BAYESIAN METHODS IN CASE-CONTROL STUDIES WITH APPLICATIONS IN GENETIC EPIDEMIOLOGY

By

LI ZHANG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2006

To my husband, Xin, and my parents.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to both of my advisors (Professor Malay Ghosh and Professor Bhramar Mukherjee) for their immense help at every stage of my research. I remain grateful for their constant encouragement, and mental support throughout the hardship of my graduate study at the University of Florida. Without their patience, guidance and encouragement, none of this work would have been possible. As mentors, their wisdom, kindness and enthusiasm benefitted me greatly in both my research work and life. Their valuable insights and ideas directly and significantly contributed to the work in this dissertation.

I would also like to give special thanks to Professor Rongling Wu for many fruitful discussions, great help and providing the dataset analyzed in Chapter 3 of this dissertation. I also extend my gratitude to Professor Michael Daniels and Professor Paul Duncan for serving on my committee. I appreciate their constructive suggestions and precious time. I thank all the other professors in our department for their help throughout my graduate study.

I would like to convey my appreciation to Dr. Nilanjan Chatterjee (who is a Senior Investigator at the National Cancer Institute) for being my mentor during my training fellowship at the National Cancer Institute and for providing us a wonderful dataset which directly motivated the work in Chapter 4 of this dissertation. I would like to take this opportunity to thank my fellow graduate students in the Department of Statistics at the University of Florida. In particular, I thank Dr. Samiran Sinha (currently on the faculty at Texas A& M University) for many helpful discussions and for his contribution to the work in Chapter 4 of this dissertation. I thank my friend

Dr. Yan Gong (who is a faculty member at the University of Florida) for sharing her expertise in genetics.

I thank the College of Liberal Arts and Sciences at the University of Florida for awarding me the Keene Dissertation Fellowship Award, which provided a wonderful opportunity for me to focus solely on my research during the last stage of my dissertation.

Last, but not the least, my sincere thanks go to my family for their endless love, continuous support and encouragement during my life. This work is dedicated to all of them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

BAYESIAN METHODS IN CASE-CONTROL STUDIES WITH APPLICATIONS IN GENETIC EPIDEMIOLOGY

By

Li Zhang

August 2006

Chair: Malay Ghosh
Cochair: Bhramar Mukherjee
Major Department: Statistics

The fundamental idea behind case-control studies is to compare selected persons having a disease (the cases) with those not having the disease (the controls) by assessing to what extent they have been exposed to the disease's possible risk factors. The natural likelihood to use for a case-control study is a "retrospective" likelihood, i.e. a likelihood based on the probability of exposure given disease status. I prove the equivalence of posterior inference for the log odds ratios parameters based on prospective and retrospective likelihoods in stratified case-control studies in which some of the exposure variables could be missing completely at random.

My dissertation also addresses three problems in the domain of genetic epidemiology to explore a variety of disease-gene association and gene-environment interaction.

First, I consider the problem of detecting association between a disease and a candidate gene in the presence of population admixture. I propose a two-stage parametric Bayesian approach implemented via Markov chain Monte Carlo (MCMC) numerical integration technique, which first estimates the posterior probability of different unknown population substructures and then integrates this information into

a disease-gene association model through the technique of Bayesian model averaging. Thus, the uncertainty in estimating the population substructure is taken into account while providing credible intervals for parameters in the disease-gene association model.

Second, I present a Bayesian semiparametric approach to model the effect of stratification variables under the assumption of gene-environment independence in the control population conditional on some other covariates to study the gene-environment interaction. I take account of stratum heterogeneity in the exposure distribution by adopting the Dirichlet process mixture (DPM) of normal prior to the distribution of the environmental exposure and a flexible model for the distribution of the genetic factor. I illustrate the methods by applying them to an Israeli ovarian cancer study to investigate the effect of BRCA1/2 mutations, oral contraceptive use and parity in the development of ovarian cancer.

Third, I consider analysis of unmatched case-control studies in which binary exposures are potentially misclassified. I describe a relative simple approach to adjust the estimation of the parameters of interest in gene-environment association studies in the presence of misclassification and by exploiting the G-E independence assumption. Concluding remarks and directions for future work are included in the end.

CHAPTER 1
OVERVIEW

1.1 Introduction: The Frequentist Development in Case-Control Studies

The goal of an epidemiologic study is to find the causes of a disease and to assess the degree of association between the disease and its potential risk factors. Case-control studies are perhaps the most dominant form of analytical research in epidemiology, especially in cancer epidemiology. The fundamental idea behind such investigations is to compare selected persons having a disease (the cases) with those not having the disease (the controls) by assessing to what extent they have been exposed to the disease's possible risk factors. The ultimate goal often is to evaluate the hypothesis that one or more of the exposure variables is a cause of the disease. There are several popular study designs to ascertain disease-exposure association. A case-control study is *retrospective* in the sense that separate random samples from case and control populations are collected first and then exposure information is ascertained for the selected subjects. In such a study design, one collects exposure information conditional on the disease status of the subject. A cohort study, on the other hand, is *prospective* in nature as an initially healthy cohort is followed over time to assess the disease incidence rate and possible disease-exposure association.

