)$ evaluated at γ^*. Moreover, in practice, the asymptotic standard error of $\hat{\beta}_{EG}^{(k)}$ is

$$\mathrm{SE}\left(\hat{\beta}_{EG}^{(k)}\right) = \frac{1}{\sqrt{n}}\sqrt{\hat{\nabla}_k\Sigma_{\hat{\gamma}}\hat{\nabla}_k}, \tag{3.3}$$

where $\hat{\nabla}_k$ is the gradient of the function $\beta_{EG}^{(k)}(\gamma)$ evaluated at $\hat{\gamma}$. Therefore, it is sufficient to analyze case-control data based on the parameterization ξ.

With the new parameterization ξ, we first re-write the retrospective log-likelihood function in a much simpler form

$$\ell(\gamma) = \sum_{i,j,k} n_{ijk}\log\gamma_{ijk},$$

which shows even more clearly that, without any additional assumption, the model can only be partially identified, leaving the parameter θ not identified. Let $\hat{\gamma}^{(U)}$ denote the unconstrained maximum likelihood estimator of γ. We can easily solve the equation $\partial\ell(\gamma)/\partial\gamma = 0$ with respect to γ, yielding $\hat{\gamma}_{ijk}^{(U)} = n_{ijk}/n_i$ for $i, j = 0, 1$ and $k = 0, \ldots, K$. The asymptotic distribution of $\hat{\gamma}^{(U)}$ is

$$\sqrt{n}\left(\hat{\gamma}^{(U)} - \gamma^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\ \mathbf{B}_{\gamma^*}^{-1}\right), \tag{3.4}$$

where $\mathbf{B}_{\gamma^*}$ is the unconstrained Fisher information, i.e., the negative of the expec-

tation of the second derivative of the log-likelihood function $\ell(\gamma)$, which is given in Appendix A.

Next, we look for the mathematical representation of the GEI assumption with the new parameterization of the model. Under the GEI assumption, we have $\iota_{00}\iota_{1k} = \kappa_0\delta_0\kappa_1\delta_k = \iota_{10}\iota_{0k}$, for $k = 1,\ldots,K$. Note that ι_{jk} can be expressed by ξ through $\iota_{jk}(\xi) = (1-\theta)\gamma_{0jk} + \theta\gamma_{1jk}$. Thus, the GEI assumption can be enforced onto ξ through the following constraints

$$\mathbf{g}(\xi) = (g_1(\xi),\ldots,g_K(\xi))^T = \mathbf{0},$$

where, for $k = 1,\ldots,K$,

$$g_k(\xi) = \iota_{00}(\xi)\iota_{1k}(\xi) - \iota_{10}(\xi)\iota_{0k}(\xi).$$

These equality constraints define a GEI subspace of ξ, within which ξ can be mapped to ϕ_r and the mapping is invertible.

Finally, since each component of ξ is a probability, the following inequality constraints apply naturally on ξ:

$$0 \leq \gamma_{ijk} \leq 1, \tag{3.5}$$

for $i, j = 0, 1$, $k = 0,\ldots,K$, and

$$0 \leq \theta \leq 1. \tag{3.6}$$

Therefore, the problem is now transformed to finding the estimate of ξ that maximizes $\ell(\gamma)$ subject to the equality constraints $\mathbf{g}(\xi) = 0$ and the inequality constraints (3.5) and (3.6).

3.4 Estimation with known disease prevalence

We begin with the special case where the disease prevalence is known, say from an external source. Suppose the true value of θ is known to be θ^*. First, we can plug θ^* into $\mathbf{g}(\xi)$ and thus the constraints induced by the GEI assumption now become $\mathbf{g}(\gamma,\theta^*) = 0$, which are equations concerning γ only. Secondly, the inequality

constraint (3.6) is no longer in effect. Thus, the problem now becomes finding the estimate of γ that maximizes $\ell(\gamma)$ subject to the equality constraint $\mathbf{g}(\gamma, \theta^*) = 0$ and the inequality constraints (3.5).

3.4.1 Theoretical properties of the estimator

We first study the asymptotic distribution of the constrained maximum likelihood estimator of γ. Given that the true value of θ is known, within a sufficiently small neighborhood of the true value of γ, the inequality constraints are automatically satisfied. Thus, in order to study the asymptotic properties of the estimator of γ, we can ignore the inequality constraints (3.5) and treat the problem as estimating parameters subject to equality constraints only, which has been studied by Aitchison and Silvey [1].

Consider an auxiliary function $T(\gamma, \lambda) = (1/n)\ell(\gamma) + \lambda^T \mathbf{g}(\gamma; \theta^*)$, where λ is a Lagrange multiplier vector of length K. The maximum point of $T(\gamma, \lambda)$ can be found by solving $\partial T(\gamma, \lambda)/\partial(\gamma, \lambda) = \mathbf{0}$ jointly with respect to γ and λ, which leads to the following $5K + 2$ equations:

$$\frac{1}{n}\mathbf{s}(\gamma) + \mathbf{J}_{\gamma;\theta^*}\lambda = \mathbf{0}, \tag{3.7}$$

$$\mathbf{g}(\gamma; \theta^*) = \mathbf{0}, \tag{3.8}$$

where $\mathbf{s}(\gamma)$ is the gradient of the log-likelihood function, and $\mathbf{J}_\xi$ is the Jacobian of $\mathbf{g}(\xi)$ with respect to γ, both of which are given in Appendix A.

Suppose $(\hat{\gamma}, \hat{\lambda})$ is the solution to the equations (3.7) and (3.8). Applying the theory in [1], as sample size n goes to infinity with the case-to-control ratio n_1/n_0 fixed, the asymptotic distribution of $(\hat{\gamma}, \hat{\lambda})$ is

$$\sqrt{n}\begin{pmatrix} \hat{\gamma} - \gamma^* \\ \hat{\lambda} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{P}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{22} \end{pmatrix}\right), \tag{3.9}$$

where

$$\begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{B}_{\gamma^*} & -\mathbf{J}_{\gamma^*,\theta^*} \\ -\mathbf{J}_{\gamma^*,\theta^*}^T & \mathbf{0} \end{pmatrix}.$$

We have shown earlier that $\mathbf{B}_{\gamma^*}$ represents the unconstrained Fisher information. Analogously, we refer to the matrix on the right hand side of the above equation as the 'constrained Fisher information' that accounts for the GEI assumption.

Finally, it is easy to see from the equation (3.9) that the variance of the asymptotic distribution of $\hat{\gamma}$ is $\Sigma_\gamma = \mathbf{P}_{11}$. We can then apply the (3.2) to derive the asymptotic distributions of the constrained maximum likelihood estimators of the interaction effects, $\hat{\beta}_{EG}^{(k)}$ for $k = 1, \ldots, K$, under the GEI assumption with known disease prevalence.

3.4.2 Numerical algorithm

In practice, we need to compute the maximum likelihood estimate $\hat{\gamma}$ numerically, since there is no closed form solution. In general, there are various numerical algorithms, such as the augmented Lagrangian algorithm [5] and the sequential quadratic programming algorithm [9], for solving the constrained optimization problem. For this particular problem, however, naively implementing these algorithms using the existing tools like the *NLopt* library [16] encounters numerical errors mainly because the target function has no definition when the inequality constraints (3.5) are violated. More care needs to be taken for successful implementations of these advanced algorithms, which can be tricky. Alternatively, we find that a simple variant of the Newton's method as proposed in [1] is easy to implement and works well in practice.

We first find the constrained maximum likelihood estimate of γ subject to the equality constraints $\mathbf{g}(\gamma, \theta^*) = 0$, ignoring the inequality constraints (3.5) for the moment. Specifically, we begin by setting the initial value of γ to be $\gamma^{(U)}$, as it is expected that the constrained and the unconstrained maximum likelihood estimate of γ should be close, if not equal, to each other. We then iteratively update the value of γ by

$$\begin{pmatrix} \gamma \\ \lambda \end{pmatrix} \leftarrow \begin{pmatrix} \gamma \\ \mathbf{0} \end{pmatrix} + c \begin{pmatrix} \mathbf{B}_\gamma & -\mathbf{J}_{\gamma,\theta^*} \\ -\mathbf{J}_{\gamma,\theta^*}^T & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{n}\mathbf{s}(\gamma) \\ \mathbf{g}(\gamma, \theta^*) \end{pmatrix}, \tag{3.10}$$

where $c \in (0, 1]$ is a tuning parameter that controls the step size of each update and prevent the values of γ from leaving their plausible range. We find $c = 0.5$ works

very well in practice. Also, although we update the value of λ in each iteration, we don't need to keep track of it because it won't be used in the next iteration. This process is repeated until convergence or termination, as summarized below in Algorithm 1. The termination criteria include reaching the preset maximum number of iterations and running into numerical errors, both of which occur very rarely.

Algorithm 1 Find $\hat{\gamma}$ under the GEI assumption with known θ.

Set the initial value $\gamma^{(0)}$
For $m = 1, 2, \ldots, M$
    Compute $\gamma^{(m)}$ given $\gamma^{(m-1)}$ using the equation (3.10)
    If any error occurs:
        Set $m = M$ and break
    If $||\gamma^{(m)} - \gamma^{(m-1)}||_\infty < \varepsilon$
        Break
If $m < M$
    Output $\hat{\gamma} = \gamma^{(m)}$

We now discuss the effect of the inequality constraints (3.5). If Algorithm 1 converges and successfully finds a constrained maximum likelihood estimate $\hat{\gamma}$, then the inequality constraints (3.5) should be naturally satisfied, otherwise the log-likelihood does not even exist. However, it is possible, though very rarely occurred in our simulation studies, that Algorithm 1 may terminate due to the violation of these inequality constraints before it finds the final estimate. In that case, we need to use a smaller step size by decreasing the value of c to prevent our algorithm moving too fast.

Finally, having the constrained maximum likelihood estimate $\hat{\gamma}$, we can plug it into the equations (3.1) and (3.3), with $\Sigma_\gamma = \mathbf{P}_{11}$, to get the constrained maximum likelihood estimates for the interaction effects and their associated standard errors.

3.5 Estimation with unknown disease prevalence

We now come back to our original problem in the more common settings where the disease prevalence is unknown. It needs to be learned indirectly through the

constraints imposed by the GEI assumption.

3.5.1 Parameter identification

Before we study estimation, we first discuss the identifiability of the parameters. In general settings where no assumption is made about the joint distribution of genotype and environmental exposure, it is well known that neither the joint distribution of genotype and environmental exposure, ι, nor the intercept parameter of the logistic model, β_0, is identifiable from case-control data. Under the GEI assumption, Chatterjee and Carroll [3] claimed that these parameters are theoretically identifiable from the retrospective likelihood of case-control data, except for some boundary situations where the logistic model for $Pr(D|G,E)$ depends only on G or only on E, but not both. In other words, either $(\beta_G, \beta_E, \beta_{EG}) = (0, \beta_E, 0)$ or $(\beta_G, \beta_E, \beta_{EG}) = (\beta_G, 0, 0)$, corresponding to either only the main effect of G or only the main effect of E, respectively. However, this conclusion is incomplete since we find that there are situations where the originally non-identifiable parameters may still not be fully identified under the GEI assumption, even though the disease risk is affected by both genotype and environmental exposure.

To learn the identifiability of the parameters under the GEI assumption, it is sufficient to study the identifiability of θ, which is the only parameter in the parameterization ξ that is not identifiable in general. We first consider the simplest setting where G and E are both binary. In this case, the constraints imposed by the GEI assumption reduce to just one single equation as follows

$$\begin{aligned} 0 \; = & (\gamma_{100}\theta + \gamma_{000}(1-\theta))(\gamma_{111}\theta + \gamma_{011}(1-\theta)) - \\ & (\gamma_{101}\theta + \gamma_{001}(1-\theta))(\gamma_{110}\theta + \gamma_{010}(1-\theta)). \end{aligned}$$

It is easy to see that this equation is quadratic in θ. Therefore, it may produce a 'twin' solution comprised of the true disease prevalence as well as an erroneous value that also lies between 0 and 1. That is, there could exist two different values of ξ within the GEI subspace such that they only differ in the value of θ. Correspondingly, it is possible that two different values of ϕ both satisfy the GEI assumption and result in the same case-control sampling probabilities. For example, it can be easily verified that the following two values of ϕ_r lead to the same

case-control sampling probabilities:

$$\phi_r^{(1)}:\ \beta_G=\beta_E=\beta_{EG}=\log 2,\ \beta_0=\log 1.5,\ \kappa=(0.5,0.5)\,,\ \delta=(0.8,0.2)\,,$$
$$\phi_r^{(2)}:\ \beta_G=\beta_E=\beta_{EG}=\log 2,\ \beta_0=\log 5.5,\ \kappa=\left(\frac{13}{28},\frac{15}{28}\right),\ \delta=\left(\frac{52}{67},\frac{15}{67}\right).$$

Next, in more general settings where the environmental exposure takes more than two categories, the GEI-induced constraint can be viewed as multiple quadratic equations in θ, which can have at most two solutions. If there exist two values of θ both satisfying all of these equations, then these equations differ from one another only up to some constants, which is an extremely rare but theoretically possible situation.

Therefore, even under the GEI assumption, we may still not be able to fully identify all parameters from case-control data, even for some settings that are not included in the boundary situations defined in [3]. In theory, the originally non-identifiable parameters become 'almost' identified under the GEI assumption, though a 'twin' of the true value may be present as well.

3.5.2 Theoretical properties of the estimator

We now study the asymptotic distribution of the constrained maximum likelihood estimator of ξ. Within a sufficiently small neighborhood of the true value of ξ, the inequality constraints (3.5) and (3.6) are automatically satisfied. Thus, in order to study the asymptotic properties of the estimator, we can ignore those inequality constraints and treat the problem as estimating parameters arising from a partially identified model subject to only equality constraints, which has been studied in Chapter 2.

We still use the Lagrange multiplier method, and consider an auxiliary function $T(\xi,\lambda)=(1/n)\ell(\gamma)+\lambda^T\mathbf{g}(\xi)$. Since θ is now also an unknown parameter, we need to solve $\partial T(\gamma,\theta,\lambda)/\partial(\gamma,\theta,\lambda)=\mathbf{0}$ jointly with respect to γ, θ and λ to find

the maximum point of $T(\xi,\lambda)$. Thus, we get the following $5K+3$ equations

$$\begin{aligned} \frac{1}{n}\mathbf{s}(\gamma)+\mathbf{J}_\xi\lambda &= \mathbf{0}, && (3.11)\\ \mathbf{K}_\xi\lambda &= \mathbf{0}, && (3.12)\\ \mathbf{g}(\xi) &= \mathbf{0}, && (3.13) \end{aligned}$$

where $\mathbf{s}(\gamma)$ is again the gradient of the log-likelihood function, and $\mathbf{J}_\xi$ and $\mathbf{K}_\xi$ are the Jacobians of $\mathbf{g}(\xi)$ with respect to γ and θ, respectively, which are all given in Appendix A.

Suppose $(\hat\gamma,\hat\theta,\hat\lambda)$ is the solution of the equations (3.11) - (3.13). Then applying the theory developed in Chapter 2, as the sample size n goes to infinity with the case-to-control ratio n_1/n_0 fixed, the asymptotic distribution of $(\hat\gamma,\hat\theta,\hat\lambda)$ is

$$\sqrt{n}\begin{pmatrix}\hat\gamma-\gamma^*\\ \hat\theta-\theta^*\\ \hat\lambda\end{pmatrix}\xrightarrow{d}\mathcal{N}\left(\begin{pmatrix}\mathbf{0}\\ 0\\ \mathbf{0}\end{pmatrix},\begin{pmatrix}\mathbf{Q}_{11}&\mathbf{Q}_{12}&\mathbf{0}\\ \mathbf{Q}_{21}&\mathrm{Q}_{22}&\mathbf{0}\\ \mathbf{0}&\mathbf{0}&-\mathbf{Q}_{33}\end{pmatrix}\right), \qquad (3.14)$$

where

$$\begin{pmatrix}\mathbf{Q}_{11}&\mathbf{Q}_{12}&\mathbf{Q}_{13}\\ \mathbf{Q}_{21}&\mathrm{Q}_{22}&\mathbf{Q}_{23}\\ \mathbf{Q}_{31}&\mathbf{Q}_{32}&\mathbf{Q}_{33}\end{pmatrix}^{-1}=\begin{pmatrix}\mathbf{B}_{\gamma^*}&\mathbf{0}&-\mathbf{J}_{\xi^*}\\ \mathbf{0}&0&-\mathbf{K}_{\xi^*}\\ -\mathbf{J}_{\xi^*}^T&-\mathbf{K}_{\xi^*}^T&\mathbf{0}\end{pmatrix}.$$

Similarly as before, we refer to the matrix on the right hand side of the above equation as the 'constrained Fisher information' that accounts for the GEI assumption in the case that the disease prevalence is unknown.

We find from the equation (3.14) that the variance of the asymptotic distribution of $\hat\gamma$ is $\Sigma_\gamma=\mathbf{Q}_{11}$. Again, we apply the (3.2) to derive the asymptotic distributions of the constrained maximum likelihood estimators of the interaction effects, $\hat\beta_{EG}^{(k)}$ for $k=1,\ldots,K$, under the GEI assumption with unknown disease prevalence.

3.5.3 Numerical algorithm

In practice, the maximum likelihood estimate $\hat{\xi}$ has no closed form expression and still needs to be computed numerically. For this particular problem, we use the numerical algorithm proposed in Chapter 2, supplemented with a one-dimensional grid search if necessary, which is easy to implement and works well in practice.

First, we search globally for the constrained maximum likelihood estimate of ξ subject to the equality constraints, ignoring the inequality constraints for the moment. We call this a global search because each component of ξ is allowed to take any value. We set the initial value of γ to be $\gamma^{(U)}$ for the same reason as before. Without any additional information on disease prevalence, the initial value of θ is set to be 0.5, the midpoint of its plausible range. Therefore, the algorithm begins with an initial value $\xi^{(0)} = (\gamma^{(U)}, 0.5)$. We then iteratively update the value of ξ by

$$\begin{pmatrix} \gamma \\ \theta \\ \lambda \end{pmatrix} \leftarrow \begin{pmatrix} \gamma \\ \theta \\ \mathbf{0} \end{pmatrix} + c \begin{pmatrix} \mathbf{B}_\gamma & \mathbf{0} & -\mathbf{J}_\xi \\ \mathbf{0} & 0 & -\mathbf{K}_\xi \\ -\mathbf{J}_\xi^T & -\mathbf{K}_\xi^T & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{n}\mathbf{s}(\gamma) \\ 0 \\ \mathbf{g}(\xi) \end{pmatrix}. \tag{3.15}$$

Still, there is no need to record the value of λ for the next iteration. This process is repeated until convergence or termination. The termination criteria again include reaching the preset maximum number of iterations and running into numerical errors, both of which occur occasionally.