Case-control studies design became popular in the 1920's. Initially, there were doubts regarding the validity of using case-control data to extract information on the relative risks of the disease, i.e., the odds of the occurrence of a disease for those exposed relative to those unexposed. Cornfield (1951) demonstrated that the exposure odds ratio for cases versus controls equals the disease odds ratio for exposed versus unexposed, and that the latter in turn approximates the ratio of disease rates or the relative risk of the disease provided that the disease is rare. To understand

Table 1–1: Case-control data with a binary exposure variable

Disease Status	Exposed	Non-Exposed	Total
Case	n_{11}	n_{10}	n_1
Control	n_{01}	n_{00}	n_0
Total	e_1	e_0	N

this issue in the simplest setting, consider a case-control study with a single binary exposure variable X ($X = 1$ exposed, and $X = 0$ unexposed) and let D denote the disease status ($D = 1$ for cases, $D = 0$ for controls). Table 1–1 presents the data layout and cell frequencies for each disease-exposure combination. One may note that

$$\begin{aligned} & \frac{P(X=1|D=1)P(X=0|D=0)}{P(X=1|D=0)P(X=0|D=1)} \quad \text{(the exposure odds ratio)} \\ = & \frac{P(D=1|X=1)P(D=0|X=0)}{P(D=0|X=1)P(D=1|X=0)} \quad \text{(the disease odds ratio)} \\ \approx & \frac{P(D=1|X=1)}{P(D=1|X=0)} \quad \text{(relative risk).} \end{aligned} \tag{1–1}$$

The approximation holds for a rare disease, as $P(D = 0|X = 1) \approx P(D = 0|X = 0) \approx 1$. So the disease odds ratio, say, $\psi = \exp(\beta)$ (where β denotes the log-odds ratio parameter), is the same as the exposure odds ratio which approximates the relative risk of the disease for a rare disease. Therefore, an odds ratio of 1 implies that there is no association between the disease and the exposure, whereas an odds ratio other than 1 implies that exposure is either synergistic or antagonistic with the disease. Also, one estimates ψ by $\hat{\psi} = n_{11}n_{00}/(n_{10}n_{01})$ and $\beta = \log(\psi)$ by $\hat{\beta} = \log(\hat{\psi})$.

1.1.1 The Mantel-Haenszel Era

It is well-known that for a large sample, $\hat{\beta} - \beta$ has an asymptotic normal distribution with mean 0 and variance $(1/n_{11} + 1/n_{10} + 1/n_{01} + 1/n_{00})$ (Agresti, 2001). For a small sample size, exact inference is based on a noncentral hypergeometric

Table 1–2: Series of 2×2 table for stratified case-control data

Disease Status	Exposed	Non-Exposed	Total
Case	n_{11i}	n_{10i}	n_{1i}
Control	n_{01i}	n_{00i}	n_{0i}
Total	e_{1i}	e_{0i}	N_i

distribution,

$$\Pr(n_{11}|n_1, n_0, e_1, e_0; \psi) = \frac{\binom{n_1}{n_{11}}\binom{n_0}{e_1 - n_{11}}\psi^{n_{11}}}{\sum_u \binom{n_1}{u}\binom{n_0}{e_1 - u}\psi^u}, \tag{1–2}$$

which is the conditional distribution of paired binomial data given the marginal totals (the marginal totals are considered as approximately ancillary in the sense that they do not contain any information about the parameter of interest ψ). One can use Fisher's exact test to test $H_0 : \psi = \psi_0$ against $H_1 : \psi > \psi_0$, by calculating the upper tail probability under the distribution shown in (1–2),

$$p_u = \sum_{u \geq n_{11}} P(u|n_1, n_0, e_1, e_0; \psi_0). \tag{1–3}$$

Similarly, to test H_0 against H_1: $\psi < \psi_0$ one should calculate the corresponding lower tail probability.

Mantel and Haenszel (1959) proposed an alternative to Fisher's exact test. Assuming a common odds ratio across a series of 2×2 tables, they proposed an estimator for the common odds ratio. Specifically, suppose one has I such tables and the i-th table is represented by the data layout in Table 1–2. The Mantel-Haenszel (MH) odds ratio estimate is given by

$$\hat{\psi}_{MH} = \exp(\hat{\beta}_{MH}) = \frac{\sum_i n_{11i} n_{00i}/N_i}{\sum_i n_{01i} n_{10i}/N_i}. \tag{1–4}$$

To test for homogeneity of odds ratios across the tables, i.e., $H_0 : \psi_1 = \psi_2 = ... = \psi_I$, the MH test statistic is

$$\chi^2 = \frac{\{|\sum_i (n_{11i} - E(n_{11i}|\hat{\psi}_{MH}))| - \frac{1}{2}\}^2}{\sum_i Var(n_{11i}|\hat{\psi}_{MH})}, \tag{1–5}$$

which has an approximate χ^2 distribution with $I - 1$ degrees of freedom.