Next, we consider the inequality constraints. As discussed earlier in Section 3.4.2, we can tune the value of c to prevent values of γ from leaving their plausible range, and the inequality constraints (3.5) are naturally satisfied. In practice, however, the inequality constraint (3.6) may sometimes be violated. That is, the above algorithm may sometimes result in an estimate with $\hat{\theta} \notin [0,1]$, especially when the sample size is not very large. In that case, we perform a one dimensional grid search to approximate the optimal value of θ over the fixed interval $[0,1]$ that maximizes the log-likelihood $\ell(\gamma(\theta))$. For any fixed value of θ, we can apply Algorithm 1 to find the constrained maximum likelihood estimate of γ and obtain the maximum log-likelihood corresponding to that given value of θ. Moreover, this one-dimensional grid search is also performed in the rare situation when the global

search step fails to converge.

In summary, we use Algorithm 2, which combines the primary step of global search and the supplementary step of one-dimensional grid search, to obtain the maximum likelihood estimate $\hat{\xi}$. We then plug it into the equations (3.1) and (3.3), with $\Sigma_\gamma = \mathbf{Q}_{11}$, to get the constrained estimates for the interaction effects and their associated standard errors.

Algorithm 2 Find $\hat{\gamma}$ under the GEI assumption with unknown θ.

Set the initial value $\xi^{(0)} = (\gamma^{(0)}, 0.5)$
For $m = 1, 2, \ldots, M$
    Compute $\xi^{(m)}$ given $\xi^{(m-1)}$) using equation (3.15)
    If any error occurs:
        Set $m = M$ and break
    If $||\xi^{(m)} - \xi^{(m-1)}||_\infty < \varepsilon$
        Break
If $m < M$ and $\theta^{(m)} \in [0, 1]$
    Output $\hat{\xi} = \xi^{(m)}$
Else
    Set $w = -\infty$
    For $\theta = 0, 0.01, \ldots, 0.99, 1$
        Apply Algorithm 1 to find the estimate $\hat{\gamma}(\theta)$.
        If $\ell(\hat{\gamma}(\theta)) > w$
            Set $w = \ell(\hat{\gamma}(\theta))$
            Set $\hat{\xi} = (\hat{\gamma}(\theta), \theta)$
    Output $\hat{\xi}$

3.6 Extension: a reduced logistic model

In some cases, it may make practical sense to treat the categorical environmental exposure as ordinal rather than nominal. When the environmental exposure is ordinal, a reduced logistic regression model can be useful:

$$\text{logit}Pr(D = 1|E, G) = \beta_0 + \beta_G G + \beta_E E + \beta_{EG} EG.$$

It should be noted that this reduced model is just a special case of the saturated model, where $\beta_E^{(k)} = k\beta_E^{(1)}$ and $\beta_{EG}^{(k)} = k\beta_{EG}^{(1)}$, for $k = 2,\ldots,K$. Thus, assuming a reduced model introduces additional $2K-2$ constraints on γ,

$$\mathbf{h}(\gamma) = (h_{0,2}(\gamma),\ldots,h_{0,K}(\gamma),h_{1,2}(\gamma),\ldots,h_{1,K}(\gamma))^T = \mathbf{0},$$

where, for $j = 0,1$ and $k = 2,\ldots,K$,

$$h_{j,k}(\gamma) = \log\frac{\gamma_{1jk}}{\gamma_{0jk}} + (k-1)\log\frac{\gamma_{1j0}}{\gamma_{0j0}} - k\log\frac{\gamma_{1j1}}{\gamma_{0j1}}.$$

Combining these equality constraints with those imposed by the GEI assumption, we now have in total $3K-2$ constraint equations. The asymptotic theories and the numerical algorithms developed in Sections 3.4 and 3.5 are still applicable with some minor modifications. Suppose $\lambda_{\mathbf{h}}$ is the Lagrange multiplier associated with $\mathbf{h}(\gamma)$, and $\lambda_{\mathbf{g}}$ is the Lagrange multiplier associated with $\mathbf{g}(\xi)$. Let $\mathbf{H}_\gamma$ denote the Jacobian of $\mathbf{h}(\gamma)$ with respect to γ, the form of which is given in Appendix A. We discuss the extensions in three scenarios.

No GEI assumption

When we only assume a reduced logistic regression model in a general setting without the GEI assumption, the problem becomes finding the estimate of γ that maximizes $\ell(\gamma)$ subject to the equality constraint $\mathbf{h}(\gamma) = \mathbf{0}$ and the inequality constraints (3.5). This also fits into the framework developed in Section 3.4. Thus, the asymptotic distribution of $(\hat{\gamma},\hat{\lambda}_{\mathbf{h}})$ is

$$\sqrt{n}\begin{pmatrix}\hat{\gamma}-\gamma^*\\ \hat{\lambda}_{\mathbf{h}}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{R}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_{22}\end{pmatrix}\right), \tag{3.16}$$

where

$$\begin{pmatrix}\mathbf{R}_{11} & \mathbf{R}_{12}\\ \mathbf{R}_{21} & \mathbf{R}_{22}\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{B}_{\gamma^*} & -\mathrm{H}_{\gamma^*}\\ -\mathbf{H}_{\gamma^*}^T & 0\end{pmatrix}.$$

The right hand side of the above equation is the 'constrained Fisher information' that accounts for only assuming a reduced model. It should then be

used to modify the equation (3.10) in Algorithm 1 for updating the value of γ. This new algorithm can be used to numerically compute the estimate $\hat{\gamma}$.

GEI assumption with known θ

When we assume a reduced model under the GEI assumption with known disease prevalence, the problem is to find the estimate of γ that maximizes $\ell(\gamma)$ subject to the equality constraints $(\mathbf{h}(\gamma),\ \mathbf{g}(\gamma,\theta^*)) = \mathbf{0}$ and the inequality constraints (3.5). Following the work in Section 3.4, the asymptotic distribution of $(\hat{\gamma},\hat{\lambda}_{\mathbf{h}},\hat{\lambda}_{\mathbf{g}})$ is

$$\sqrt{n}\begin{pmatrix}\hat{\gamma}-\gamma^*\\ \hat{\lambda}_{\mathbf{h}}\\ \hat{\lambda}_{\mathbf{g}}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{P}'_{11} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & -\mathbf{P}'_{22} & -\mathbf{P}'_{23}\\ \mathbf{0} & -\mathbf{P}'_{32} & -\mathbf{P}'_{33}\end{pmatrix}\right), \quad (3.17)$$

where

$$\begin{pmatrix}\mathbf{P}'_{11} & \mathbf{P}'_{12} & \mathbf{P}'_{13}\\ \mathbf{P}'_{21} & \mathbf{P}'_{22} & \mathbf{P}'_{23}\\ \mathbf{P}'_{31} & \mathbf{P}'_{32} & \mathbf{P}'_{33}\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{B}_{\gamma^*} & -\mathbf{H}_{\gamma^*} & -\mathbf{J}_{\gamma^*;\theta^*}\\ -\mathbf{H}^T_{\gamma^*} & \mathbf{0} & \mathbf{0}\\ -\mathbf{J}^T_{\gamma^*;\theta^*} & \mathbf{0} & \mathbf{0}\end{pmatrix}.$$

The right hand side of the above equation is the 'constrained Fisher information' that accounts for assuming a reduced model and the GEI assumption with known disease prevalence. Again, it is used to modify the equation (3.10) in Algorithm 1 for updating the value of γ. This new algorithm is then used to numerically compute the estimate $\hat{\gamma}$ in practice for this problem.

GEI assumption with unknown θ

When we assume a reduced model under the GEI assumption with unknown disease prevalence, the problem becomes finding the estimate of ξ that maximizes $\ell(\gamma)$ subject to the equality constraints $(\mathbf{h}(\gamma),\ \mathbf{g}(\xi)) = \mathbf{0}$ and the inequality constraints (3.5) and (3.6). Following the work in Section 3.5, the

asymptotic distribution of $(\hat{\gamma}, \hat{\theta}, \hat{\lambda}_{\mathbf{h}}, \hat{\lambda}_{\mathbf{g}})$ is

$$\sqrt{n}\begin{pmatrix} \hat{\gamma}-\gamma^* \\ \hat{\theta}-\theta^* \\ \hat{\lambda}_{\mathbf{h}} \\ \hat{\lambda}_{\mathbf{g}} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ 0 \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{Q}'_{11} & \mathbf{Q}'_{12} & \mathbf{0} & \mathbf{0} \\ \mathbf{Q}'_{21} & \mathrm{Q}'_{22} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{Q}'_{33} & -\mathbf{Q}'_{34} \\ \mathbf{0} & \mathbf{0} & -\mathbf{Q}'_{43} & -\mathbf{Q}'_{44} \end{pmatrix}\right), \tag{3.18}$$

where

$$\begin{pmatrix} \mathbf{Q}'_{11} & \mathbf{Q}'_{12} & \mathbf{Q}'_{13} & \mathbf{Q}'_{14} \\ \mathbf{Q}'_{21} & \mathrm{Q}'_{22} & \mathbf{Q}'_{23} & \mathbf{Q}'_{24} \\ \mathbf{Q}'_{31} & \mathbf{Q}'_{32} & \mathbf{Q}'_{33} & \mathbf{Q}'_{34} \\ \mathbf{Q}'_{41} & \mathbf{Q}'_{42} & \mathbf{Q}'_{43} & \mathbf{Q}'_{44} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{B}_{\gamma^*} & \mathbf{0} & -\mathbf{H}_{\gamma^*} & -\mathbf{J}_{\xi^*} \\ \mathbf{0} & 0 & \mathbf{0} & -\mathbf{K}_{\xi^*} \\ -\mathbf{H}^T_{\gamma^*} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{J}^T_{\xi^*} & -\mathbf{K}^T_{\xi^*} & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Then the right hand side of the above equation is the 'constrained Fisher information' that accounts for assuming a reduced model and the GEI assumption with unknown disease prevalence. We then use it to modify the equation (3.15) in Algorithm 2 for updating the value of ξ. This new algorithm is used to numerically compute the estimate $\hat{\xi}$ in this situation.

3.7 Efficiency gain

In this section, we investigate the benefit of exploiting the GEI assumption in terms of estimation efficiency. Whereas related work [3] has relied on simulation studies to assess this benefit at only a limited number of parameter settings, using the theories in previous sections allow us to directly evaluate the efficiency gain at a very large number of values for the underlying parameters.

3.7.1 The special binary case

First, we consider the special case where both genotype and environmental exposure are binary. In this case, we can show that there is no efficiency gain by exploiting the GEI assumption if the disease prevalence is unknown. When the

GEI assumption is assumed and the disease prevalence is unknown, the variance of the asymptotic distribution of the estimator of γ is $\mathbf{Q}_{11}$ as given in the equation (3.14). Following the results given in the proof of Lemma 2 in Chapter 2, we are able to express $\mathbf{Q}_{11}$ by the block matrices of the corresponding constrained Fisher information:

$$\mathbf{Q}_{11} = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{J}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{J}^T\mathbf{B}^{-1} + \mathbf{W}$$

where subscripts are omitted for brevity and

$$\mathbf{W} = \mathbf{B}^{-1}\mathbf{J}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathrm{K}^T\left[\mathrm{K}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathrm{K}^T\right]^{-1}\mathrm{K}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{J}^T\mathbf{B}^{-1}.$$

Note that K reduces to a scalar in this special case. Thus, it can be easily verified that

$$\mathbf{W} = \mathbf{B}^{-1}\mathbf{J}(\mathbf{J}^T\mathbf{B}^{-1}\mathbf{J})^{-1}\mathbf{J}^T\mathbf{B}^{-1}.$$

It then follows that $\mathbf{Q}_{11} = \mathbf{B}^{-1}$, where $\mathbf{B}^{-1}$ is the variance of the asymptotic distribution of the unconstrained estimator of γ. This result suggests that the unconstrained and the constrained estimators of γ asymptotically follow the same distribution. Correspondingly, the unconstrained and the constrained estimators of $\beta_{EG}^{(k)}$, $k = 1, \ldots, K$, are asymptotically equivalent, as these parameters are connected with ξ only through γ. Thus, there is no efficiency gain by exploiting the GEI assumption when the disease prevalence is unknown in the special binary case. This matches the finding by Chen and Chen [4] that estimating the intercept term in the prospective relationship uses up the additional information inherent in the GEI assumption in the special binary case.

3.7.2 The saturated model

Next, we consider a saturated logistic model in the scenario where the environmental exposure has four levels. Some exploratory experiments show that the magnitude of efficiency gain is likely to be affected by the magnitude of the baseline risk β_0. Thus, we examine two different values of β_0 (logit0.003 *vs.* logit0.3) separately. For each value of β_0, we randomly generate 10000 sets of values for the remaining parameters. First, the marginal distributions of genotype and environmental exposure, κ and δ, are generated using flat Dirichlet distributions that are

uniform over the simplex, respectively. Secondly, all components of the main effects, β_G and β_E, and the interaction effects, β_{EG}, are generated separately using a standard normal distribution.

For each parameter setting, we first compute the variances of the asymptotic distributions of the estimators of γ resulting from the traditional method (TRAD), the method exploiting the GEI assumption with unknown disease prevalence (GEI-U), and the method exploiting the GEI assumption with known disease prevalence (GEI-K), which are $\mathbf{B}^{-1}$ defined in the equation (3.4), $\mathbf{Q}_{11}$ defined in the equation (3.9), and $\mathbf{P}_{11}$ defined in the equation (3.14), respectively. We then use the equation (3.2) to deduce the asymptotic variances of the corresponding estimators of the interaction effects $\beta_{EG}^{(k)}$, for $k = 1, \ldots, K$. We finally summarize results by boxplots in Figure 3.1, where the asymptotic variance ratios of the GEI-U estimators to the TRAD estimators are shown in left panels, and the asymptotic variance ratios of the GEI-K estimators to the GEI-U estimators are shown in right panels.

From Figure 3.1, we can see that exploiting the GEI assumption does lead to efficiency gain for the estimation of the interaction effects. This benefit is more likely to be substantial when the baseline risk is small. Moreover, knowing the disease prevalence can further lead to more efficiency gain, which also tends to be greater when the baseline risk is small.

3.7.3 The reduced model

Finally, we consider a reduced logistic model in three different scenarios, where the environmental exposure has three, four, and five levels, respectively. Again, we look at two different values of β_0 (logit0.003 *vs.* logit0.3) separately. For each value of β_0, we still randomly generate 10000 sets of values for the remaining parameters. Parameter values are generated in the same manner as before, except that β_E and β_{EG} are now two scalars.

For each parameter setting, we compute the asymptotic variances of the TRAD estimator, the GEI-U estimator and the GEI-K estimator of γ, which are $\mathbf{R}_{11}$ defined in the equation (3.16), $\mathbf{Q}'_{11}$ defined in the equation (3.18) and $\mathbf{P}'_{11}$ defined in the equation (3.17), respectively. The asymptotic variances of the corresponding three estimators of the interaction effects β_{EG} are then deduced. Figure 3.2 sum-

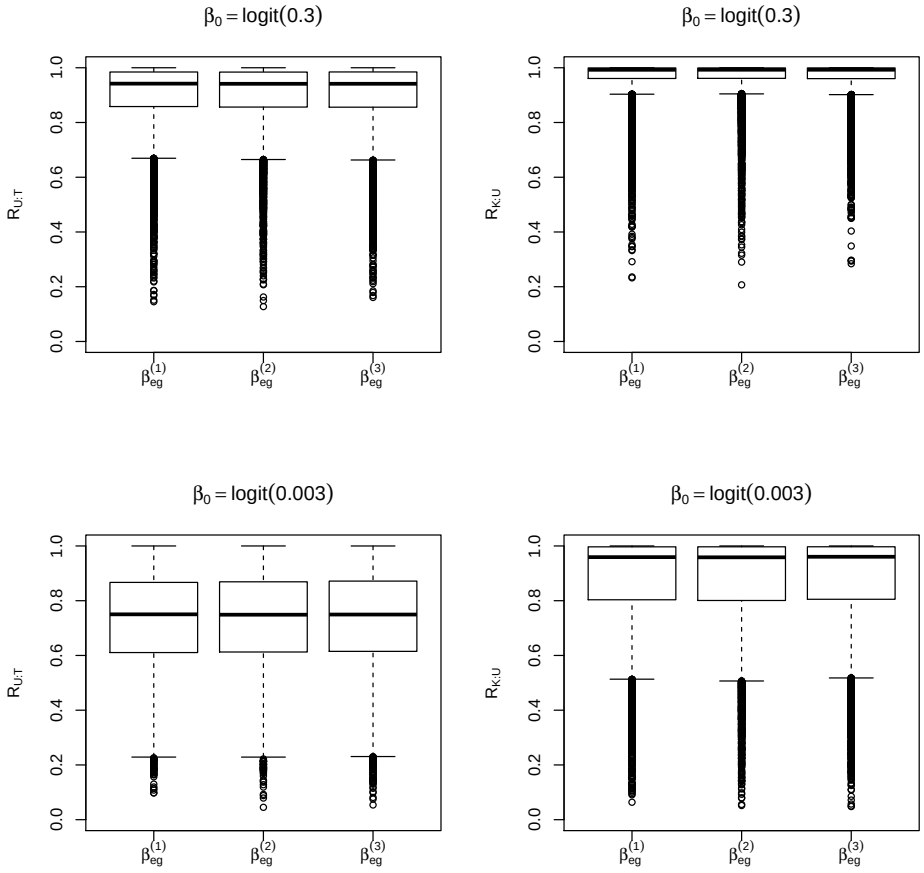

Figure 3.1: Comparison of the TRAD method, the GEI-U method, and the GEI-K method in terms of efficiency, with a saturated logistic model. The comparison is based on 10000 randomly generated parameter settings under the scenario where the environmental exposure has four levels. The asymptotic variance ratios of the GEI-U estimator to the TRAD estimator, $R_{U:T}$, are shown in left panels, and the asymptotic variance ratios of the GEI-K estimator to the GEI-U estimator, $R_{K:U}$, are shown in right panels.

marizes results, where the asymptotic variance ratios of the GEI-U estimator to the TRAD estimator are shown in left panels, and the asymptotic variance ratios of the

GEI-K estimator to the GEI-U estimator to the GEI-U estimator are shown in right panels.

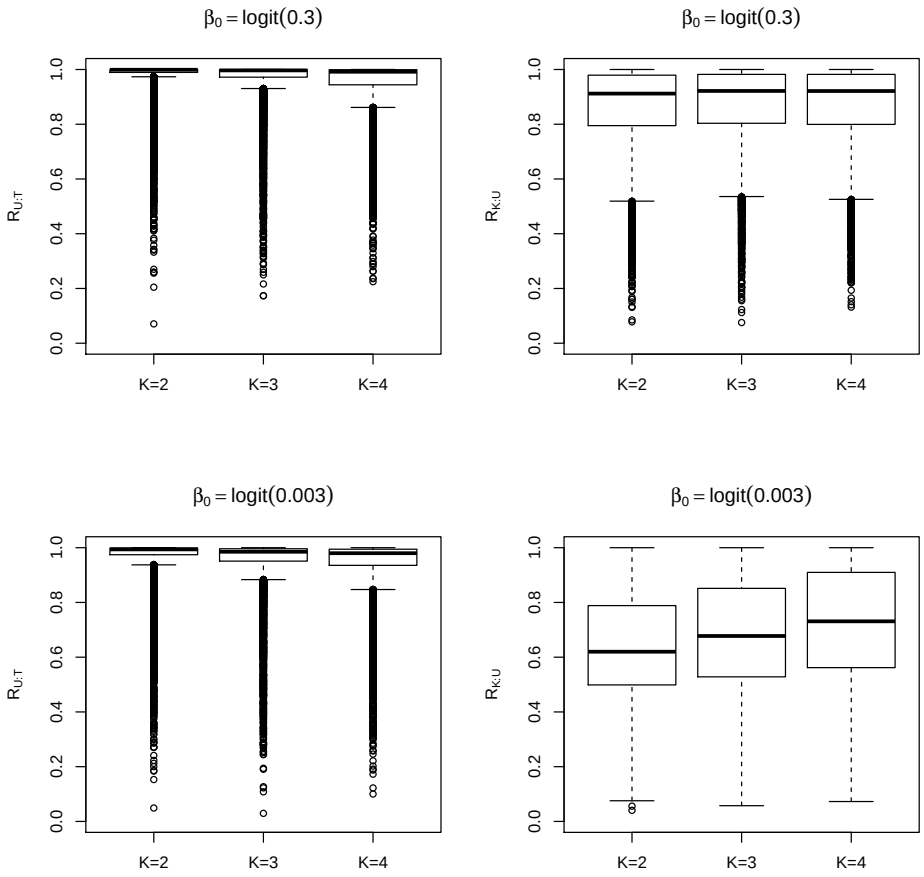

Figure 3.2: A similar figure to Figure 3.1, but shown with a reduced logistic model in scenarios where the environmental exposure has three ($K = 2$), four ($K = 3$), and five ($K = 4$) levels, respectively.

From Figure 3.2, we can see that exploiting the GEI assumption still leads to efficiency gain with a reduced logistic model, although the effect is less pronounced compared to the case with a saturated logistic model. Also, we find the trend that the benefit of exploiting the GEI assumption becomes greater when the environmental exposure has more levels. Interestingly, a reversed trend is observed in the

right bottom panel of Figure 3.2, which suggests that the benefit of knowing the disease prevalence seems to decrease as the levels of the environmental exposure increases.

3.8 Simulation studies

In this section, we conduct simulation studies to compare the performance of the traditional prospective logistic method (TRAD), the method exploiting the GEI assumption with unknown disease prevalence (GEI-U), and the method exploiting the GEI assumption with known disease (GEI-K) under different scenarios. We set the tuning parameters in Algorithm 1 and Algorithm 2 to be $c = 0.5$, $\varepsilon = \exp(-10\log 10)$, and $M = 1000$.

3.8.1 The special binary case

We first consider the special case where both genotype and environmental exposure are binary. We use the parameter setting: $\kappa = (0.95, 0.05)$, $\delta = (0.6, 0.4)$, $\beta_0 = \text{logit}0.005$, $\beta_G = \log 1.5$, $\beta_E = \log 1.2$, and $\beta_{EG} = \log 3$, so that the GEI assumption indeed holds. We consider three different sample sizes $n \in \{500, 1000, 2000\}$, with equal numbers of controls and cases. For each sample size, we apply the TRAD method, the GEI-U method, and the GEI-K method on 10000 randomly generated samples to obtain estimates and the corresponding 95% confidence intervals for the interaction effect β_{EG}.

We first present some summaries of the 10000 generated datasets. The GEI-K method successfully converges for all generated datasets. When the sample size is $500/1000/2000$, there are $80/0/0$ datasets containing at least one zero cell count, which are not used for comparison. The GEI-K method converges for all of the remaining $9920/10000/10000$ datasets. On the other hand, the GEI-U method directly finds the constrained maximum likelihood estimates of γ without the need of the one-dimensional grid search for $5137/5168/5323$ of those datasets. For the other $4783/4832/4677$ datasets, the one-dimensional grid search is performed to find estimates. Among these datasets, on-boundary estimates ($\hat{\theta} \in \{0, 1\}$) are found for $4779/4832/4677$ datasets.

We then summarize simulation results by three key indices in Table 3.1: the

bias and the mean squared error of the estimators, plus the coverage probabilities of the 95% confidence intervals. First, we can see that the GEI-U estimator is a little bit biased in practice because some estimates are computed differently by the one-dimensional grid search in order to ensure that estimates of the disease prevalence actually make practical sense. However, this empirical bias reduces as sample size increases. Secondly, the coverage probability of the GEI-U 95% confidence interval is even slightly greater than the nominal level. Thirdly, compared to the TRAD method, the GEI-U method yields slightly better mean squared errors. Finally, we find that the GEI-K method performs the best among the three methods. It leads to an unbiased estimator with substantially lower mean squared errors. It is interesting to see that knowing the disease prevalence has no effect on the efficiency of the estimators of the interaction effect without any additional assumption, but greatly improves the efficiency when the GEI assumption is made.

Table 3.1: Comparison between the TRAD method, the GEI-U method, and the GEI-K method in terms of the bias and the mean squared error (MSE) of the estimators of β_{EG}, and the coverage probability of the 95% confidence intervals in the special binary case.

n	Bias			MSE			Coverage		
	TRAD	GEI-U	GEI-K	TRAD	GEI-U	GEI-K	TRAD	GEI-U	GEI-K
500	0.061	0.257	0.034	0.651	0.584	0.201	0.959	0.959	0.957
1000	0.028	0.184	0.015	0.294	0.221	0.091	0.953	0.973	0.956
2000	0.019	0.127	0.010	0.137	0.097	0.045	0.954	0.976	0.949

We also compare the length of the 95% confidence intervals, as shown in Figure 3.3 and 3.4. When the sample size is small, the GEI-U 95% confidence intervals may sometimes be shorter than their TRAD counterparts. When the sample size gets larger, the 95% confidence intervals of these two estimators have about the same length. This observation matches our earlier result that these two estimators are asymptotically equivalent in the special binary case. Again, the GEI-K method outperforms the other two methods, as it results in much shorter 95% confidence intervals. More importantly, as we can see from Table 3.1, the coverage probability

of the GEI-K 95% confidence intervals is still maintained at the nominal level.

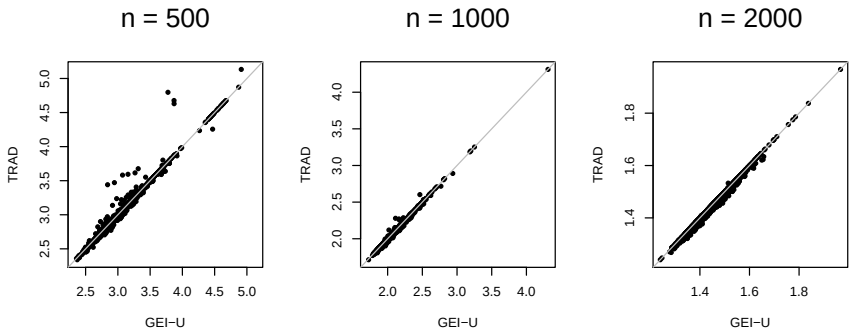

Figure 3.3: Comparison of the TRAD method and the GEI-U method in terms of the length of the 95% confidence interval in the special binary case. For better visualization, we only present results for a random sample of 1000 datasets from all 10000 simulated datasets. The grey line is the identity line.

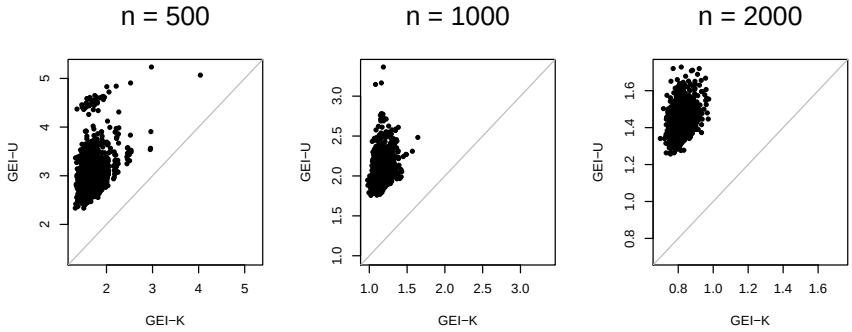

Figure 3.4: Comparison of GEI-U method and the GEI-K method in terms of the length of the 95% confidence interval in the special binary case. We still present results for a random sample of 1000 datasets from all 10000 simulated datasets for better visualization. The grey line is the identity line.

3.8.2 The saturated model

We now consider a saturated disease risk model concerning a binary genotype and an environmental exposure having four categories ($K = 3$) under the parameter setting: $\kappa = (0.9, 0.1)$, $\delta = (0.4, 0.3, 0.2, 0.1)$, $\beta_0 = \text{logit } 0.005$, $\beta_G = 0$, $\beta_E = (\log 1.1, \log 1.3, \log 1.5)$, and $\beta_{EG} = (\log 1.2, \log 1.6, \log 2)$. We still consider three different sample sizes $n \in \{500, 1000, 2000\}$, with equal numbers of controls and cases. For each sample size, we apply the TRAD method, the GEI-U method, and the GEI-K method on 10000 randomly generated samples to obtain the estimates and the corresponding 95% confidence intervals for the interaction effects, $\beta_{EG}^{(1)}$, $\beta_{EG}^{(2)}$, and $\beta_{EG}^{(3)}$.

We present some summaries about the 10000 generated datasets. When the sample size is $500/1000/2000$, there are $921/74/1$ datasets containing at least one zero cell count, which are not used for comparison. The GEI-K method converges for all of the remaining $9079/9926/9999$ datasets. On the other hand, the GEI-U method directly finds the constrained maximum likelihood estimates of γ without the need of the one dimensional grid search for $4541/5056/5107$ of those datasets. For the other $4528/4870/4892$ datasets, the one dimensional grid search is performed to find estimates. Among these datasets, non-boundary estimates are found for only $2/4/1$ datasets.

The three summary indices, the bias and the mean squared error of the estimators as well as the coverage probabilities of the 95% confidence intervals, are reported in Table 3.2. The comparisons between the GEI-U method and the TRAD method, and between the GEI-K method and the GEI-U method, in terms of the length of the 95% confidence intervals, are shown in Figure 3.5 and 3.6, respectively. First, we can see that the GEI-K method still performs the best among the three methods with respect to all these aspects. Secondly, the coverage probability of the GEI-U 95% confidence interval is slightly smaller than the nominal level. Thirdly, compared to the TRAD method, the GEI-U method yields comparable mean squared error even when the sample size is small, and better mean squared error as sample size increases. Finally, the GEI-U 95% confidence intervals are nearly always shorter than their TRAD method counterparts. For the parameter setting used in this simulation study, the asymptotic variance ratios for $\beta_{EG}^{(1)}$, $\beta_{EG}^{(2)}$

are 0.53, 0.65 and 0.88 between the GEI-U and the TRAD estimators, and 0.97, 0.69 and 0.43 between the GEI-U and the GEI-K estimators. Correspondingly, when the sample size is sufficiently large, the GEI-U 95% confidence intervals are 27%, 19%, and 6% shorter than their TRAD counterparts, and the GEI-K 95$ confidence intervals 2%, 17%, and 35% shorter than their GEI-U counterparts, for $\beta_{EG}^{(1)}$, $\beta_{EG}^{(2)}$, and $\beta_{EG}^{(3)}$, respectively. These theoretical results approximately match the empirical results shown in Figure 3.5 and 3.6.

Table 3.2: Comparison of the TRAD method, the GEI-U method, and the GEI-K method in terms of the bias and the mean squared error (MSE) of the estimators, and the coverage probability of the 95% confidence intervals in the scenarios with a saturated disease risk model.

	n	Bias			MSE			Coverage		
		TRAD	GEI-U	GEI-K	TRAD	GEI-U	GEI-K	TRAD	GEI-U	GEI-K
$\beta_{EG}^{(1)}$	500	0.020	0.016	0.007	0.642	0.610	0.321	0.950	0.938	0.958
	1000	0.000	0.002	-0.005	0.284	0.250	0.147	0.952	0.941	0.955
	2000	0.002	0.013	-0.001	0.138	0.106	0.069	0.949	0.943	0.952
$\beta_{EG}^{(2)}$	500	0.057	0.059	0.010	0.744	0.749	0.317	0.959	0.929	0.960
	1000	0.024	0.043	0.005	0.335	0.331	0.144	0.953	0.915	0.956
	2000	0.013	0.048	0.000	0.156	0.136	0.068	0.954	0.931	0.957
$\beta_{EG}^{(3)}$	500	0.007	0.022	0.000	0.904	0.990	0.405	0.971	0.927	0.963
	1000	0.056	0.091	-0.007	0.520	0.547	0.184	0.961	0.910	0.953
	2000	0.037	0.104	-0.001	0.248	0.241	0.085	0.953	0.928	0.951

3.8.3 The reduced model

Next, we consider a reduced disease risk model for a binary genotype and a four-category environmental exposure under the parameter setting: $\kappa = (0.9, 0.1)$, $\delta = (0.4, 0.3, 0.2, 0.1)$, $\beta_0 = \text{logit}0.005$, $\beta_G = 0$, $\beta_E = \log 1.3$, and $\beta_{EG} = \log 3$. Again, we consider three different sample sizes $n \in \{500, 1000, 2000\}$ and set the num-

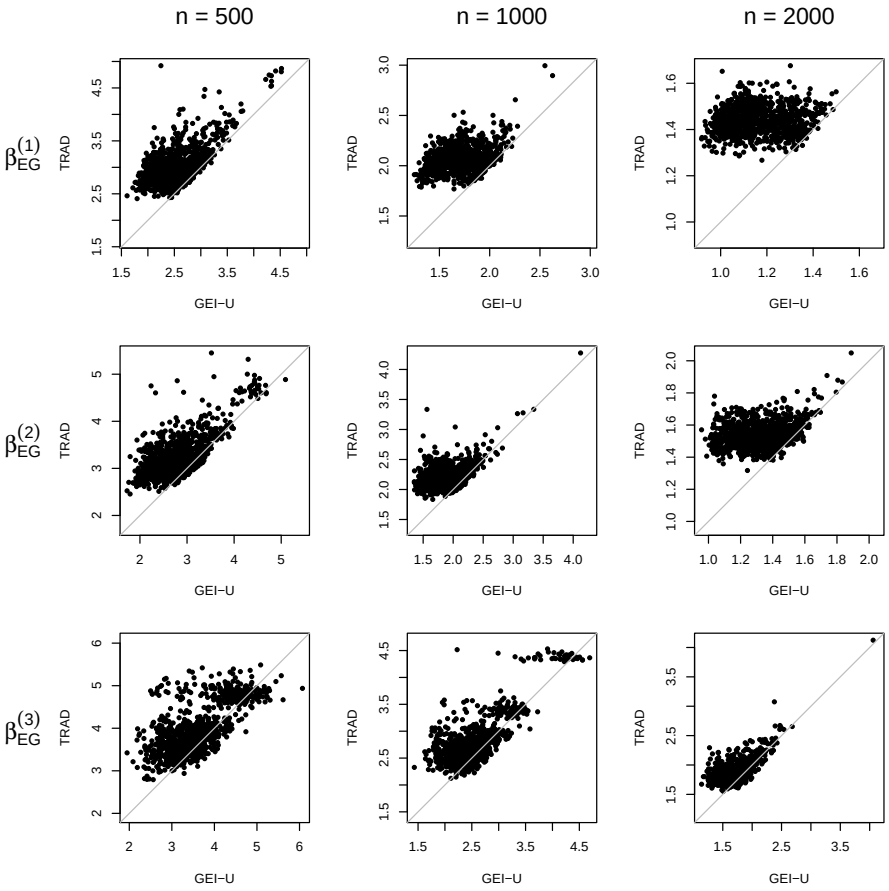

Figure 3.5: Comparison between the TRAD method and the GEI-U method in terms of the length of the 95% confidence interval for the scenario with a saturated disease risk model. Results are only presented for a random sample of 1000 datasets from all 10000 simulated datasets. The grey line is the identity line.

bers of controls and cases to be equal. We compare the performance of the TRAD method, the GEI-U method, and the GEI-K method still based on 10000 random samples. It should be noted that, with a reduced model, we can estimate the interaction effect β_{EG} even when the dataset contains zero cell counts. We find that the GEI-K method fails to converge for $1/0/0$ dataset using the current setting of

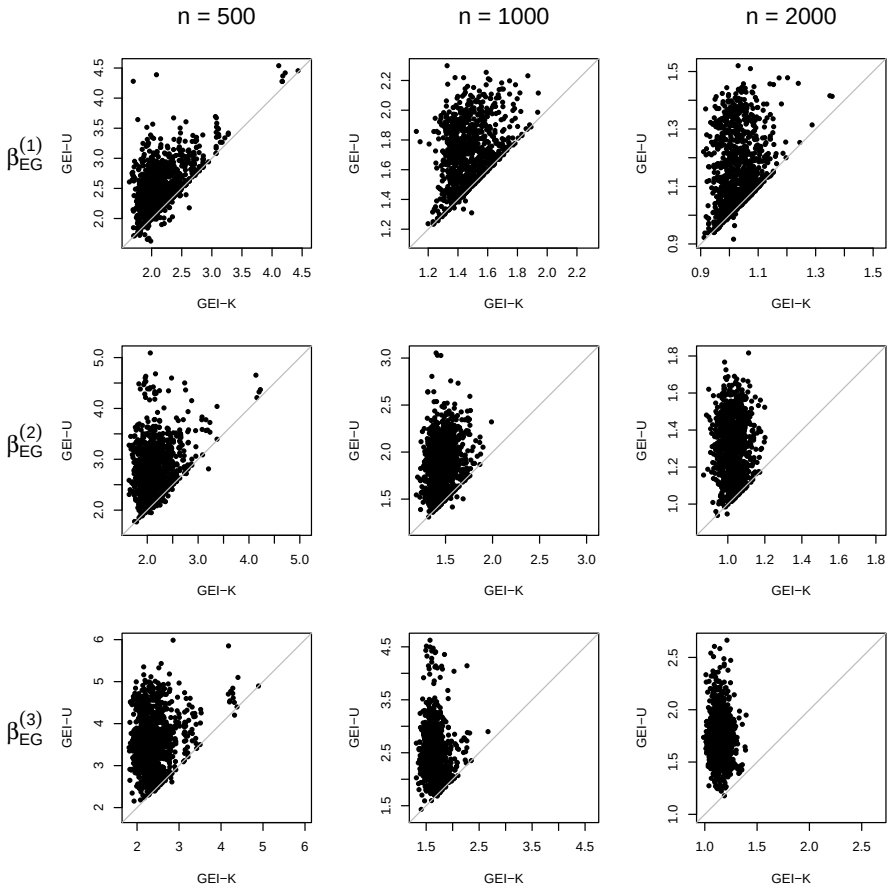

Figure 3.6: Comparison between the GEI-U method and the GEI-K method in terms of the length of the 95% confidence interval for the scenario with a saturated disease risk model. Results are only presented for a random sample of 1000 datasets from all 10000 simulated datasets. The grey line is the identity line.

tuning parameters. Among the remaining $9999/10000/10000$ generated datasets, the GEI-U method can directly find the constrained maximum likelihood estimates of γ without the need of the one dimensional grid search for $6941/7588/8366$ datasets. For the other $3058/2412/1634$ datasets, the one dimensional grid search is performed, and non-boundary estimates are found for only $1/0/0$ datasets.