Mantel and Haenszel presented no variance formula for their estimator and referred to the work by Cornfield (1956) for calculation of the interval estimates. Robins, Breslow and Greenland (1986) and Phillips and Holland (1987) independently proposed variance estimator of the MH estimator covering the two different types of asymptotic structure: 1) a small number of tables with large frequencies, and 2) a large number of tables with small frequencies. The main idea is the following. First, $E(R_i) = \psi_i E(S_i)$, where $R_i = n_{11i}n_{00i}/N_i$, $S_i = n_{01i}n_{10i}/N_i$, and ψ_i denotes the true odds ratio in table i. Thus $\hat{\psi}_{MH}$ is the solution of the unbiased estimating equation $R - \psi S = 0$, with $R = \sum_i R_i$ and $S = \sum_i S_i$, assuming a common value ψ for ψ_i. Second, under paired binomial sampling, the variances of the individual contributions to this estimating equation satisfy

$$N_i^2 Var(R_i - \psi S_i) = \frac{1}{2} E\{(n_{11i}n_{00i} + \psi n_{01i}n_{10i})(n_{11i} + n_{00i} + \psi(n_{01i} + n_{10i}))\}. \tag{1–6}$$

Now, with one step Taylor expansion,

$$\hat{\beta}_{MH} = \log(\hat{\psi}_{MH}) = \log(\psi) + \frac{R - \psi S}{E(R)} + o_p\Big(\frac{Var(R)}{E^2(R)} + \frac{Var(S)}{E^2(S)}\Big). \tag{1–7}$$

The last two equations together yield

$$Var(\hat{\beta}_{MH}) \doteq \frac{Var(R - \psi S)}{E^2(R)} = \frac{\sum_i Var(R_i - \psi S_i)}{E^2(R)}. \tag{1–8}$$

However, the MH methods concern the effects of a single binary risk factor. One may extend the methods to a single categorical exposure and then to multiple categorical exposures only by considering each factor at a time after stratification

with respect to levels of the other factors. Continuous exposures cannot be handled in this framework unless one categorizes them.

1.1.2 Logistic Regression in Case-Control Studies

Methods to evaluate simultaneous effects of multiple quantitative risk factors started being developed in the 1960's. Cornfield *et al.* (1961) noted that if the multivariate distribution of exposure $\boldsymbol{X}$ among persons with and without disease D were normal with separate means but a common covariance matrix, then the probability of developing disease for an individual with values $\boldsymbol{X} = \boldsymbol{x}$ was given by the logistic response curve

$$\Pr(D = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})}. \tag{1–9}$$

Day and Kerridge (1967) confirmed that logistic regression was efficient in a semeparametric sense. They noted that the full joint likelihood with exposure variables having an arbitrary distribution $p(\boldsymbol{x})$ can be written as $p(D, \boldsymbol{X}) = \Pr(D|\boldsymbol{X})p(\boldsymbol{x})$, and the two factors in the likelihood could be maximized separately, leading to semiparametric efficiency of the logistic model.

A key feature of the logistic model for case-control studies is that the regression coefficients $\boldsymbol{\beta}$ have a nice risk interpretation (Seigel and Greenhouse 1973) in the following sense:

$$\frac{\Pr(D = 1|\boldsymbol{X} = \boldsymbol{x}_1)\Pr(D = 0|\boldsymbol{X} = \boldsymbol{x}_0)}{\Pr(D = 0|\boldsymbol{X} = \boldsymbol{x}_1)\Pr(D = 1|\boldsymbol{X} = \boldsymbol{x}_0)} = \exp\{\boldsymbol{\beta}^T(\boldsymbol{x}_1 - \boldsymbol{x}_0)\}. \tag{1–10}$$

Thus $\boldsymbol{\beta}^T(\boldsymbol{x}_1 - \boldsymbol{x}_0)$ represents the log relative risk for a subject with exposure $\boldsymbol{x}_1$ versus one with exposure $\boldsymbol{x}_0$. But the natural likelihood for case-control sampling is the "retrospective" likelihood, and is of the form $p(\boldsymbol{X}|D)$ rather than $\Pr(D|\boldsymbol{X})$ which is the form of a "prospective" likelihood obtained from a cohort study. As Mantel and Haenszel (1959) stated in their seminal paper: "a primary goal is to reach the same

conclusions in a retrospective study as would have been obtained from a prospective study, if one had been done."

Prospective logistic regression analysis is indeed more convenient than fitting retrospective models. In a retrospective formulation, modeling the distribution of the exposure may pose certain challenges, especially when the exposure is high dimensional or a mixture of discrete and continuous variables. But the use of the prospective model in analyzing case-control data needed more theoretical validation which was provided by Anderson (1972) and Prentice and Pyke (1979). I will discuss this issue in greater detail in the next section.

1.1.3 Equivalence of Prospective and Retrospective Models in Case-Control Studies

As stated in (1–10) the prospective logistic regression model may be used to induce a retrospective model, which also turns out to be of a logistic form (Prentice and Pyke, 1979). Beginning with (1–10) and defining

$$\alpha = \log\left\{\frac{\Pr(D=1|\boldsymbol{x}_0)}{\Pr(D=0|\boldsymbol{x}_0)}\right\} - \boldsymbol{\beta}^T\boldsymbol{x}_0 \tag{1–11}$$

one can recover (1–9). Similarly the odds ratio representation (1–10) allows one to calculate

$$p(\boldsymbol{X}=\boldsymbol{x}|D=d) = \frac{\exp\{\gamma(\boldsymbol{x}) + d\boldsymbol{\beta}^T\boldsymbol{x}\}}{\int \exp\{\gamma(\boldsymbol{x}) + d\boldsymbol{\beta}^T\boldsymbol{x}\}d\boldsymbol{x}}, \quad d=0,1, \tag{1–12}$$

where $\gamma = \gamma(\boldsymbol{x}) = \log\{\Pr(\boldsymbol{X}=\boldsymbol{x}|D=0)/\Pr(\boldsymbol{X}=\boldsymbol{x}_0|D=0)\}$ for all $\boldsymbol{x}$. Further more, if $\boldsymbol{X}$ has K distinct values, the integration becomes summation over all K distinct values. The prospective model (1–9) and the retrospective model (1–12) are precisely equivalent provided that α in (1–9) and γ in (1–12) are unrestricted.