The three summary indices, the bias and the mean squared error of the estimators as well as the coverage probabilities of the 95% confidence intervals, are reported in Table 3.3. The comparisons between the GEI-U method and the TRAD method, and between the GEI-K method and the GEI-U method, in terms of the length of the 95% confidence intervals, are shown in Figure 3.7 and Figure 3.8, respectively. First of all, as expected, the GEI-K method still performs the best among the three methods. Moreover, the GEI-U estimators are only slightly biased. The GEI-U method outperforms the TRAD method by achieving substantially lower mean squared errors even when the sample size is small. Also, the GEI-U 95% confidence intervals have coverage probabilities slightly above the nominal level, and are generally shorter than their TRAD counterparts. For this parameter setting, the asymptotic variance ratios for β_{EG} are 0.80 between the GEI-U and the TRAD estimators, and 0.36 between the GEI-K and the GEI-U estimators, corresponding to 10% and 40% reductions in the length of 95% confidence intervals, respectively. These theoretical results are observed empirically in Figure 3.7 and 3.8.

Table 3.3: Comparison of the TRAD method, the GEI-U method, and the GEI-K method in terms of the bias and the mean squared error (MSE) of the estimators, and the coverage probability of the 95% confidence intervals in the scenarios with a reduced disease risk model.

n	Bias			MSE			Coverage		
	TRAD	GEI-U	GEI-K	TRAD	GEI-U	GEI-K	TRAD	GEI-U	GEI-K
500	0.041	0.074	0.018	0.106	0.063	0.027	0.957	0.977	0.952
1000	0.026	0.041	0.010	0.049	0.031	0.013	0.947	0.970	0.953
2000	0.016	0.021	0.006	0.023	0.015	0.006	0.953	0.968	0.951

3.8.4 The violation of the GEI assumption

Finally, we test the performance of the proposed GEI-based methods when the GEI assumption does not hold. We consider a reduced disease risk model as an

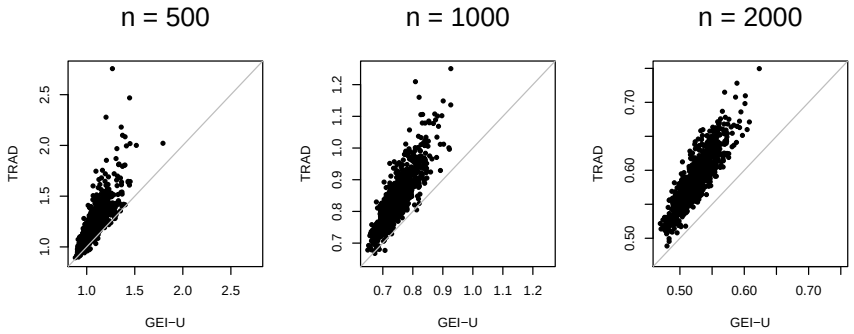

Figure 3.7: Comparison between the TRAD method and the GEI-U method in terms of the length of the 95% confidence interval for the scenario with a reduced disease risk model. Results are only presented for a random sample of 1000 datasets from all 10000 simulated datasets. The grey line is the identity line.

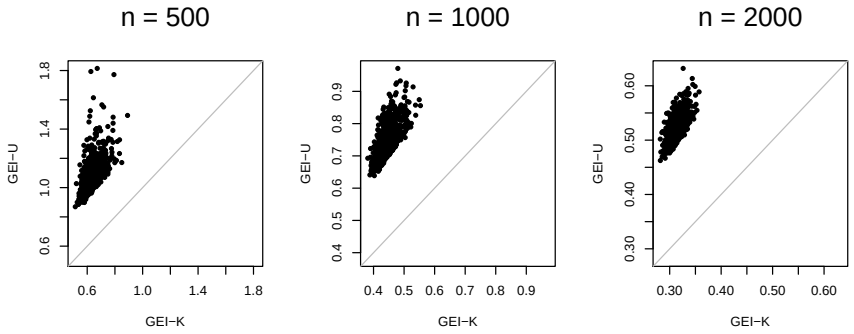

Figure 3.8: Comparison between the GEI-U method and the GEI-K method in terms of the length of the 95% confidence interval for the scenario with a saturated disease risk model. Results are only presented for a random sample of 1000 datasets from all 10000 simulated datasets. The grey line is the identity line.

example. The parameter setting for β is the same as the previous section, i.e., $(\beta_0, \beta_G, \beta_E, \beta_{EG}) = (\text{logit}0.005, 0, \log 1.3, \log 3)$. For the joint distribution of geno-

type and environmental exposure, we consider two scenarios:

$$(a): \quad \iota = (0.364, 0.269, 0.178, 0.089, 0.036, 0.031, 0.022, 0.011),$$
$$(b): \quad \iota = (0.38, 0.27, 0.17, 0.08, 0.02, 0.03, 0.03, 0.02),$$

where scenario (a) corresponds to a slight violation of the GEI assumption as the odds ratio between genotype and any pair of environmental exposure levels ranges from 0.8 to 1.3, and scenario (b) corresponds to a serious violation of the GEI assumption as the odds ratio between genotype and any pair of environmental exposure levels ranges from 0.2 to 4.8. For each scenario, we generate 10000 random samples of 500 controls and 500 cases. The GEI-K method converges for all datasets generated under both scenarios. Under the scenario (a), GEI-U estimates can be directly obtained by the global search step of Algorithm 2 for 7520 out of 10000 generated datasets, and on-boundary estimates are found by one-dimensional grid search for all the other 2480 datasets. Under the scenario (b), GEI-U estimates can be directly obtained by the global search step of Algorithm 2 for only 2646 out of 10000 generated datasets, and on-boundary estimates are found by one-dimensional search for all the other 7354 datasets. We see that data seem to be less likely to 'support' the GEI assumption when the assumption is seriously violated.

Table 3.4 summaries the simulation results for scenarios (a) and (b) with the bias and the mean squared error of the estimators as well as the coverage probabilities of the 95% confidence intervals. When the GEI assumption is slightly violated, the GEI-U method still performs pretty well, although slightly worse compared to the results shown in Table 3.3. It again leads to smaller mean squared errors compared to the TRAD method, and the coverage probability of the 95% confidence interval is also maintained at the nominal level. When the GEI assumption is seriously violated, however, the GEI-U estimator is greatly biased. Consequently, the proposed method produces much larger mean squared errors than the TRAD method. Moreover, the coverage probability of the 95% confidence interval is much lower than the nominal level. Finally, we can see that the GEI-K method is very sensitive to the violation of the GEI assumption and performs very poorly when the assumption is seriously violated, even much worse than the GEI-U

method.

Table 3.4: Comparison of the TRAD method, the GEI-U method, and the GEI-K method when the GEI assumption is violated, assuming a reduced disease risk model. Scenarios (a) and (b) correspond to the situations where the GEI assumption is slightly and seriously violated, respectively. Simulation results of 10000 random samples are summarized in terms of the bias and the mean squared error (MSE) of the estimators, and the coverage probability of the 95% confidence intervals.

	Bias			MSE			Coverage		
	TRAD	GEI-U	GEI-K	TRAD	GEI-U	GEI-K	TRAD	GEI-U	GEI-K
Scenario (a)	0.023	0.092	0.080	0.046	0.036	0.020	0.954	0.958	0.916
Scenario (b)	0.023	0.366	0.569	0.050	0.152	0.344	0.950	0.629	0.007

3.9 Data analysis

We now consider a real dataset for the application of the proposed GEI-U method. The dataset is taken from Garca-Closas et al. [7], who investigate the associations of polymorphisms in *NAT* and *GST* genes with bladder cancer risk and their interactions with cigarette smoking among subjects participating in the Spanish Bladder Cancer Study. We focus on the joint effect of genetic variation in *NAT2* and smoking habit on bladder cancer risk, and restrict the analysis to subjects who had complete information on *NAT2* genotype (rapid/intermediate vs. slow acetylator) and smoking habit (never, occasional, former, and current), resulting in a total of 1134 cases and 1130 controls. The observed cell frequencies are presented in Table 3.5.

We now apply the proposed GEI-U method to this dataset. The estimated interaction effects, accompanied by their 95% confidence intervals, are 0.529 with $(-0.305, 1.362)$ for the interaction between *NAT2* slow acetylator and occasional smokers, 0.628 with $(0.158, 1.098)$ for the interaction between *NAT2* slow acetylator and former smokers, and 0.403 with $(-0.092, 0.898)$ for the interaction between *NAT2* slow acetylator and current smokers. For comparison, the correspond-

Table 3.5: Data from a case-control study concerning the interaction of *NAT2* genotype and smoking habit on bladder cancer. *NAT2* genotype is coded as 0 for rapid/intermediate acetylator and 1 for slow acetylator. Smoking status is coded as 0 for never smoker, 1 for occasional smoker, 2 for former smoker, and 3 for current smoker.

	$G=0$				$G=1$			
	$E=0$	$E=1$	$E=2$	$E=3$	$E=0$	$E=1$	$E=2$	$E=3$
$D=0$	131	37	212	113	199	48	240	150
$D=1$	66	16	161	163	91	32	310	295

ing estimates and 95% confidence intervals from the TRAD method are 0.530 with $(-0.303, 1.362)$, 0.628 with $(0.160, 1.096)$, and 0.407 with $(-0.088, 0.902)$, respectively. Thus, we can see that making the GEI assumption is not very helpful for this particular dataset. Moreover, the estimated disease prevalence from the GEI-U method is 0.51, which is not very convincing since bladder cancer is a relatively rare disease. This might suggest that the GEI assumption is not satisfied for this problem. However, it should also be reminded that, even when the GEI assumption actually holds, exploiting the assumption may lead to no efficiency gain for some parameter settings, as can be seen from Figure 3.1.

3.10 Conclusion

In this chapter, we have studied the problem of estimating parameters arising from a logistic regression model for case-control data. This problem was previously studied by Chatterjee and Carroll [3] using the profile likelihood technique. We approach the problem in a different way by treating it as a constrained maximum likelihood estimation problem. It should be noted that both methods are based on the same retrospective likelihood for the case-control data, and thus should yield identical estimates. However, the present contribution is useful because, compared to the profile likelihood method in [3], our method is easier to implement. Moreover, as our method gives the explicit form for the asymptotic variance of the estimator, it can help with the planning of a case-control study.

By switching the position of genotype and environmental exposure, our method is readily applicable to the problem where the environmental exposure is binary but the genotype has three levels (recessive, co-dominant/incomplete-dominant, dominant). In general, with modern genotyping methods, we often do not know what trait is associated with a specific genetic marker/SNP within a gene, but have many of these markers within a gene (looks like categories but with no meaning to order). Then the challenge becomes trying to find a suitable model to assess joint impact of these markers with environmental exposure (typically forced to be represented as dichotomous) on disease. Our method may be naturally applied to approach this kind of problems.

Many aspects of exploiting the gene-environmental independence assumption in a case-control study have been well discussed by Chatterjee and Carroll [3]. Particularly, they briefly described the possibility of generalizing their method to address population stratification. When the genotype and environmental exposure are independent conditional on some stratum variables, the environmental exposure and the stratum variables can be combined as the new 'environmental variable'. This new variable is independent of the genetic factor, and thus the GEI-based method can still be applied. Our method can be generalized in a similar way to address population stratification as well.

Chapter 4

Bayesian Inference in Case-Control Studies

4.1 Introduction

In this chapter, we study methods for analyzing case-control data that exploit the GEI assumption in a Bayesian framework. This chapter is organized as follows. We first present another parameterization of the underlying model for case-control data, which has a reduced form directly incorporating the GEI assumption and also connects to the data structure closely. We then develop a Bayesian framework for analyzing case-control data under the GEI assumption, and conduct simulation studies to illustrate the performance of the proposed Bayesian method in situations where the GEI assumption indeed holds. Next, we generalize the proposed Bayesian method to allow uncertainty around the GEI assumption. We conduct more simulation studies to investigate the performance of the generalized Bayesian method in situations where the GEI assumption may or may not hold. Finally, we consider two real datasets for the application of the proposed methods, and give some concluding thoughts at the end of this chapter.

4.2 Another reparameterization

In Chapter 3, we presented a reparameterization of the underlying model that is closely connected to the case-control data structure. It was used to facilitate the analysis of case-control data from the frequentist perspective. With that parameterization, however, it is difficult to set an appropriate prior structure reflecting the GEI assumption, since the assumption is defined through some constraint equations. Thus, that parameterization is not suitable for Bayesian analysis. On the other hand, the original parameterization has a reduced form that directly incorporates the GEI assumption. Nonetheless, it may still not be ideal for Bayesian analysis in terms of computational efficiency, as it is not very closely connected to the actual data structure. Therefore, we need a new parameterization of the model that has a reduced form directly incorporating the GEI assumption and connects to the data structure as closely as possible.

Let θ still denote the disease prevalence. We use $\gamma_0 = (\gamma_{000}, \gamma_{001}, \ldots, \gamma_{01K})$ and $\gamma_1 = (\gamma_{100}, \ldots, \gamma_{11K})$ to represent the vectors of sampling probabilities underlying controls and cases, respectively. Further, we define $\rho = (\rho_1, \ldots, \rho_K)$ as the vector of gene-environment associations in the source population, where $\rho_k = (\iota_{1k}\iota_{00})/(\iota_{10}\iota_{0k})$ is the ratio of the odds of having genotype $G = 1$ in the $E = k$ subpopulation to the odds of having genotype $G = 1$ in the $E = 0$ subpopulation, for $k = 1, \ldots, K$. Finally, let $\gamma_{00} = (\gamma_{000}, \ldots, \gamma_{00K}, \gamma_{01+})$ denote another vector of probabilities in the control population, where $\gamma_{01+} = 1 - \sum_k \gamma_{00k}$ is the combined probability of having genotype $G = 1$.

Next, we are going to show that the new parameterization $\psi = (\theta, \rho, \gamma_{00}, \gamma_1)$ is a reparameterization of the original parameterization $\phi = (\iota, \beta_0, \beta_G, \beta_E, \beta_{EG})$. Since we have already shown in Chapter 3 that $\xi = (\theta, \gamma_0, \gamma_1)$ is a reparameterization of ϕ, it is sufficient to show the connection between ξ and ψ. On one hand, as $\iota_{jk} = (1-\theta)\gamma_{0jk} + \theta\gamma_{1jk}$, we can write ρ_k in terms of ξ through

$$\rho_k(\xi) = \frac{\{(1-\theta)\gamma_{000} + \theta\gamma_{100}\}\{(1-\theta)\gamma_{01k} + \theta\gamma_{11k}\}}{\{(1-\theta)\gamma_{00k} + \theta\gamma_{10k}\}\{(1-\theta)\gamma_{010} + \theta\gamma_{110}\}},$$

for $k = 1, \ldots, K$. The values of the remaining components of ψ, $(\theta, \gamma_{00}, \gamma_1)$, are readily available from ξ. Thus, the value of ψ can be uniquely determined given

a value of ξ. On the other hand, in order to determine the value of ξ given a value of ψ, we only need to find the value of $(\gamma_{010},\ldots,\gamma_{01K})$ as the values of the other components of ξ can be directly found in ψ. We first rearrange the above expression for ρ_k to get an equation with respect to γ_{010} and γ_{01k}:

$$-\rho_k \iota_{0k}\gamma_{010} + \iota_{00}\gamma_{01k} = \frac{\theta}{1-\theta}\left(-\iota_{00}\gamma_{11k} + \rho_k \iota_{0k}\gamma_{110}\right),$$

for $k = 1,\ldots,K$, where the value of $(\iota_{00},\ldots,\iota_{0K})$ is known given ψ. Combining these equations with the constraint that $\sum_{j,k}\gamma_{0jk} = 1$, we end up with in total $K+1$ equations for $K+1$ unknown parameters $(\gamma_{010},\ldots,\gamma_{01K})$. Furthermore, these $K+1$ equations can be rewritten in matrix form as

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ -\rho_1\iota_{01} & \iota_{00} & 0 & \cdots & 0 \\ -\rho_2\iota_{02} & 0 & \iota_{00} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\rho_K\iota_{0K} & 0 & 0 & \cdots & \iota_{00} \end{pmatrix} \begin{pmatrix} \gamma_{010} \\ \gamma_{011} \\ \gamma_{012} \\ \vdots \\ \gamma_{01K} \end{pmatrix} = \begin{pmatrix} 1-(\gamma_{000}+\cdots+\gamma_{00K}) \\ \frac{\theta}{1-\theta}\left(-\iota_{00}\gamma_{111}+\rho_1\iota_{01}\gamma_{110}\right) \\ \frac{\theta}{1-\theta}\left(-\iota_{00}\gamma_{112}+\rho_2\iota_{02}\gamma_{110}\right) \\ \vdots \\ \frac{\theta}{1-\theta}\left(-\iota_{00}\gamma_{11K}+\rho_K\iota_{0K}\gamma_{110}\right) \end{pmatrix}.$$

Then it is clear that this linear equation system always has a unique solution for $(\gamma_{010},\ldots,\gamma_{01K})$, due to the non-singularity of the matrix of coefficients. Thus, the value of ξ can be uniquely determined given a value of ψ. Therefore, ψ can be considered as a reparameterization of ξ, and in turn a reparameterization of ϕ.

If we let $\mathbb{P}^m$ denote the standard m-simplex, then the parameter space for ψ is $\Phi = \mathbb{P}^{2K+1}\times\mathbb{R}^{2K+2}$. With the reparameterization from ϕ to ψ, the original parameter space Φ is mapped to a new parameter space Ψ, which is only a proper subset of the space $\mathbb{P}^1\times\mathbb{R}_+^K\times\mathbb{P}^{K+1}\times\mathbb{P}^{2K+1}$ because some values of ψ in the larger space don't have correspondents in Φ. Moreover, for any given value of ψ in the larger space, it is easy to test for its membership in Ψ. Now, the retrospective likelihood for case-control data in terms of the new parameterization ψ is

$$L(\psi) = \left\{\prod_{j,k}\gamma_{1jk}^{n_{1jk}}\right\}\times\left\{\prod_k \gamma_{00k}^{n_{00k}}\right\}\times\left\{\prod_k\left(\gamma_{01k}(\psi)\right)^{n_{01k}}\right\}.$$

Finally, the GEI assumption can be easily represented by setting every component of ρ to be one.

4.3 Bayesian framework

We now develop a Bayesian framework for analyzing case-control data under the GEI assumption, based on the parameterization of ψ.

We begin by specifying a prior distribution for ψ over its parameter space Ψ. First, making the GEI assumption is equivalent to stating that $\psi \in \Psi_I$, where Ψ_I denotes the subset of Ψ corresponding to $\rho = \mathbf{1}$. Thus, the prior distribution has zero-density when $\psi \notin \Psi_I$. Next, we use a flat Dirichlet distribution of order $2K+2$ that is uniform over the simplex, $\mathrm{Dir}(1,\ldots,1)$, as the prior distribution for γ_1. Also, the same flat Dirichlet distribution would be used as the prior distribution for γ_0 if we were going to specify a prior distribution for ξ. That inspires a Dirichlet distribution of order $K+2$, $\mathrm{Dir}(1,\ldots,1,K+1)$, as the prior distribution for γ_{00}. Finally, we set an appropriate prior distribution for θ to reflect any knowledge we have on the disease prevalence, which may come from an external source. In this chapter, we simply use a standard uniform distribution as the non-informative prior distribution for θ, assuming no extra information is available. In summary, the prior density for ψ, denoted by $\tau(\psi)$, is

$$\tau(\psi) = C_\tau \cdot (\gamma_{01+})^K \cdot \mathbf{1}\{\psi \in \Psi_I\}.$$

where C_τ is a normalizing constant not depending on ψ. Finally, the posterior density of ψ, denoted by $p(\psi)$, is proportional to the product of the prior density $\tau(\psi)$ and the likelihood function $L(\psi)$, and thus can be expressed as:

$$p(\psi) = C_p \cdot (\gamma_{01+})^K \cdot \prod_k \gamma_{00k}^{n_{00k}} \cdot \prod_k (\gamma_{01k}(\psi))^{n_{01k}} \cdot \prod_{j,k} \gamma_{1jk}^{n_{1jk}} \cdot \mathbf{1}\{\psi \in \Psi_I\},$$

where C_p is another normalizing constant also not depending on ψ.

To compute the posterior distribution for Bayesian inference, we use the technique of importance sampling. The importance sampling method is often used as a variance reduction technique for approximating integrals. However, it does more than just that. It also provides an alternative way to simulate from complex distri-

butions (Robert and Casella [25]). Briefly, the method simulates a random sample from a proposal distribution along with corresponding importance weights to form a weighted sample numerically representing the original distribution. We choose the importance sampling technique over other Monte Carlo methods such as the rejection sampling method and the Metropolis-Hastings method for two reasons. First, it is computationally more efficient for our problem since a good proposal distribution can be established. Secondly, comparing to some other Markov Chain Monte Carlo methods like the Metropolis-Hastings method, the importance sampling technique comes with a relatively objective measure to evaluate the quality of the simulated Monte Carlo sample. We will illustrate these two points in the remainder of this section.

To ensure the performance of importance sampling, we need a good proposal distribution for ψ. First, we fix $\rho = \mathbf{1}$ to make sure that $\psi \in \Psi_I$. With the set-up of a conjugate Dirichlet prior distribution for the multinomial distribution, the posterior distribution for the case cell probabilities γ_1 can be easily obtained as a Dirichlet distribution with updated parameters, $\text{Dir}\left((n_{100}+1,\ldots,n_{11K}+1)\right)$, which in turn serves as a good proposal distribution for γ_1. Similarly, a reasonable proposal distribution for γ_{00} would be another Dirichlet distribution with updated parameters, $\text{Dir}\left((n_{000}+1,\ldots,n_{00K}+1,n_{01+}+K+1)\right)$. Finally, we propose θ from a distribution whose density function is $u(\theta)$. Therefore, the density function of the proposal distribution of ψ is

$$q(\psi) = C_q \cdot (\gamma_{01+})^{n_{01+}+K} \cdot u(\theta) \cdot \prod_{j,k} \gamma_{1jk}^{n_{1jk}} \cdot \prod_{k} \gamma_{00k}^{n_{00k}},$$

where C_q is the product of the normalizing constants of the two Dirichlet distributions, which does not depend on ψ. Note that $q(\psi)$ is positive over its support, which is a superset of Ψ_I. Thus, this proposal distribution of ψ is valid for the importance sampling.

We now discuss two candidate proposal distributions for θ. The first option is to use the prior distribution of θ, which is a standard uniform distribution in most cases without external information on the disease prevalence. This proposal distribution may not be very efficient because we know the posterior distribution of θ is more concentrated, especially when the sample size is very large. Nevertheless,

this simple proposal distribution works well for reasonable sample sizes in practice. The second option is to use the result of the frequentist approach as developed in Chapter 3. More specifically, we propose θ from a truncated normal distribution on the interval $[0,1]$, with mean being the constrained maximum likelihood estimate of θ and standard deviation being the corresponding standard error. This is a more aggressive proposal distribution that works more efficiently for most datasets, especially when the sample size is large. However, when the constrained maximum likelihood estimate of θ deviates far from the true value for some datasets due to random variation, this proposal distribution can be very inefficient. In practice, we can try both proposal distributions and it is very unlikely that both of them perform poorly.

For each value of ψ generated by the proposal distribution, the corresponding importance weight is $p(\psi)/(mq(\psi))$, where m is the total number of points generated. However, this exact weight can not be directly obtained since the normalizing constant in $p(\psi)$, C_p, is intractable. In practice, we compute the relative weight

$$w = \frac{\prod_k [\gamma_{01k}(\psi)]^{n_{01k}}}{[\gamma_{01+}]^{n_{01+}}} \cdot \mathbf{1}(\psi \in \Psi_I),$$

and renormalize all weights to have them sum to one. Having obtained a weighted sample numerically representing the posterior distribution of ψ, by transformation we can induce another weighted sample numerically representing the posterior distribution of ϕ. Either directly based on this weighted sample or based on an unweighted sample by further resampling, we can obtain the posterior mean and the corresponding equal-tailed 95% credible intervals for Bayesian inference.

Finally, we introduce the effective sample size (ESS) as a measure for assessing the quality of a weighted sample. It is defined in [25] as

$$ESS = \frac{(\sum_{i=1}^m w_i)^2}{\sum_{i=1}^m w_i^2}$$

Briefly, it gives the sample size of a simple random sample that would convey the same amount of information about the target posterior distribution as the weighted sample. Thus, we are going to keep track of the effective sample size of the weighted sample in our importance sampling algorithm and keep iterating until

it is greater than some threshold.

4.4 A simulation study

We now conduct a simulation study to investigate the performance of the proposed Bayesian method that exploits the GEI assumption (BGEI), and also compare it with the traditional prospective logistic method (TRAD).

We focus on the situation where both genotype and environmental exposure are binary, and consider parameter settings from a factorial experiment, where we fix the value of δ, β_G and β_E, and consider all combinations of assignments of two different values each for κ, β_0, and β_{EG}. In particular, we first set parameters $\delta = (0.6, 0.4)$, $\beta_G = \log 1.5$, and $\beta_E = \log 1.2$. We then consider $\kappa = (0.95, 0.05)$ for a rare genotype and $\kappa = (0.7, 0.3)$ for a common genotype, $\beta_0 = \text{logit } 0.005$ for a rare disease and $\beta_0 = \text{logit } 0.2$ for a common disease, and $\beta_{EG} = \log 1.1$ for a weak gene-environment interaction effect and $\beta_{EG} = \log 3$ for a strong gene-environment interaction effect. Thus, we end up with eight parameter settings used in the simulation study, as shown in Table 4.1.

Table 4.1: Parameter settings used in the simulation study in Section 3.8.

	κ	δ	β_0	β_G	β_E	β_{EG}
ϕ_1	$(0.95, 0.05)$	$(0.6, 0.4)$	logit 0.005	$\log 1.5$	$\log 1.2$	$\log 1.1$
ϕ_2	$(0.95, 0.05)$	$(0.6, 0.4)$	logit 0.005	$\log 1.5$	$\log 1.2$	$\log 3$
ϕ_3	$(0.95, 0.05)$	$(0.6, 0.4)$	logit 0.2	$\log 1.5$	$\log 1.2$	$\log 1.1$
ϕ_4	$(0.95, 0.05)$	$(0.6, 0.4)$	logit 0.2	$\log 1.5$	$\log 1.2$	$\log 3$
ϕ_5	$(0.7, 0.3)$	$(0.6, 0.4)$	logit 0.005	$\log 1.5$	$\log 1.2$	$\log 1.1$
ϕ_6	$(0.7, 0.3)$	$(0.6, 0.4)$	logit 0.005	$\log 1.5$	$\log 1.2$	$\log 3$
ϕ_7	$(0.7, 0.3)$	$(0.6, 0.4)$	logit 0.2	$\log 1.5$	$\log 1.2$	$\log 1.1$
ϕ_8	$(0.7, 0.3)$	$(0.6, 0.4)$	logit 0.2	$\log 1.5$	$\log 1.2$	$\log 3$

For each parameter setting, we consider three different choices for sample size, $n \in \{1000, 3000, 9000\}$, and set the numbers of cases and controls to be equal. For each sample size, we apply the TRAD method and the BGEI method on 2000

randomly generated samples to obtain point estimates and the corresponding 95% confidence/credible intervals for the interaction effects β_{EG}. For the BGEI method, the proposal distribution for θ is chosen to be the truncated normal distribution, as described in Section 4.3, and the threshold for the effective sample size is set to be 5000.

We first examine the computational efficiency of the importance sampling algorithm. Table 4.2 shows the 10-, 25-, 50-, 75-, and 90-percentiles of the importance sampling sample sizes actually used in our simulation study. In order to achieve effective sample size of at least 5000, the importance sampling algorithm requires sample size less than 20000 for 75% of the generated datasets in all settings considered. Moreover, except for one setting, the required importance sampling sample size is not greater than 30000 for 90% of the generated datasets. Therefore, we can see the importance sampling algorithm computes the posterior distribution very efficiently for most datasets.

Next, we summarize our simulation results with three key indices in Table 4.3: the bias and the mean square errors of the estimator, and the coverage probability of the 95% confidence/credible interval. First, we can see that, when the sample size is small, the BGEI estimator is biased for the parameter settings ϕ_2 and ϕ_6, both corresponding to the situation when the disease is rare and the gene-environment interaction effect is strong. For the remaining parameter settings, the BGEI estimator is nearly unbiased. Secondly, the BGEI method achieves smaller mean squared error than the TRAD method in practice, especially when the sample size is relatively small. Yet this advantage diminishes as sample size increases, which matches our conclusion in Chapter 3 that exploiting the GEI assumption does not improve asymptotic estimation efficiency in the special binary case. Lastly, the coverage probability of the BGEI equal-tailed 95% credible interval is maintained at the nominal level for a majority of parameter settings considered, but is slightly lower than the nominal level for a few parameter settings.

We also compare the BGEI method with the TRAD method in terms of the length of the 95% credible/confidence intervals, as shown in Figure 4.1. The first impression is that the length of the BGEI 95% credible interval is far more variable than that of the TRAD 95% confidence interval. The BGEI 95% credible intervals are shorter than their TRAD counterparts for most datasets, especially when the

sample size is small. But, when the gene-environment interaction effect is weak, there are a few datasets for which the BGEI 95% credible intervals are substantially longer than the TRAD 95% confidence intervals. However, as the sample size gets larger, the BGEI 95% credible interval and the TRAD 95% confidence interval tend to have similar length. This pattern can be seen more clearly from the last column in Figure 4.1.

Finally, we connect current results with earlier results in Section 3.8, where the parameter setting ϕ_2 and the sample size $n = 1000$ were also used in that simulation study, to compare the Bayesian GEI-based method (BGEI) and the frequentist GEI-based method (GEI-U as described in Chapter 3). Both estimators are biased and have similar mean squared errors. However, the two methods result in different interval estimates. By sacrificing a little bit in coverage probability, the BGEI 95% credible intervals are shorter than the GEI-U 95% confidence intervals.

Table 4.2: The computational efficiency of the proposed importance sampling algorithm summarized by percentiles of the importance sampling sample sizes for achieving effective sample size of at least 5000.

	n	Importance sampling sample sizes				
		10-%tile	25-%tile	50-%tile	75-%tile	90-%tile
ϕ_1	1000	8000	8000	9000	11000	16000
	3000	8000	8000	9000	11000	16000
	9000	8000	8000	9000	12000	20000
ϕ_2	1000	7000	7000	8000	10000	13000
	3000	6000	6000	7000	7000	9000
	9000	6000	6000	6000	7000	8000
ϕ_3	1000	8000	8000	9000	10000	14000
	3000	8000	8000	9000	11000	14000
	9000	8000	8000	9000	10000	15000
ϕ_4	1000	8000	9000	11000	14000	22000
	3000	7000	8000	9000	12000	19000
	9000	7000	7000	8000	9000	11000
ϕ_5	1000	9000	9000	10000	14000	24000
	3000	9000	9000	10000	15000	30000
	9000	9000	9000	12000	19000	58100
ϕ_6	1000	7000	8000	8000	10000	15000
	3000	7000	7000	8000	9000	14000
	9000	7000	7000	7000	8000	12000
ϕ_7	1000	9000	9000	10000	13000	21000
	3000	8000	9000	10000	13000	20000
	9000	8000	9000	10000	13000	22000
ϕ_8	1000	8000	9000	11000	14000	20000
	3000	9000	10000	11000	12000	15000
	9000	9000	10000	10000	11000	12000

Table 4.3: Comparison of the performance between the TRAD method and the proposed BGEI method for the estimation of the interaction effects, in terms of the bias and the mean squared error (MSE) of the estimators, and the coverage probability of the 95% credible/confidence intervals.

	n	Bias		MSE		Coverage	
		TRAD	BGEI	TRAD	BGEI	TRAD	BGEI
ϕ_1	1000	0.014	0.024	0.330	0.252	0.950	0.951
	3000	-0.005	0.015	0.097	0.080	0.952	0.946
	9000	-0.001	0.019	0.033	0.028	0.943	0.942
ϕ_2	1000	0.028	0.268	0.275	0.204	0.965	0.926
	3000	0.012	0.179	0.093	0.074	0.949	0.908
	9000	0.002	0.106	0.032	0.026	0.942	0.909
ϕ_3	1000	-0.013	-0.022	0.360	0.271	0.957	0.955
	3000	0.000	-0.008	0.114	0.092	0.945	0.954
	9000	-0.001	-0.004	0.035	0.028	0.959	0.957
ϕ_4	1000	0.062	-0.032	0.457	0.294	0.959	0.959
	3000	0.029	-0.023	0.136	0.105	0.951	0.953
	9000	0.008	-0.026	0.043	0.042	0.952	0.946
ϕ_5	1000	0.012	0.026	0.075	0.064	0.950	0.952
	3000	0.005	0.023	0.025	0.021	0.948	0.939
	9000	0.004	0.024	0.007	0.007	0.956	0.927
ϕ_6	1000	0.011	0.142	0.076	0.063	0.950	0.925
	3000	0.001	0.085	0.025	0.021	0.950	0.930
	9000	0.000	0.047	0.009	0.007	0.950	0.930
ϕ_7	1000	0.004	0.001	0.076	0.064	0.954	0.956
	3000	0.000	-0.001	0.024	0.020	0.961	0.960
	9000	0.001	-0.001	0.008	0.007	0.951	0.953
ϕ_8	1000	0.009	-0.044	0.085	0.077	0.950	0.947
	3000	0.007	-0.019	0.029	0.029	0.948	0.946
	9000	0.000	-0.009	0.009	0.009	0.952	0.952

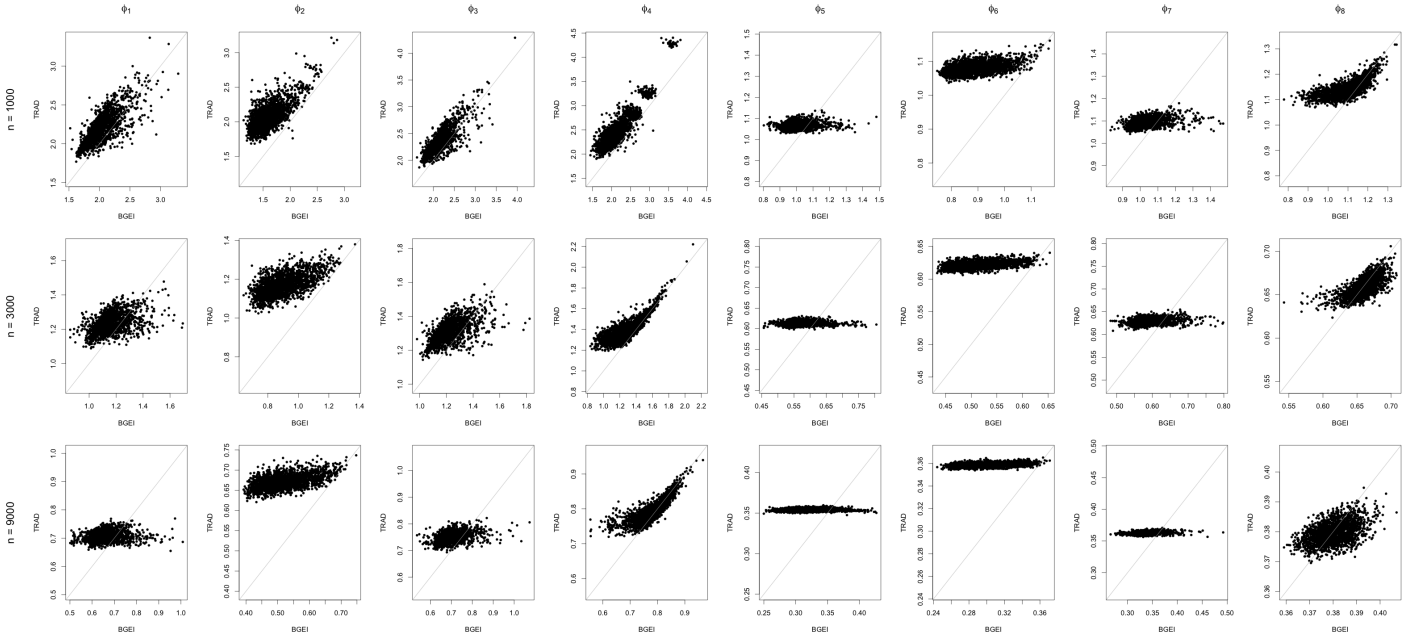

Figure 4.1: Comparison of traditional method (TRAD) and the proposed Bayesian GEI-based method (BGEI) in terms of the length of the 95% confidence/credible interval. The grey lines are the identity line.