Anderson (1972) provides a deeper look into the proposition of retrospective data being analyzed by a prospective model. Suppose a discrete exposure variable $\boldsymbol{X}$ takes K distinct values $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_K$. There are $n = n_0 + n_1$ samples with n_0 controls and n_1

cases. Let n_{0k} and n_{1k} denote the number of controls and cases observed, respectively, with $\boldsymbol{X} = \boldsymbol{z}_k$. Denote $p^*_{1k} = 1 - p^*_{0k} = \Pr(D = 1|\boldsymbol{X} = \boldsymbol{z}_k)$, which is specified by the logistic model (1–9), and the marginal probabilities corresponding to the exposure are given by $q^*_k = \Pr(\boldsymbol{X} = \boldsymbol{z}_k)$. Assuming the marginal disease probabilities $\Pr(D = d) = \pi_d$ are known and by using $\Pr(\boldsymbol{X}|D) = \Pr(D|\boldsymbol{X}) \Pr(\boldsymbol{X})/ \Pr(D)$, the case-control likelihood is proportional to

$$L^*_1 L^*_2 = \{ \prod_{d=0}^{1} \prod_{k=1}^{K} (p^*_{dk})^{n_{dk}} \} \{ \prod_{k=1}^{K} (q^*_k)^{n_{+k}} \}, \tag{1–13}$$

where $n_{+k} = n_{0k} + n_{1k}$. But the parameters are constrained by fixed marginal probabilities of disease: $\sum_k p^*_{dk} q^*_k = \pi_d$, for $d = 0, 1$. Anderson (1972) discovered that estimates and covariance matrix for the coefficients $\boldsymbol{\beta}$ were identical to those of ordinary logistic regression involving maximization of L^*_1 alone.

Prentice and Pyke (1979) extended Anderson's (1972) results on logistic discrimination and generalized the findings of Breslow and Powers (1978) on the equivalence of odds ratio estimators when both prospective and retrospective logistic models are applied to case-control studies. They started from another factorization of the likelihood.

Again, let us consider n_0 controls and n_1 cases but an arbitrary exposure variable $\boldsymbol{x}$. The retrospective likelihood function is

$$L = \prod_{j:cases} P(\boldsymbol{x}_j|D = 1) \prod_{j:controls} P(\boldsymbol{x}_j|D = 0). \tag{1–14}$$

Denote S as a sampling indicator ($S = 1$, an individual is selected in the case-control sample; $= 0$, otherwise). Because conditional on disease status, sampling is

independent of exposure, by Bayes's theorem,

$$\begin{aligned} P(\boldsymbol{x}|D=d) &= P(\boldsymbol{x}|D=d,S=1) \\ &= \frac{P(D=d|\boldsymbol{x},S=1)P(\boldsymbol{x}|S=1)}{P(D=d|S=1)} \\ &= P(D=d|\boldsymbol{x},S=1)P(\boldsymbol{x}|S=1)\frac{n}{n_d}. \end{aligned} \tag{1–15}$$

As in Mantel (1973), we can obtain

$$P(D=1|\boldsymbol{x},S=1) = \frac{P(S=1|D=1)P(D=1|\boldsymbol{x})}{\sum_{d=0}^{1} P(S=1|D=d)P(D=d|\boldsymbol{x})}, \tag{1–16}$$

by the fact that sampling is independent of exposure within cases and controls. This is the conditional probability of an individual in cases, given exposure $\boldsymbol{x}$ and was sampled for the study.

Since $P(D=1|S=1)/P(D=0|S=1) = n_1/n_0$, inserting (1–9) into (1–16), one obtains

$$P(D=1|\boldsymbol{x},S=1) = \frac{\exp(\delta+\boldsymbol{\beta}^T\boldsymbol{x})}{1+\exp(\delta+\boldsymbol{\beta}^T\boldsymbol{x})}, \tag{1–17}$$

where $\delta = \alpha + \log\left(\frac{n_1\pi_0}{n_0\pi_1}\right)$.

Now substituting (1–15) into (1–14), one obtains

$$L \propto L_1 \times L_2, \tag{1–18}$$

where

$$\begin{aligned} L_1 &= \prod_{d=0}^{1}\prod_{j=1}^{n_d} P(D=d|\boldsymbol{x}_{dj},S=1) \\ L_2 &= \prod_{d=0}^{1}\prod_{j=1}^{n_d} P(\boldsymbol{x}_{dj}|S=1). \end{aligned} \tag{1–19}$$

Note that the parameters δ, $\boldsymbol{\beta}$ and $q(\boldsymbol{x})(= P(\boldsymbol{x}|S=1))$ are restricted by $n_d/n = \int P(\boldsymbol{x}|S=1)P(D=d|\boldsymbol{x},S=1)d\boldsymbol{x}$.

Prentice and Pyke (1979) demonstrated that the solution to the unconstrained maximization problem, with $(\hat{\delta}, \hat{\boldsymbol{\beta}})$ from the ordinary logistic regression coefficients based on L_1 and $\widehat{q(\boldsymbol{x})} = s/n$ (which is assigned to any value of $\boldsymbol{x}$ that is observed with multiplicity s) the sample $\boldsymbol{X}$ distribution, actually satisfied the constraints and thus yielded the desired estimates. They further showed that the estimating equations derived from L_1 were unbiased and, using estimating equation theory, confirmed that the usual covariance matrix for $\hat{\boldsymbol{\beta}}$ remained valid under case-control sampling. Because the intercept δ was a free parameter, it did not matter that the π_i's were unknown.

Carroll, Wang and Wang (1995) extended the Prentice and Pyke (1979) results to validate fitting of prospective logistic regression models to case-control data in the presence of measurement error and partial missingness in exposure values. They showed that, in general, using prospectively derived standard errors is at worst asymptotically conservative; in addition, they derived a simple sufficient condition guaranteeing that prospective standard errors are asymptotically correct.

Roeder, Carroll and Lindsay (1996) extended the Prentice and Pyke (1979) results to the case where covariates are measured with error. They proved that the prospective and retrospective models generate the same profile likelihood for the log odds ratio. By using a mixture model, the relationship between the true covariate X and the response D can be modeled appropriately for both complete and reduced data. The likelihood depends on the marginal distribution of X and the measurement error density $[W|X, D]$. The latter is modeled parametrically based on the validation sample. The marginal distribution of the true covariate is modeled using a nonparametric mixture distribution.

Seaman and Richardson (2004) presented an alternative proof of equality of the two profile likelihoods in the absence of measurement error, where they applied the multinomial-Poisson (MP) transformation. Furthermore, they proved that a Bayesian

analysis which uses the prospective likelihood and assumes a uniform prior distribution for the log odds that an individual with baseline exposure is diseased, is exactly equivalent to an analysis that uses the retrospective likelihood and assumes a Dirichlet prior distribution for the exposure probabilities in the control group. This means that Bayesian analysis of case-control studies may, like the classical frequentist analysis, be carried out using a prospective model, thus significantly reducing its complexity.

Seaman and Richardson (2004), like Prentice and Pyke (1979), considered unmatched case-control problems. They left the open question of similar equivalence results in the context of matched case-control problems and also for situations with missing data. In my dissertation I address the problem of extending the equivalence results to stratified case-control studies in which some of the exposure variables could be missing completely at random.

1.1.4 Matched Case-Control Studies

So far I have concentrated on unmatched case-control study designs, but my dissertation will involve some matched case-control settings as well, so I briefly review the matched study design. Matching is often implemented as a design strategy to eliminate effects due to confounding. In a matched case-control study, controls are matched with a case (or several cases) on the basis of some matching factors (confounding variables) such as age, gender, region, ethnicity etc. There are two types of matching commonly used. One is *frequency matching*, in which the number of controls are selected according to the number of cases in broad homogeneous strata defined by the values of matching factors to maintain a specific case:control ratio in each stratum. The other is *individual matching*, in which controls are selected individually corresponding to each selected case by matching with respect to certain factors.

The simplest situation of matched data arises when one case is matched with one control, and they are categorized on the basis of a binary exposure. Suppose one

Table 1–3: Matched case-control data with a binary exposure variable

Disease Status	Exposed	Non-Exposed
Case	m_{11}	m_{10}
Control	m_{01}	m_{00}

has m_{11}, m_{10}, m_{01} and m_{00} matched pairs under different levels of D and X as shown in Table 1–3. Let π be the conditional probability of observing a matched pair with an exposed case and unexposed control given a discordant pair.

$$\pi = \frac{P(X=1|D=1)P(X=0|D=0)}{P(X=1|D=1)P(X=0|D=0)+P(X=0|D=1)P(X=1|D=0)} = \frac{\psi}{\psi+1}. \tag{1–20}$$

Note that $m_{10}|m_{10}, m_{01} \sim Bin(m_{10}+m_{01}, \pi)$. So the Mantel Haenszel estimator of the common odds ratio parameter, the MLE of ψ, is m_{10}/m_{01}. Note that when $\psi = 1$, $\pi = 1/2$. Hence the test statistic to test $H_0 : \psi = 1$ is

$$\chi^2 = \frac{(|m_{10} - E_{H_0}(m_{10}|m_{10}+m_{01})| - \frac{1}{2})^2}{Var_{H_0}(m_{10}|m_{10}+m_{01})}, \tag{1–21}$$

which is known as McNemar's (1947) test. One of the potential problems with this estimator and this test is that it uses only the discordant pairs of observations and discards the information contained in the concordant set.

In the case of $1 : M$ matching, the Mantel-Haenszel estimator of common odds ratio is

$$\hat{\psi}_{MH} = \frac{\sum_{r=1}^{M}(M-r+1)m_{1r-1}}{\sum_{r=1}^{M} r m_{0r}}, \tag{1–22}$$

where m_{1r} is the number of matched sets where the case and r controls are exposed; and m_{0r} is the number of matched sets where the case is unexposed but r controls are exposed. The test statistic for testing $H_0 : \psi = 1$ is

$$\chi^2 = \frac{\left\{|\sum_{r=1}^{M}(m_{1r-1} - \frac{rt_r}{M+1})| - \frac{1}{2}\right\}^2}{\sum_{r=1}^{M} rt_r \frac{(M-r+1)}{(M+1)^2}}, \tag{1–23}$$

where $t_r = m_{1r} + m_{0r}$.