4.5 Relaxation of the GEI assumption

Researchers sometimes have concerns about the validity of the GEI assumption. Thus, we discuss a variant of the proposed Bayesian method that relaxes the GEI assumption.

4.5.1 Two established methods

We first briefly review two established methods for relaxing the GEI assumption.

- **The empirical Bayes method**

  Mukherjee and Chatterjee [21] proposed a simple stochastic framework to trade off between bias and efficiency in a data-adaptive way. They showed that the magnitude of the uncertainty parameter can be estimated from the data itself. This estimate of the uncertainty parameter can then be used in an empirical Bayes fashion to obtain a shrinkage estimator that "shrinks" the maximum likelihood estimators of gene-environment interaction parameters under a general model to those obtained under the model that assumes the independence assumption.

  Particularly, in the simple set-up of a case-control study with a binary genetic factor and a binary environmental exposure, Mukherjee and Chatterjee [21] considered two commonly used estimators of the interaction effect β_{EG}, the one obtained from using all case-control data

  $$\hat{\beta}_{EG}^{(CC)} = \log \frac{n_{001} n_{010} n_{100} n_{111}}{n_{101} n_{110} n_{000} n_{011}},$$

  and the other obtained from using data of cases alone

  $$\hat{\beta}_{EG}^{(CO)} = \log \frac{n_{100} n_{111}}{n_{101} n_{110}}.$$

  The empirical Bayes estimator of β_{EG} was then proposed as the following weighted estimator

  $$\hat{\beta}_{EG}^{(EB)} = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}_{CC}^2} \hat{\beta}_{EG}^{(CC)} + \frac{\hat{\sigma}_{CC}^2}{\hat{\tau}^2 + \hat{\sigma}_{CC}^2} \hat{\beta}_{EG}^{(CO)},$$

where

$$\hat{\tau}^2 = \left(\log \frac{n_{000} n_{011}}{n_{001} n_{010}} \right)^2$$

is an estimate for the conceptual prior variability of the gene-environment association in the population of controls, which measures the uncertainty about the independence (among controls) assumption, and

$$\hat{\sigma}^2_{CC} = \sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{k=0}^{1} \frac{1}{n_{ijk}}$$

is the estimated variance of the case-control estimator $\hat{\beta}_{EG}^{(CC)}$. When the data provide evidence in favor of GEI assumption in the control population ($\hat{\tau}^2 \to 0$), we have $\hat{\beta}_{EG}^{(EB)} \to \hat{\beta}_{EG}^{(CC)}$. When the uncertainty regarding the assumption becomes stronger ($\hat{\tau}^2 \to \infty$), we have $\hat{\beta}_{EG}^{(EB)} \to \hat{\beta}_{EG}^{(CO)}$.

- **The full Bayesian method**

  Mukherjee et al. [22] proposed a proper full Bayesian approach for analyzing studies of gene-environment interaction, which provides a natural way to incorporate uncertainty around the assumption of GEI in the population of controls. In particular, they demonstrated their method in the simple set-up with a binary genotype and a binary environmental exposure. Let $\gamma_0 = (\gamma_{000}, \gamma_{001}, \gamma_{010}, \gamma_{011})$ and $\gamma_1 = (\gamma_{100}, \gamma_{101}, \gamma_{110}, \gamma_{111})$ denote the underlying sampling probabilities for controls and cases, respectively. The prior distributions on γ_i are assumed to be independent Dirichlet distributions, namely, $\gamma_i \sim \text{Dir}(\alpha_i)$ with $\alpha_i = (\alpha_{i00}, \alpha_{i01}, \alpha_{i10}, \alpha_{i11})$, $i = 0, 1$. Through a multinomial-Dirichlet conjugate analysis, the posterior distribution on γ_i can then be easily derived in closed form as a Dirichlet distribution with updated parameters, namely, $\gamma_i | \mathbf{n}_i \sim \text{Dir}(\mathbf{n}_i + \alpha_i)$, where $\mathbf{n}_i = (n_{i00}, n_{i01}, n_{i10}, n_{i11})$, $i = 0, 1$. Therefore, the posterior distribution of β_{EG} can be obtained using extremely inexpensive computation.

  Next, Mukherjee et al. [22] proposed to reflect prior belief on the assumption of GEI in the control population only through prior specification. More specifically, they defined the strength of the Dirichlet prior on the control

and case probability vector as $s_i = \sum_{j=0}^{1}\sum_{k=0}^{1}\alpha_{ijk}$ for $i = 0,1$. Different choices of s_0 and s_1 induce different variances on the logarithm of the gene-environment associations in controls and cases, respectively. Also, to make the corresponding prior distribution on β_{EG} roughly centered around zero, it was implicitly assumed in [22] that the two independent Dirichlet prior distributions are symmetric. For reflecting different degrees of belief on the assumption, the two parameter vectors, γ_0 and γ_1, are treated asymmetrically in the prior specification. The prior distribution on γ_1 is chosen to be fairly non-informative, say $\alpha_1 = (5,5,5,5)$. On the other hand, the Dirichlet prior on γ_0 can have varying strength s_0 to induce different prior variance around the assumption of independence (among controls).

However, the above two methods both focus on the other form of the GEI assumption that asserts GEI in the population of controls. This form of the assumption is not very natural, particularly as acquisition of genotype and environmental exposure are temporally antecedent to the disease. Therefore, we study the method for relaxing the more natural assumption of GEI within the source population in the next section.

4.5.2 A generalized Bayesian framework

In this section, we present a generalized Bayesian method that extends the Bayesian framework developed in Section 4.3 to further incorporate uncertainty around the GEI assumption.

Now, instead of asserting that $\rho = \mathbf{1}$, we model ρ *a priori* by a multivariate log-normal distribution

$$\rho \sim \ln\text{Norm}(\mathbf{0},\ \sigma_\rho^2\mathbf{I}),$$

where the common standard deviation σ_ρ can be specified to reflect different levels of uncertainty about the GEI assumption. Assuming the priors for other parameters are the same as before, the density of the prior distribution of ψ now becomes

$$\tilde{\tau}(\psi) = C_{\tilde{\tau}} \cdot (\gamma_{01+})^K \cdot \prod_{k=1}^{K} \frac{\mathrm{d}\left(\log(\rho_k)/\sigma_\rho\right)}{\rho_k} \cdot \mathbf{1}\{\psi \in \Psi\},$$

where $\mathrm{d}(\cdot)$ is the density function of the standard normal distribution. Note that the parameter space for ψ is now enlarged from Ψ_I to the entire Ψ. The posterior density of ψ now becomes

$$\tilde{p}(\psi) = C_{\tilde{p}} \cdot (\gamma_{01+})^K \cdot \prod_{k=1}^{K} \frac{\mathrm{d}\left(\log(\rho_k)/\sigma_\rho\right)}{\rho_k} \cdot \prod_k \gamma_{00k}^{n_{00k}} \cdot \prod_k (\gamma_{01k}(\psi))^{n_{01k}} \cdot \prod_{j,k} \gamma_{1jk}^{n_{1jk}} \cdot \mathbf{1}\{\psi \in \Psi\}.$$

Computationally, we need to extend the importance sampling algorithm to also have ρ proposed from its prior distribution. Moreover, since the GEI assumption may not hold, it would be better to use a standard uniform distribution as the proposal distribution for θ. Thus, we end up with a proposal distribution for ψ with the following density function

$$\tilde{q}(\psi) = C_{\tilde{q}} \cdot (\gamma_{01+})^{n_{01+}+K} \cdot \prod_{k=1}^{K} \frac{\mathrm{d}\left(\log(\rho_k)/\sigma_\rho\right)}{\rho_k} \cdot \prod_{j,k} \gamma_{1jk}^{n_{1jk}} \cdot \prod_k \gamma_{00k}^{n_{00k}}.$$

We can then compute the relative importance weight

$$\tilde{w} = \frac{\prod_k [\gamma_{01k}(\psi)]^{n_{01k}}}{[\gamma_{01+}]^{n_{01+}}} \cdot \mathbf{1}(\psi \in \Psi)$$

for each value of ψ generated from the proposal distribution and renormalize all weights to have them sum up to one. As before, we want the effective sample size of the weighted sample greater than some threshold to ensure a good approximation to the posterior distribution.

4.6 Another simulation study

In this section, we conduct another simulation study to compare the performance of the three methods discussed in Section 4.5 in situations where the GEI assumption may or may not hold.

We still focus on the set-up with a binary genotype and a binary environmental exposure. We assume that the marginal distributions for genotype and environmental exposure are $\kappa = (0.95, 0.05)$ and $\delta = (0.6, 0.4)$, respectively. The coefficients

in the logistic regression model are: $\beta_0 = \text{logit}\, 0.2$, $\beta_G = \log 1.5$, $\beta_E = \log 1.2$, and $\beta_{EG} = \log 3$. We now consider three different scenarios: (a) $\rho = 1$, (b) $\rho = 1.1$, and (c) $\rho = 3$, corresponding to no violation, a slight violation, and a serious violation of the GEI assumption, respectively. For each scenario, we generate 2000 datasets of 500 cases and 500 controls.

We analyze each generated case-control dataset using the empirical Bayes (EB) method, the full Bayesian (FB) method with three different choices of s_0, the Bayesian GEI (BGEI) method , and the generalized Bayesian (gBGEI) method with two different choices of σ_ρ. For three FB methods, we fix the case Dirichlet parameter α_1 as $(5,5,5,5)$ with $s_1 = 20$, and set the control Dirichlet parameter α_0 at $(5,5,5,5)$, $(20,20,20,20)$, and $(80,80,80,80)$, corresponding to values of s_0 at 20 (FB-20), 80 (FB-80), and 320 (FB-320), respectively. This setting was also used in [22]. For two gBGEI methods, the standard deviation σ_ρ is set at 0.1 (gBGEI-0.1) and 1.0 (gBGEI-1.0) for small and great uncertainty around the GEI assumption, respectively. Table 4.4 summarizes our simulation results in three indices: the bias and the mean squared error of the estimators as well as the coverage probabilities of the 95% credible/confidence intervals.

When the GEI assumption holds or is only slightly violated, our proposed Bayesian methods all result in unbiased or nearly unbiased estimators. The coverage probabilities of the 95% credible intervals are all maintained at the nominal level. Moreover, the BGEI method produces the smallest mean squared error, although the mean squared error of the gBGEI-0.1 estimator is very close. On the other hand, the EB estimator is slightly biased and the three FB estimators are all greatly biased. The coverage probabilities of the 95% credible/confidence intervals resulting from the EB method and the three FB methods are all lower than the nominal level. Lastly, the performance of the FB method gets worse (more bias, greater mean squared error, and lower coverage probability) with stronger prior belief on the assumption of GEI in the control population.

When the GEI assumption is seriously violated, both the BGEI method and the gBGEI-0.1 method perform poorly as they strongly rely on the validity of the GEI assumption. The gBGEI-1.0 method performs much better as it allows more uncertainty around the GEI assumption. On the other hand, the performance of the EB method is relatively invariant to the violation of the GEI assumption, as it

Table 4.4: Comparison between different methods in terms of the bias and the mean squared error (MSE) of the estimators of β_{EG}, and the coverage probability of the 95% credible/confidence intervals in situations where there is no violation ($\rho = 1$), a slight violation ($\rho = 1.1$), and a serious violation ($\rho = 3$) of the GEI assumption

		Bias	MSE	Coverage
$\rho = 1$	EB	-0.116	0.318	0.910
	FB-20	-0.357	0.299	0.935
	FB-80	-0.665	0.542	0.658
	FB-320	-0.785	0.695	0.340
	BGEI	-0.001	0.275	0.968
	gBGEI-0.1	-0.004	0.278	0.966
	gBGEI-1.0	-0.016	0.343	0.964
$\rho = 1.1$	EB	-0.072	0.276	0.932
	FB-20	-0.326	0.290	0.933
	FB-80	-0.575	0.435	0.750
	FB-320	-0.658	0.516	0.518
	BGEI	0.089	0.263	0.959
	gBGEI-0.1	0.080	0.267	0.960
	gBGEI-1.0	0.019	0.353	0.952
$\rho = 3$	EB	0.249	0.397	0.898
	FB-20	-0.203	0.226	0.966
	FB-80	-0.141	0.125	0.980
	FB-320	-0.034	0.086	0.983
	BGEI	0.796	0.791	0.489
	gBGEI-0.1	0.760	0.741	0.557
	gBGEI-1.0	0.074	0.360	0.961

automatically adjusts itself to rely less on the assumption. Finally, it is interesting to see that the FB methods work surprisingly well for this setting. The performance of the FB method gets better with stronger prior belief on the assumption of GEI

in the control population. However, the true gene-environment odds ratio in the control population is about 1.8, which does not justify the good performance of the FB method. Therefore, we feel that the performance of the FB method can be hard to predict.

4.7 Data analysis

We now examine two datasets to illustrate our methods. One dataset, taken from Hwang et al. [15], reveals an interaction effect of maternal smoking during pregnancy and a *Taq1* polymorphism at the transforming growth factor alpha (TGFα) locus on oral clefts. In this study, 113 infants with cleft palate were identified as cases ($D = 1$) and 281 infants without cleft palate were selected as controls ($D = 0$). All subjects were tested to ascertain whether they are carriers ($G = 1$) or non-carriers ($G = 0$) of any rare *Taq1* alleles. Smoking status during pregnancy, with $E = 1$ for smokers and $E = 0$ for non-smokers, was obtained by interview.

The other dataset considered concerns the joint effect of NAT2 genotype and cigarette smoking on bladder cancer, which was first published in Gu et al. [10] and was later reanalyzed by Gustafson and Burstyn [13]. The dataset consists of 502 cases ($D = 1$) and 512 controls ($D = 0$). The two categories of genotype are rapid ($G = 0$) and slow ($G = 1$) acetylator. The smoking status in this study is categorized as either heavy ($E = 1$) or never/light ($E = 0$). Both datasets are summarized in Table 4.5.

For each dataset, we apply ten estimators to quantify the gene-environment interaction, including the traditional case-control estimator (TRAD), the seven estimators considered in Section 4.6 (EB, BGEI, gBGEI-0.1, gBGEI-1.0, FB-20, FB-80, FB-320), and another generalized Bayesian estimator with $\sigma_\rho = 3.0$ (gBGEI-3.0). The results are summarized in Table 4.6. We again emphasize that the proposed Bayesian methods address the GEI assumption in the source population, but the empirical Bayes method and the full Bayesian method concern the GEI assumption in the control population. Given that the oral cleft is a rare disease, however, the two assumptions are approximately equivalent.

For both datasets, we see some similar patterns. First, the BGEI estimate is quite different from the TRAD estimate. Also, the gBGEI estimate is closer to

Table 4.5: Two datasets considered in Section 4.7. One was reported in [15] concerning the effect of TGF α genotype and maternal smoking during pregnancy on oral cleft. Genotype is coded as 0 = common Taq1 allele, 1 = rare Taq1 allele. Smoking status is coded as 0 = non-smoker, 1 = smoker. The other was used by [10] to investigate the effect of NAT2 genotype and smoking during pregnancy on oral cleft. Genotype is coded as 0 = rapid acetylator , 1 = slow acetylator. Smoking status is coded as 0 = never/light, 1 = heavy.

	Data from [15]				Data from [10]			
	$G=0$		$G=1$		$G=0$		$G=1$	
	$E=0$	$E=1$	$E=0$	$E=1$	$E=0$	$E=1$	$E=0$	$E=1$
$D=0$	167	69	34	11	172	58	230	52
$D=1$	60	32	12	9	106	83	156	157

the BGEI estimate when the uncertainty around the GEI assumption is weak, and closer to the TRAD estimate when the uncertainty around the GEI assumption is strong. Secondly, comparing to the BGEI estimate, the EB estimate is even more different from the TRAD estimate. Finally, the FB method tends to give the estimate that is most far away from the TRAD estimate when a very strong prior strength is assumed.