Let us now focus on logistic regression models in matched case-control studies. In the simplest setting, the data consist of I strata and there are M_i controls matched with a case, for stratum S_i, $i = 1, \cdots, I$. As before, one assumes a prospective logistic incidence model for disease

$$P(D = 1|\boldsymbol{z}, S_i) = \frac{\exp\{\alpha_i + \boldsymbol{\beta}^T(\boldsymbol{z} - \boldsymbol{z}_0)\}}{1 + \exp\{\alpha_i + \boldsymbol{\beta}^T(\boldsymbol{z} - \boldsymbol{z}_0)\}}, \tag{1–24}$$

where α_i's are stratum specific intercept terms. Without loss of generality, assuming that the first subject in each stratum is a case and rest of the subjects are controls, conditioning on the sufficient statistics $\sum_{j=1}^{M_i+1} D_{ij}$ for α_i, one obtains the conditional likelihood

$$\begin{aligned} L_c &= \prod_{i=1}^{I} \prod_{j=1}^{M_i+1} P(D_{ij}|\boldsymbol{z}_{ij}, S_i, \sum_{j=1}^{M_i+1} D_{ij} = 1)\} \\ &= \prod_{i=1}^{I} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{z}_{i1})}{\sum_{j=1}^{M_i+1} \exp(\boldsymbol{\beta}^T \boldsymbol{z}_{ij})}. \end{aligned} \tag{1–25}$$

This method is known as conditional logistic regression (CLR). Breslow (1996) illustrated that unmatched analysis of matched data based on unconditional full likelihood led to biased and inconsistent estimates of the relative risk parameters. The difference between unconditional and conditional analysis depends on the degree of association between the exposure and the matching variables. It is indeed important to acknowledge the matched study design into any model proposed for matched data.

1.2 Bayesian Analysis of Case-Control Studies

Since the methods I propose in my dissertation are mostly based on the Bayesian paradigm, I will now present a brief account of the current state of the art in Bayesian methods for case-control studies. In spite of the vast literature in the frequentist domain, Bayesian methods for analyzing case-control data were first proposed in the 1980's. With the arrival of Markov chain Monte Carlo (MCMC) techniques in the 1990's, it became possible to address more complex and unorthodox data scenarios

like missingness and measurement error in the context of a case-control study even in a Bayesian framework.

Zelen and Parker (1986), Nurminen and Mutanen (1987), Marshall (1988), and Ashby *et al.* (1993) developed Bayesian methods for analyzing case-control studies with only a single binary exposure variable. All of them used versions of the following model:

Let ϕ and γ be the probabilities of exposure in control and case populations, respectively. The retrospective likelihood is

$$l(\phi, \gamma) \propto \phi^{n_{01}}(1-\phi)^{n_{00}}\gamma^{n_{11}}(1-\gamma)^{n_{10}}, \tag{1–26}$$

where n_{01} and n_{00} are the number of exposed and unexposed observations in a control population, whereas n_{11} and n_{10} denote the same for a case population.

Independent conjugate prior distributions for ϕ and γ are assumed to be $Beta(u_1, u_2)$ and $Beta(v_1, v_2)$ respectively. After reparametrization, one obtains the posterior distribution of the log odds ratio parameter, $\beta = \log\{\gamma(1-\phi)/(\phi(1-\gamma))\}$ as

$$\begin{aligned} & p(\beta|n_{11}, n_{10}, n_{01}, n_{00}) \\ \propto \quad & \exp\{(n_{11}+v_1)\beta\}\int_0^1 \frac{\phi^{n_{11}+n_{01}+v_1+u_2-1}(1-\phi)^{n_{10}+n_{00}+v_2+u_1-1}}{\{1-\phi+\phi\exp(\beta)\}^{n_{11}+n_{10}+v_1+v_2}}d\phi. \end{aligned} \tag{1–27}$$

The posterior density of β does not exist in closed from, but may be evaluated by numerical integration.

Since interest often lies in the hypothesis $\beta = 0$, Zelen and Parker (1986) recommended calculating the ratio of the two posterior probabilities $p(\beta)/p(0)$ at selected deviates β. When β is set at the posterior mode, a large value of this ratio will indicate concentration of the posterior away from 0 and one would infer disease-exposure association. However, the critical value suggested for this ratio is completely arbitrary. They also provided a normal approximation to the posterior distribution of β to avoid

numerical computation, and discussed the problem of choosing a prior distribution based on some prior data on exposure information in a Bayesian framework.

Nurminen and Mutanen (1987) considered a more general parametrization in terms of the odds ratio $\psi = \exp(\beta)$ which covers risk ratio and risk differences. They provided a complicated exact formula for the cumulative distribution function of this general comparative parameter, which can be related to Fisher's exact test for comparing two proportions in sampling theory. The Bayesian point estimates were considered as posterior median and mode, whereas inference was based on highest posterior density interval for the comparative parameter of interest.

Marshall (1988) provided a closed-form expression for the moments of the posterior distribution of the odds ratio. He mentioned that an approximation to the exact posterior density of the odds ratio parameter can be obtained by power series expansion of the hypergeometric functions involved in the expression for the density, but acknowledged the problem of slow convergence in adopting this method. Instead Marshall used Lindley's (1964) result for the approximate normality of log(odds ratio) which works very well over a wide range of situations. In the absence of exposure information, Marshall recommended using independent priors on the parameters. He suggested that a perception about the value of the odds ratio should guide the choice of prior parameters rather than attempting to exploit the exposure proportions as suggested in Zelen-Parker. Inference again is based on posterior credible intervals.

Müller and Roeder (1997) proposed a semiparametric Bayesian approach to case-control studies having continuous exposures with measurement error. They used a Bayesian non-parametric model for the joint marginal distribution of the true exposure (where available), the surrogate and the measurement error. Their methods are intrinsically designed for continuous exposure. Müller *et al.* (1999) proposed a hierarchical Bayesian approach for combining the data from a case-control study and a prospective cohort study, and to estimate the absolute risk of the disease. They

modeled the retrospective distribution of the exposure variable given the disease status, and accounted for parameter heterogeneity across studies by using a hierarchical Bayesian approach.

Diggle, Morris and Wakefield (2000) presented the first Bayesian analysis for individually matched case-control data (appropriate nuisance parameters are introduced to represent the separate effect of matching in each matched set to recognize the study design). They considered matched data when exposure of primary interest is defined by the spatial location of an individual relative to a point or line source of pollution.

Seaman and Richardson (2001) extended the binary exposure model of Zelen-Parker to any number of categorical exposures, by simply replacing the binomial likelihoods in (1–26) by a multinomial likelihood, and then adopting a MCMC strategy with respect to a baseline category. They also adapted the Müller-Roeder approach to the setting with categorical exposures and illustrate that under certain specific choices of a discrete Dirichlet prior on the exposure distribution, Zelen-Parker and Müller-Roeder approaches became approximately equivalent.

Ghosh and Chen (2002) developed general Bayesian inferential techniques for matched case-control problems in the presence of one or more binary exposure variables. Their model was more general than that of Zelen and Parker (1986), and was based on an unconditional likelihood rather than a conditional likelihood unlike Diggle, Morris and Wakefield (2000). The general Bayesian methodology based on the full likelihood that they proposed worked beyond the logit link. Their procedure included not only the probit and the complementary log links but also some new symmetric as well as skewed links. The propriety of posteriors was proved under a very general class of priors that need not always be proper.

Sinha *et al.* (2005a) presented a unified semiparametric Bayesian approach to matched case-control studies with missing exposure. They assumed a Dirichlet process prior with a mixing normal distribution on the distribution of the stratum effects on the exposure distribution. The proposed method possessed certain attractive robustness properties under varying degrees of stratum heterogeneity in the exposure distribution.

Sinha *et al.* (2004) considered matched case-control studies with multiple disease states. They further extended their methods to model multivariate exposure with association and partial missingness (Sinha *et al.*, 2005b). To summarize, they presented an ensemble of methods to handle unorthodox data scenarios in matched case-control studies.

1.3 Topics of This Dissertation

A resurgence of interest has been recently expressed in genetic case-control studies (Risch and Merikangas, 1996; Morton and Collins, 1998; Sullivan *et al.*, 2001) to explore a variety of disease-gene association and gene-environment interaction. The Bayesian pathways have remained less explored in the case-control context mainly because of the computing needs for implementing the models.

In genetic case-control studies, accounting for population substructure is a critical issue in a population where admixture of several ancestry has taken place. A systematic difference in ancestry in cases and controls can lead to false discovery of association. In my dissertation, I propose a two-stage parametric Bayesian approach which integrates the model uncertainty into a disease-gene association model through the technique of Bayesian model averaging, where the analysis is not limited to binary genotypes irrespective of whether or not the disease is rare.

Many human diseases result from the interplay of genetic factors and environmental exposures. One may exploit the gene-environment independence in order to derive more efficient estimation techniques than the traditional logistic regression analysis.

I provide Bayesian nonparametric methods to capture stratification effects on the distribution of environmental exposures under the gene-environment independence assumption in the *control* population. Also in a Bayesian paradigm I can effectively use the prior knowledge while modeling the individual genotype frequencies in each stratum and thus relax the stringent logistic assumption. My objective will be not only to estimate the interaction effect parameter, but also to estimate the effects of the genetic factor and environmental exposures as well.

Measurement error in exposure assessment is one of the major source of bias in epidemiological studies. When ignored, these errors bias our point and interval estimates of effect, and invalidate p-values of hypotheses tests. Less attention has been given to the influence of misclassification on the assessment of interactions between two or more factors. Based on sensitivity and specificity of the genetic and environmental factors, I describe a relative simple approach to adjust the estimation of the parameters of interest in gene-environment association studies in the presence of misclassification while exploiting the G-E independence assumption.

The outline of the rest of my dissertation is as the following. In Chapter 2, I present a general result which shows that the posterior inference for the parameters from a multinomial likelihood is exactly equivalent to that from the corresponding Poisson likelihood with an arbitrary proper prior for the parameters of interest and independent uniform priors for the latent parameters. The result is then extended to prove the equivalence of posterior inference for the odds ratio parameter based on prospective and retrospective likelihoods in stratified case-control studies where some of the exposure variables could be missing completely at random.

In Chapter 3, I propose a parametric Bayesian approach to examine the association between a candidate gene and the occurrence of a disease in the presence of population admixture. Two unmatched case-control simulation studies based on an

admixed Argentinean population as described in Sala *et al.* (1998, 1999) are performed to illustrate the methods and computing scheme. The method is also applied to a real dataset coming from a genetic association study on obesity.