The two datasets might differ in term of the validity of the GEI assumption. For the first dataset, the $TGF\alpha$ genotype is probably independent of maternal smoking during pregnancy, as the EB estimate is very different from the TRAD estimate. Thus, comparing to the TRAD method, all methods that strongly rely on the assumption yield different point estimates and shorter interval estimates. For the second dataset, the $NAT2$ genotype might not be independent of maternal smoking during pregnancy, since the estimate resulting from the empirical Bayes method is less different from the TRAD estimate. Therefore, unlike the first dataset, all methods produce more similar results.

Table 4.6: The point and interval estimates of the gene-environment interaction β_{eg} for the two datasets considered in Section 4.7, obtained by ten different methods.

	Dataset of [15]		Dataset of [10]	
	Estimate	95% CI	Estimate	95% CI
TRAD	0.586	$(-0.628, 1.799)$	0.651	$(0.093, 1.208)$
EB	0.374	$(-0.628, 1.376)$	0.516	$(-0.077, 1.110)$
BGEI	0.474	$(-0.637, 1.598)$	0.555	$(0.032, 1.114)$
gBGEI-0.1	0.484	$(-0.607, 1.610)$	0.577	$(0.044, 1.115)$
gBGEI-1.0	0.552	$(-0.663, 1.730)$	0.650	$(0.095, 1.210)$
gBGEI-3.0	0.585	$(-0.612, 1.790)$	0.650	$(0.109, 1.198)$
FB-20	0.435	$(-0.640, 1.446)$	0.626	$(0.092, 1.172)$
FB-80	0.185	$(-0.820, 1.130)$	0.580	$(0.052, 1.079)$
FB-320	0.088	$(-0.805, 0.971)$	0.496	$(0.030, 0.970)$

4.8 Conclusion

In this chapter, we have developed a Bayesian framework for analyzing case-control data under the GEI assumption, and also generalized it to relax the GEI assumption. We have shown through two real dataset applications that the generalized Bayesian method does indeed serve as a compromise between the traditional case-control method and the proposed Bayesian method exploiting the exact GEI assumption.

We have seen in Chapter 3 that knowing the disease prevalence in addition to the GEI assumption can further improve estimation efficiency. This is also true for the proposed Bayesian methods. Moreover, the proposed Bayesian methods allow more flexibility to incorporate prior knowledge on the disease prevalence through an appropriate prior distribution. Therefore, rather than knowing the exact disease prevalence, any knowledge about the disease prevalence, like the rare disease assumption, may help improve estimation efficiency. The more concentrated prior distribution assumed on the disease prevalence, the better efficiency can be

achieved.

Finally, we have also briefly reviewed the empirical Bayes method and the full Bayesian method for relaxing the GEI assumption. However, these two methods both depend on the other form of the GEI assumption that asserts gene-environment independence in the population of controls. That version of GEI assumption is not as natural as the GEI assumption considered throughout this thesis. The assumption of GEI in the control population is often justified under the GEI assumption if the disease is rare. In that case, however, we feel that it would be more appropriate to directly incorporate the GEI assumption and the rare disease assumption into the analysis, which can be easily achieved by the proposed Bayesian methods. We expect that methods exploiting different versions of the GEI assumption would result in similar estimates if the prior distribution asserts the disease to be very rare. But some discrepancy could be observed if the prior distribution allows possibility of the disease prevalence being less extreme.

Chapter 5

Bayesian Inference in Case-Only Studies

5.1 Introduction

In this chapter, we study the analysis of case-only data under the GEI assumption and the rare disease assumption. The case-only design is a special variant of the standard case-control design, where only data on cases are collected. It exploits the GEI assumption for studying the gene-environment interaction, assuming that the disease is rare. This chapter is organized as follows. We first show how to adapt the Bayesian methods proposed in Chapter 4 for analyzing case-only data. We then investigate the prior distribution of the systematic bias of the traditional case-only method under different levels of disease prevalence, while the other assumption still holds true. Simulation studies are conducted to compare the performance of the proposed Bayesian methods with the traditional case-only method. Finally, we apply the proposed Bayesian methods on two real datasets.

5.2 Bayesian case-only methods

We first describe the application of the Bayesian methods developed in Chapter 4 on case-only data, since the case-only study can be viewed as a special type of case-control study where all control cell frequencies are zeros.

First, the likelihood in terms of ψ for case-only data is simply $L(\psi) = \prod_{j,k} \gamma_{1jk}^{n_{1jk}}$. Next, we still fix $\rho = \mathbf{1}$ to reflect the GEI assumption, and set the prior distributions for γ_{00} and γ_1 as $\mathrm{Dir}(1,\ldots,1,K+1)$ and $\mathrm{Dir}(1,\ldots,1)$, respectively. Moreover, to reflect the rare disease assumption, we acknowledge that a smooth prior on θ, such as a Beta distribution putting most of its mass very close to zero, might make more practical sense. However, for the sake of illustration, we initially assume that θ is uniformly distributed over $(0,\nu)$, where ν is a threshold slightly greater than zero that controls the prior belief concerning the rare disease assumption. We will use both specifications for our data analysis in Section 5.5. In summary, the prior density $\tau(\psi)$ is proportional to $(\gamma_{01+})^K$ when $\psi \in \Psi_I$ and $0 < \theta < \nu$, and zero otherwise. Finally, the posterior density $p(\psi)$ is:

$$p'(\psi) = C_{p'} \cdot (\gamma_{01+})^K \cdot \prod_{j,k} \gamma_{1jk}^{n_{1jk}} \cdot \mathbf{1}\{\psi \in \Psi_I\} \cdot \mathbf{1}\{0 < \theta < \nu\},$$

where $C_{p'}$ is a normalizing constant that does not depend on ψ. The interaction effects are estimated by their posterior means.

Computationally, we still use the technique of importance sampling to numerically represent the posterior distribution. The proposal distribution for γ_1 is $\mathrm{Dir}(n_{100}+1,\ldots,n_{11K}+1)$. The proposal distribution for γ_{00} is the same as its prior distribution, $\mathrm{Dir}(1,\ldots,1,K+1)$, since $n_{001} = \cdots = n_{01K} = 0$. Lastly, the proposal distribution for θ is also its prior distribution $\mathrm{U}(0,\nu)$. With these proposal distribution, it can be easily verified that the density function of the proposal distribution for ψ is some constant. Therefore, all proposed values of ψ, such that $\psi \in \Psi_I$ and $0 < \theta < \nu$, are equally weighted.

Next, we discuss the limiting case when we have an infinite amount of case-only data. In this case, the posterior distribution of case cell probabilities converges to a point mass at its true value, say γ_1^*. Correspondingly, the region of plausible values of ψ, where $\psi \in \Psi_I$, $0 < \theta < \nu$, and $\gamma_1 = \gamma_1^*$, may be greatly shrunk, leading to more concentrated distributions of β_{EG}. Thus, we can investigate the limiting posterior distributions (LPDs) of β_{EG}'s to gain some insight about how much *at most* can be learned from a non-identified model [11]. To obtain the LPDs, we modify the importance sampling algorithm by having γ_1 fixed at its true γ_1^* in the importance sampling algorithm.

Finally, the generalized Bayesian method developed in Section 4.5 is also applicable for the analysis of case-only data when GEI assumption may be relaxed. We still model ρ *a priori* by a multivariate log-normal distribution with a tuning parameter σ_ρ controlling the degree of uncertainty around the GEI assumption. For case-only data, the generalized Bayesian method also assigns equal weights to all proposed values of ψ such that $\psi \in \Psi$.

5.3 Bias of the traditional case-only method

In this section, we investigate the relationship between the disease prevalence and the performance of the traditional case-only method. Particularly, in the case of binary genetic and environmental factors, we have

$$\beta_{EG} = \zeta_1 - \zeta_0, \tag{5.1}$$

where ζ_1 and ζ_0 are the logarithm of gene-environment odds ratios in the populations of cases and controls, respectively. Since ζ_1 is the target parameter estimated by the traditional case-only method and β_{EG} is the parameter of real interest, their difference $\zeta_0 = \zeta_1 - \beta_{EG}$ measures the systematic bias of the traditional case-only method.

We examine the prior distributions of ζ_0 conditional on different values of θ. Given a value of θ, we randomly generate 100000 values of ψ by sampling γ_{00} and γ_1 from their prior distributions, respectively. We only keep points such that $\psi \in \Psi_I$ and $0 < \theta < \nu$, which are more than 95% for all values of θ considered. We then calculate the corresponding value of ζ_0 and obtain a numerical representation for its distribution conditional on the given value of θ. Figure 5.1 shows the mean, accompanied with the 2.5- and 97.5-percentiles, of the prior distribution of $\zeta_0|\theta$ for different values of θ. We can see from this figure that some substantial bias can emerge when the disease is more prevalent than 0.5%.

5.4 A simulation study

We now conduct a simulation study to compare the proposed Bayesian case-only method with the traditional case-only method. Still, we focus on the case of

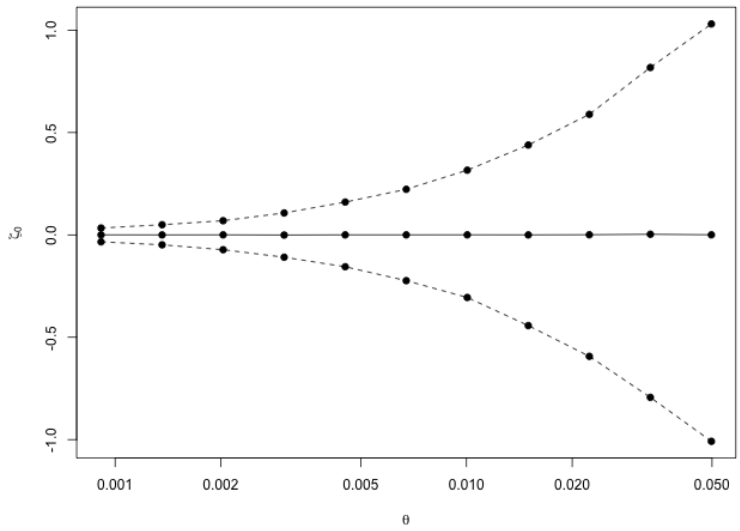

Figure 5.1: The mean and the 2.5- and 97.5-percentiles of the prior distributions of $\zeta_0|\theta$ for different values of disease prevalence θ. Note that a logarithmic scale is used for the x-axis.

a binary genotype and a binary environmental exposure. We assume that the marginal distributions for genotype and environmental exposure are $\kappa = (0.9, 0.1)$ and $\delta = (0.7, 0.3)$, respectively. We then set the main genetic and environmental effects to be $\beta_G = 0$ and $\beta_E = \log 2$. Finally, we consider combinations of two values for β_0 with two values for β_{EG}. We consider $\beta_0 = \text{logit } 0.001$ and $\beta_0 = \text{logit } 0.01$, corresponding to a very rare and a modestly rare disease, respectively. We also consider $\beta_{EG} = \log 2$ and $\beta_{EG} = \log 6$, corresponding to a modest and a strong gene-environment interaction effect, respectively. Thus, we end up with four parameter settings used in the simulation study, as shown in Table 5.1. The true disease prevalences corresponding to these parameter settings are 0.14%, 0.16%, 1.4%, and 1.6%, respectively.

We first examine the LPDs of β_{EG} under these four parameter settings. The four LPDs are obtained using $\nu = 0.01$. Table 5.2 presents the mean, the median, and the 95% equal tailed credible intervals of the prior distribution and the four LPDs of β_{EG}. We can see that the LPDs are indeed much more concentrated than

Table 5.1: Parameter settings used in the simulation study in Section 5.4.

	κ	δ	β_0	β_G	β_E	β_{EG}
ϕ_1	$(0.9, 0.1)$	$(0.7, 0.3)$	logit 0.001	0	$\log 2$	$\log 2$
ϕ_2	$(0.9, 0.1)$	$(0.7, 0.3)$	logit 0.001	0	$\log 2$	$\log 6$
ϕ_3	$(0.9, 0.1)$	$(0.7, 0.3)$	logit 0.01	0	$\log 2$	$\log 2$
ϕ_4	$(0.9, 0.1)$	$(0.7, 0.3)$	logit 0.01	0	$\log 2$	$\log 6$

the prior distribution. More importantly, each 95% LPD credible interval covers the corresponding true value of β_{EG}.

Table 5.2: Summary statistics about the prior distribution of β_{EG} and the LPDs for case-only data under four parameter settings.

	Mean	Median	2.5- and 97.5-%iles
Prior	-0.016	0.017	(-5.242, 5.053)
LPD under ϕ_1	0.700	0.693	(0.552, 0.898)
LPD under ϕ_2	1.805	1.788	(1.702, 1.979)
LPD under ϕ_3	0.682	0.676	(0.540, 0.858)
LPD under ϕ_4	1.718	1.703	(1.611, 1.935)

Next, we compare the performance of the Bayesian and the traditional case-only method in estimating β_{EG} with respect to the mean squared error (MSE) of the estimators, the coverage probability and the average length of the 95% confidence/credible intervals, based on 10000 datasets of 1000 cases. For the Bayesian method, we consider four different values, $\nu \in \{0.001, 0.003, 0.01, 0.03\}$, for the upper bound of the prior distribution on θ. Our simulation results are summarized in Table 5.3. On one hand, when we are very confident that the disease is really rare ($u = 0.001, 0.003$), the Bayesian interval estimates are very similar to the traditional interval estimates. On the other hand, when we are less certain that the disease is very rare, the Bayesian interval becomes wider to increase the coverage probability of the 95% credible interval. Moreover, we find that the Bayesian

method tends to be over-conservative. Thus, as we can see from the results for ϕ_4, even when the prior belief on the disease prevalence slightly deviates from the truth ($\nu = 0.01$ for $\theta = 0.016$), the coverage probability of the 95% credible intervals is still above the nominal level, and the average length of the intervals is comparable to that of the traditional method (0.809 versus 0.711).

Table 5.3: Comparison of the performance between the Bayesian and the traditional case-only method with respect to the mean squared error (MSE) of the estimators, the coverage probability and the average length of the 95% confidence/credible intervals, based on 10000 dataset of 1000 cases.

		ν	MSE	Coverage	Length
ϕ_1	Traditional	$0^\dagger$	0.034	0.951	0.717
	Bayesian	0.001	0.034	0.953	0.726
		0.003	0.034	0.958	0.747
		0.01	0.034	0.973	0.824
		0.03	0.036	0.994	1.090
ϕ_2	Traditional	$0^\dagger$	0.033	0.950	0.715
	Bayesian	0.001	0.033	0.953	0.721
		0.003	0.033	0.958	0.741
		0.01	0.033	0.972	0.811
		0.03	0.036	0.990	1.052
ϕ_3	Traditional	$0^\dagger$	0.035	0.943	0.719
	Bayesian	0.001	0.035	0.945	0.727
		0.003	0.035	0.953	0.749
		0.01	0.035	0.974	0.826
		0.03	0.035	0.995	1.094
ϕ_4	Traditional	$0^\dagger$	0.040	0.917	0.711
	Bayesian	0.001	0.041	0.922	0.718
		0.003	0.040	0.932	0.738
		0.01	0.039	0.961	0.809
		0.03	0.036	0.994	1.053

† The traditional method can be conceptually viewed as assuming $\nu = 0$.

Finally, we investigate the effect of sample size by considering four different choices, $n_1 \in \{500, 1000, 2000, 5000\}$, under the parameter setting ϕ_4. For the Bayesian method, we set $\nu = 0.03$. The coverage probabilities of the 95% confidence/credible intervals, based on 10000 datasets, are reported in Table 5.4. We observe that the two methods behave differently as the sample size increases. For the Bayesian method, the coverage probability is increasing and approaching 1 as the sample size gets larger. On the other hand, the traditional case-only performs worse as the sample size increases. Technically, as sample size goes to infinity, the 95% confidence interval of the traditional case-only method converges to a point mass at the true log odds ratio among cases, and thus its coverage probability converges to zero unless there is no systematic bias. In contrast, Gustafson [12] has shown that, in the partially identified context, the large-sample limit of frequentist coverage for Bayesian $(1-\alpha)$ credible intervals is one over a large subset of the parameter space, and zero over its complement, where large means having prior probability $1-\alpha$.

Table 5.4: The coverage probabilities of the Traditional and the Bayesian 95% confidence/credible intervals for different case-only sample sizes n_1, based on 10000 datasets.

	$n_1 = 500$	$n_1 = 1000$	$n_1 = 2000$	$n_1 = 5000$
Traditional	93.1%	91.2%	87.9%	77.5%
Bayesian	98.6%	99.6%	99.9%	100.0%

5.5 Data analysis

We consider two real datasets for the application of the proposed Bayesian methods. One example concerns a binary environmental exposure and the other concerns a four-category environmental exposure.

5.5.1 Analysis of colorectal cancer data

The first dataset is from a case-control study concerning the molecular epidemiology of colorectal cancer (MECC), which was previously used by Mukherjee et al.

[22]. We used the part of data concerning the interaction effect of $NAT2$ phenotype (slow vs. fast) and smoking status (never vs. ever). The observed cell counts are provided in Table 5.5.

Table 5.5: Case-control data concerning the interaction of $NAT2$ genotype (slow vs. fast) and smoking status (never vs. ever) on colorectal cancer.

	Colorectal Cancer (No)		Colorectal Cancer (Yes)	
	NAT2 (slow)	NAT2 (fast)	NAT2 (slow)	NAT2 (fast)
Smoke (never)	665	437	623	410
Smoke (ever)	584	285	475	277

The traditional case-only method gives an estimated interaction effect of -0.121 with a 95% confidence interval of $(-0.315, 0.073)$, and the traditional case-control method results in an estimated interaction effect of 0.177 with a 95% confidence interval of $(-0.092, 0.445)$. We can see that the traditional case-only method may be misleading in this data example, since its 95% confidence interval has little overlap with the case-control 95% confidence interval.

We now apply the proposed Bayesian case-only method to this dataset. We assume that the prevalence of colorectal cancer is less than 1%, i.e., the prior distribution on θ is $\theta \sim \text{Unif}(0, 0.01)$. The posterior mean of β_{EG} from the proposed Bayesian method is -0.118, with the 95% credible interval being $(-0.357, 0.124)$. We also consider a more realistic prior distribution for the prevalence of colorectal cancer, which is $\theta \sim \text{Beta}(5, 995)$. The result is very similar, with the estimated posterior mean being -0.121 and the 95% credible interval being $(-0.358, 0.117)$. Whatever prior is used for θ, we see that the proposed Bayesian method leads to more conservative credible intervals that are slightly more overlapped with the case-control confidence interval.

Finally, we apply the generalized Bayesian case-only method to this dataset. We again use $\text{Beta}(5, 995)$ as the prior distribution for disease prevalence. We first consider the setting $\sigma_\rho = 0.05$ to mimic the scenario that the GEI assumption might only be slightly violated. The estimated posterior mean is -0.124 and the 95% credible interval is $(-0.382, 0.132)$. Also, we consider the scenario where the GEI

assumption might be moderately violated with $\sigma_\rho = 0.5$. In this case, the estimated posterior mean is -0.123 and the 95% credible interval is $(-1.157, 0.900)$. As expected, the length of the 95% credible interval increases if we allow for the possibility of a stronger violation of the GEI assumption.

5.5.2 Analysis of ovarian cancer data

The second dataset is from a population-based case-control study of ovarian cancer reported by Modan et al. [20]. This study assessed whether the use of oral contraceptives and multiparity lower the risk of ovarian cancer in carriers of a BRCA1/2 mutation, as they do in non-carriers. Modan et al. [20] assumed that the the status of BRCA1/2 mutations and the reproductive risk factors are independent in controls, and did not provide a detailed breakdown. Thus, only data from the 832 cases were used. Our analysis focused on the interaction between the parity and the status of BRCA1/2 mutations, with data summarized in Table 5.6.

Table 5.6: Case-only data concerning the number of births and the status of BRCA1/2 mutations for 832 women with ovarian cancer.

	Number of Births			
	0	1-2	3-4	≥ 5
BRCA 1/2 NonCarriers	68	248	199	77
BRCA 1/2 Carriers	20	119	90	11

It was reported by Chatterjee and Carroll [3] that the estimate of the prevalence of ovarian cancer in the underlying population is about 0.00087. Thus, we set the threshold of the disease prevalence θ in the Bayesian method to be 0.2%. Making the 0-birth group the reference group, the estimated log odds ratios of the Bayesian method (and the associated 95% credible intervals) are 0.470 $(-0.062, 1.023)$ for the group of 1-2 births, 0.413 $(-0.141, 0.980)$ for the group of 3-4 births, and -0.702 $(-1.521, 0.079)$ for the group of more than 4 births. Similar to the first data analysis example, we also consider a smooth prior $\theta \sim \text{Beta}(1, 999)$. This yields estimated log odds ratios of 0.476 $(-0.068, 1.042)$ for the group of 1-2 births, 0.419 $(-0.143, 1.002)$ for the group of 3-4 births, and -0.698 $(-1.507, 0.093)$ for

the group of more than 4 births. Finally, we present the corresponding estimates and the associated 95% confidence intervals resulting from the traditional case-only method for comparison, which are 0.490 $(-0.055, 1.034)$, 0.430 $(-0.127, 0.988)$, and -0.722 $(-1.527, 0.083)$, respectively. We see that with a strong prior belief asserting that the disease is very rare, the proposed Bayesian method gives very similar results to those from the traditional case-only method. Thus, the proposed Bayesian method can be viewed in some sense as an extension of the traditional case-only method with more flexibility.

Again, we also apply the generalized Bayesian case-only method to this dataset. We use Beta$(1, 999)$ as the prior distribution for disease prevalence. When the GEI assumption might only be slightly violated with $\sigma_\rho = 0.05$, the estimated log odds ratios (and the associated 95% credible intervals) are 0.472 $(-0.078, 1.051)$ for the group of 1-2 births, 0.413 $(-0.152, 0.992)$ for the group of 3-4 births, and -0.703 $(-1.507, 0.094)$ for the group of more than 4 births. When we assume $\sigma_\rho = 0.5$ to allow a moderate violation of the GEI assumption, the estimated log odds ratios (and the associated 95% credible intervals) are 0.466 $(-0.647, 1.579)$ for the group of 1-2 births, 0.415 $(-0.729, 1.562)$ for the group of 3-4 births, and -0.698 $(-1.975, 0.528)$ for the group of more than 4 births. Again, permitting a moderate violation of the GEI assumption leads to substantially wider interval estimates.

5.6 Conclusion

We have investigated the performance of the traditional case-only method with different levels of disease prevalence, assuming that the genetic factor and the environmental exposure are independent in the target population. We have found some empirical evidence that, for most realistic parameter settings, the traditional case-only method works quite well for diseases less prevalent than 0.1%. When the disease prevalence is greater than 0.5%, however, we begin to see some substantial bias, and thus the traditional case-only method should be used with caution.

We have shown that the Bayesian framework for analyzing case-control data developed in Chapter 4 is readily applicable for analyzing case-only data. Particularly, the Bayesian case-only method allows the flexibility of incorporating dif-

ferent prior beliefs on disease prevalence. Compared to the traditional case-only method, the Bayesian case-only method leads to interval estimates that are much less likely to miss the true value if a correct, or even slightly incorrect, prior belief on disease prevalence is assumed. Moreover, we have seen from our simulation studies and data examples that, if we are confident that the disease is indeed very rare, the Bayesian case-only method gives almost identical results to the traditional case-only method. Thus, the proposed method can be viewed as a generalization of the traditional case-only method, which improves the quality of inference by taking expert opinion or previous knowledge into consideration.

Finally, we have the generalized Bayesian method which relaxes the GEI assumption and thus can be used to address situations when the GEI assumption might be violated. We have seen that allowing for the possibility of a moderate violation of the GEI assumption leads to a substantial increase in the length of the 95% credible interval. This reflects the fact that the case-only method is very sensitive to the violation of the GEI assumption.

Chapter 6

Future Work

In this thesis, we have studied the methods for analyzing case-control data that exploit the GEI assumption from both the frequentist and Bayesian perspective. Though presented in the context of gene-environment interaction studies, our methods can be applied to other case-control studies concerning two explanatory variables that are independent of each other. We have also developed the constrained maximum likelihood estimation for partially identified models, which may even have a broader application to solve other problems. In this chapter, we briefly describe a few possible directions for future research.

1. **Partially identified models with no transparent reparameterization**

   In Chapter 2, we assume that the partially identified model can be understood through a transparent re-parameterization that separates the identifiable parameters from non-identifiable parameters. Unfortunately, such a re-parameterization does not always exist. Gustafson et al. [14] gives two examples that do not admit a transparent re-parameterization. We suspect that the established theoretical results are still valid even when a transparent re-parameterization is not available. However, a rigorous proof is needed.

2. **Numerical algorithm directly incorporating inequality constraints**

   In Chapter 3, we use a two-step numerical algorithm for finding the GEI-constrained maximum likelihood estimate with unknown disease prevalence, where the second step of one-dimensional grid search is performed when the

estimate found by the first step breaks inequality constraints. That algorithm works reasonably well and suffices for our research problem, as our focus is not developing the best numerical algorithm. However, more elegant and efficient numerical algorithms that directly incorporate inequality constraints in each iteration may be desired.

3. **Reduced disease risk model for the Bayesian framework**

   In Chapter 4, we only propose the Bayesian method for a saturated disease risk model. One may also want to extend that for a reduced disease risk model. With a reduced model and the GEI assumption, it can be shown that $(\theta, \rho, \gamma_{000}, \gamma_{010}, \gamma_{011}, \gamma_{100}, \gamma_{101}, \gamma_{110}, \gamma_{111})$ serves as another parameterization of the model. A similar Bayesian framework can be developed for a reduced model based on this new parameterization.

4. **Continuous environmental exposure**

   In most gene-environment interaction studies, the environmental factors of interest are usually categorical variables with a few categories. Yet, there are situations where a continuous environmental exposure is of interest. In that case, we may convert a continuous variable to a categorical variable by grouping values, based on empirical quantiles for example. Hopefully, the loss in efficiency due to categorization can be minimized by having many, say ten, categories, and will be well compensated by the efficiency gain of exploiting the GEI assumption. More research can be conducted to look into this matter.

Bibliography

[1] J. Aitchison and S. D. Silvey. Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics*, 29(3):813 – 828, 1958. → pages 6, 8, 9, 11, 14, 18, 22, 35, 36

[2] P. S. Albert, D. Ratnasinghe, J. Tangrea, and S. Wacholder. Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology*, 154(8):687 – 693, 2001. → pages 3, 4

[3] N. Chatterjee and R. J. Carroll. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, 92(2):399 – 418, 2005. → pages 2, 31, 32, 38, 39, 45, 61, 62, 93

[4] H. Y. Chen and J. Chen. On information coded in gene-environment independence in case-control studies. *American Journal of Epidemiology*, 174(6):736 – 743, 2011. → pages 2, 28, 46

[5] A. R. Conn, N. I. M. Gould, and P. L. Toint. A globally convergent augmented lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis*, 28(2):545 – 572, 1991. → pages 36

[6] J. Dennis, S. Hawken, D. Krewski, N. Birkett, M. Gheorghe, J. Frei, G. McKeown-Eyssen, and J. Little. Bias in the case-only design applied to studies of gene-environment and gene-gene interaction: a systematic review and meta-analysis. *International Journal of Epidemiology*, 40:1329 – 1341, 2011. → pages 3

[7] M. Garca-Closas, N. Malats, D. Silverman, M. Dosemeci, M. Kogevinas, D. W. Hein, A. Tardon, C. Serra, A. Carrato, R. Garca-Closas, J. Lloreta, G. Castao-Vinyals, M. Yeager, R. Welch, S. Chanock, N. Chatterjee, S. Wacholder, C. Samanic, M. Tor, F. Fernndez, F. X. Real, and R. N. Nat2

slow acetylation and gstm1 null genotypes increase bladder cancer risk: results from the spanish bladder cancer study and meta-analyses. *Lancet*, 366(9468):649 – 659, 2005. → pages 60

[8] N. M. Gatto, U. B. Campbell, A. G. Rundle, and H. Ahsan. Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. *International Journal of Epidemiology*, 33:1014 – 1024, 2004. → pages 4

[9] P. E. Gill and E. Wong. *Mixed integer nonlinear programming*, volume 154, chapter Sequential quadratic programming method, pages 147 – 224. Springer, 2012. → pages 36

[10] J. Gu, D. Liang, Y. Wang, C. Lu, and X. Wu. Effects of n-acetyl transferase 1 and 2 polymorphisms on bladder cancer risk in caucasians. *Mutation Research*, 581:97 – 104, 2005. → pages 1, 81, 82, 83

[11] P. Gustafson. What are the limits of posterior distributions arising from nonidentified models, and why should we care? *Journal of the American Statistical Association*, 104(488):1682 – 1695, 2009. → pages 86

[12] P. Gustafson. On the behavior of bayesian credible intervals in partially identified models. *Electronic Journal of Statistics*, 6:2107 – 2124, 2012. → pages 91

[13] P. Gustafson and I. Burstyn. Bayesian inference of gene environment interaction from incomplete data: What happens when information on environment is disjoint from data on gene and disease? *Statistics in Medicine*, 30:877 – 889, 2011. → pages 81

[14] P. Gustafson, A. E. Gelfand, S. K. Sahu, W. O. Johnson, T. E. Hanson, L. Joseph, and J. Lee. On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science*, 20(2):111 – 140, 2005. → pages 6, 96

[15] S. J. Hwang, T. H. Beaty, S. R. Panny, N. A. Street, J. M. Joseph, S. Gordon, I. McIntosh, and C. A. Francomano. Association study of transforming growth factor alpha (tgf alpha) taqi polymorphismand oral clefts: indication of gene-environment interaction in a population-based sample of infants with birth defects. *American Journal of Epidemiology*, 141(7):629 – 636, 1995. → pages 81, 82, 83

[16] S. G. Johnson. The nlopt nonlinear-optimization package. URL http://ab-initio.mit.edu/nlopt. Accessed: 2016-05-31. → pages 36

[17] H. Luo, A. Bouchard-Côté, G. Cohen Freue, and P. Gustafson. The constrained maximum likelihood estimation for parameters arising from partially identified models. arXiv:1607.08826v1 [math.ST]. → pages iv

[18] H. Luo, I. Burstyn, and P. Gustafson. Gene-environment independence in case-control studies: Issues of parameteriza- tion and bayesian inference. *Statistics in Bioscience*, 7:460 – 475, 2015. → pages iv

[19] C. F. Manski. *Partial identification of probability distributions*. Springer, 2003. → pages 5

[20] B. Modan, P. Hartge, G. Hirsh-Yechezkel, A. Chetrit, F. Lubin, U. Beller, G. Ben-Baruch, A. Fishman, J. Menczer, J. P. Struewing, M. A. Tucker, and S. Wacholder. 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, 2001. → pages 93

[21] B. Mukherjee and N. Chatterjee. 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 (3):685 – 694, 2008. → pages 2, 3, 75

[22] B. Mukherjee, J. Ahn, S. B. Gruber, M. Ghosh, and N. Chatterjee. Case-control studies of gene-environment interaction: Bayesian design and analysis. *Biometrics*, 66:934 – 948, 2010. → pages 2, 3, 76, 77, 79, 92

[23] W. W. Piegorsch and C. R. Weinberg. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*, 13:153 – 162, 1994. → pages 2, 3

[24] R. L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403 – 411, 1979. → pages 1

[25] C. P. Robert and G. Casella. *Introducing Monte Carlo methods with R*. Springer, 2010. → pages 67, 68

[26] G. D. Smith and S. Ebrahim. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, 33:30 – 42, 2004. → pages 2, 4

[27] D. Spring. On the second derivative test for constrained local extrema. *The American Mathematical Monthly*, 92(9):631 – 643, 1985. → pages 15

[28] D. M. Umbach and C. R. Weinberg. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine*, 66:403 – 411, 1997. → pages 2

Appendix A

The Forms of Some Vectors and Matrices

We first define notations for two kinds of auxiliary matrices. Let $\mathbf{I}_s$ denote the identity matrix of size s, $\mathbf{E}_{s,t}$ denote the $s \times t$ all-ones matrix, and $\mathbf{O}_{s,t}$ denote the $s \times t$ zero matrix.

(a) The score function of the log-likelihood:

$$\mathbf{s}(\gamma) = \begin{pmatrix} \frac{n_{001}}{\gamma_{001}} \\ \vdots \\ \frac{n_{01K}}{\gamma_{01K}} \\ \frac{n_{101}}{\gamma_{101}} \\ \vdots \\ \frac{n_{11K}}{\gamma_{11K}} \end{pmatrix} - \begin{pmatrix} \frac{n_{000}}{\gamma_{000}} \cdot \mathbf{E}_{K,1} \\ \\ \frac{n_{100}}{\gamma_{100}} \cdot \mathbf{E}_{K,1} \\ \\ \end{pmatrix}.$$

(b) The unconstrained Fisher information:

$$\mathbf{B}_\gamma = \begin{pmatrix} \mathbf{B}_\gamma^{(0)} & \mathbf{O}_{K,K} \\ \mathbf{O}_{K,K} & \mathbf{B}_\gamma^{(1)} \end{pmatrix},$$

where, for $i = 0, 1$,

$$\mathbf{B}_{\gamma}^{(i)} = \frac{n_i}{n} \left\{ \begin{pmatrix} \frac{1}{\gamma_{i01}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\gamma_{i02}} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\gamma_{i1K}} \end{pmatrix} + \frac{1}{\gamma_{i00}} \cdot \mathbf{E}_{K,K} \right\}.$$

(c) The Jacobian of $\mathbf{g}(\xi)$ with respect to γ:

$$\mathbf{J}_{\xi} = \begin{pmatrix} (1-\theta) \cdot \mathbf{J}^{\dagger} \\ \theta \cdot \mathbf{J}^{\dagger} \end{pmatrix},$$

where

$$\mathbf{J}^{\dagger} = \begin{pmatrix} -\iota_{10} \cdot \mathbf{I}_K \\ -(\iota_{01}, \ldots, \iota_{0K}) \\ \iota_{00} \cdot \mathbf{I}_K \end{pmatrix} - \mathbf{E}_{2K+1,1} \times (\iota_{11}, \ldots, \iota_{1K}).$$

(d) The Jacobian of $\mathbf{g}(\xi)$ with respect to θ:

$$\mathbf{K}_{\xi} = \left(\frac{\partial g_1}{\partial \theta}, \ldots, \frac{\partial g_K}{\partial \theta} \right),$$

where, for $k = 1, \ldots, K$,

$$\frac{\partial g_k}{\partial \theta} = (\gamma_{100} - \gamma_{000})\iota_{1k} + (\gamma_{11k} - \gamma_{01k})\iota_{00} - (\gamma_{10k} - \gamma_{00k})\iota_{10} - (\gamma_{110} - \gamma_{010})\iota_{0k}.$$

(e) The Jacobian of $\mathbf{h}(\gamma)$, the constraints imposed by assuming a reduced model,

with respect to γ:

$$\mathbf{H}_\gamma = \begin{pmatrix} \frac{1}{\gamma_{000}}\mathbf{E}_{2K+1,1} \times (1,\ldots,K-1) & \mathbf{O}_{2K+1,K-1} \\ -\frac{1}{\gamma_{100}}\mathbf{E}_{2K+1,1} \times (1,\ldots,K-1) & \mathbf{O}_{2K+1,K-1} \end{pmatrix} + \begin{pmatrix} \mathbf{X}_0 & \mathbf{O}_{K,K-1} \\ \mathbf{O}_{K+1,K-1} & \mathbf{Y}_0 \\ \mathbf{X}_1 & \mathbf{O}_{K,K-1} \\ \mathbf{O}_{K+1,K-1} & \mathbf{Y}_1 \end{pmatrix},$$

where, for $i = 0, 1$,

$$\mathbf{X}_i = (-1)^i \begin{pmatrix} \frac{2}{\gamma_{i01}} & \frac{3}{\gamma_{i01}} & \cdots & \frac{K}{\gamma_{i01}} \\ -\frac{1}{\gamma_{i02}} & 0 & \cdots & 0 \\ 0 & -\frac{1}{\gamma_{i03}} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & -\frac{1}{\gamma_{i0K}} \end{pmatrix},$$

and

$$\mathbf{Y}_i = (-1)^i \begin{pmatrix} -\frac{1}{\gamma_{i10}} & -\frac{2}{\gamma_{i10}} & \cdots & -\frac{K-1}{\gamma_{i10}} \\ \frac{2}{\gamma_{i11}} & \frac{3}{\gamma_{i11}} & \cdots & \frac{K}{\gamma_{i11}} \\ -\frac{1}{\gamma_{i12}} & 0 & \cdots & 0 \\ 0 & -\frac{1}{\gamma_{i13}} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & -\frac{1}{\gamma_{i1K}} \end{pmatrix}.$$