In Chapter 4, I provide a novel semiparametric Bayesian approach to model stratification effects under the assumption of gene-environment independence in the control population. I illustrate the methods by applying them to data from a population-based case-control study on ovarian cancer conducted in Israel. Simulation studies are conducted to compare our method with other popular choices. The results reflect that the semiparametric Bayesian model allows incorporation of key scientific evidence in the form of a prior and offers a flexible, robust alternative when standard parametric model assumptions do not hold.

In Chapter 5, I derive analytic formulation to obtain estimates and confidence intervals for the misclassified case in a unmatched case-control set-up, which reduce back to standard analytic forms as the error probabilities reduce to zero. I adapt and extend the work of Rice and Holmans (2003) to the situation when one has a binary genetic risk factor, a binary environmental exposure, and both are potentially subject to misclassification. Concluding remarks and directions for future work are stated in Chapter 6.

CHAPTER 2
EQUIVALENCE OF POSTERIORS IN THE BAYESIAN ANALYSIS OF THE MULTINOMIAL-POISSON TRANSFORMATION

2.1 Introduction

Baker (1994) presented a general result which showed how maximum likelihood estimation of parameters from a multinomial distribution could be carried out from a corresponding Poisson likelihood by exploiting the multinomial-Poisson relationship. Henceforth, this will be referred to as the multinomial-Poisson (MP) transformation. Baker considered situations where the multinomial probabilities were ratios of functions of parameters to the sum of these functions. The motivation was to simplify the maximum likelihood computation as well as computation of the asymptotic variance-covariance matrix of the maximum likelihood estimate (MLE). Baker's result unified a large number of analyses involving log-linear models, capture-recapture models, proportional hazards models with categorical covariates, generalized Rasch models, voter plurality models, conditional logistic regression and two-stage case-control studies.

Baker's ideas were extended in the context of Bayesian analysis of case-control studies by Seaman and Richardson (2004). The natural likelihood to use for a case-control study is a "retrospective" likelihood, i.e., a likelihood based on the probability of exposure given the disease status. Prentice and Pyke (1979) showed that, when a logistic regression is assumed for the probability of a disease given certain exposures, the maximum likelihood estimators and asymptotic covariance matrix of the log odds ratios obtained from the retrospective likelihood are the same as those obtained from the prospective likelihood, i.e., that based on the probability of a disease given exposures. The objective of Seaman and Richardson (2004) was to verify a result similar to Prentice and Pyke (1979) for the posterior distribution of the log odds ratios in

a Bayesian analysis. They proved that a Bayesian analysis that uses the prospective likelihood, and assumes *a uniform prior for the log odds that an individual with baseline exposure is diseased*, is equivalent to an analysis that uses the retrospective likelihood and assumes *a Dirichlet prior for the exposure probabilities in the control group*. Earlier, an approximate equivalence result was indicated by Gustafson *et al.* (2002). Seaman and Richardson left open the question of similar equivalence for stratified case-control data with missing exposure values.

In Section 2.2 of this chapter, first based on a MP transformation in a Bayesian framework, I prove a general result which shows that the posterior inference for the parameters of a multinomial likelihood is the same as that for the corresponding Poisson likelihood with arbitrary proper priors for the parameters of interest and uniform priors for the latent parameters introduced in the Poisson likelihood. Propriety of posteriors under the assumed priors follow as an immediate consequence. In Section 2.3, I extend the results of Seaman and Richardson (2004) to stratified case-control problems where some of the exposure variables could be missing completely at random. Stratified case-control problems without any missingness can be handled as special cases. Individually matched case-control design is a special case of stratified case-control design where the matched sets define the strata. Finally, some concluding remarks are made in Section 2.4.

2.2 A General Result on Posterior Equivalence

Let $\{Y_{ij}; j \in J_i, i = 1, 2, \cdots, I\}$ denote a vector of discrete random variables with a realization $\{y_{ij}; j \in J_i, i = 1, 2, \cdots, I\}$. The subscript i indexes levels of a categorical covariate or a cross-classification of categorical covariates, and J_i (indexed by j) denotes the set of subjects in level i. I assume that the vector $\{Y_{ij}; j \in J_i\}$ follows a multinomial distribution with parameters $\{g_{ij}(\boldsymbol{\beta})/G_i(\boldsymbol{\beta})$, for $j \in J_i\}$, where $g_{ij}(\boldsymbol{\beta})$ are some functions of $\boldsymbol{\beta}$, $G_i(\boldsymbol{\beta}) = \sum_{j \in J_i} g_{ij}(\boldsymbol{\beta})$, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_q)^T$. The

likelihood function is then proportional to

$$L_M(\boldsymbol{\beta}) = \prod_{i=1}^{I} \prod_{j \in J_i} \left\{ \frac{g_{ij}(\boldsymbol{\beta})}{G_i(\boldsymbol{\beta})} \right\}^{y_{ij}}. \tag{2–1}$$

Let $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_i, \cdots, \phi_I)^T$ indicate a set of parameters. The MP transformation of (2–1) as given by Baker (1994) is the corresponding Poisson likelihood proportional to:

$$L_P(\boldsymbol{\phi}, \boldsymbol{\beta}) = \prod_{i=1}^{I} \prod_{j \in J_i} \{g_{ij}(\boldsymbol{\beta}) \exp \phi_i\}^{y_{ij}} \exp\{-g_{ij}(\boldsymbol{\beta}) \exp \phi_i\}. \tag{2–2}$$

THEOREM 1. Suppose $\sum_{j \in J_i} y_{ij} \geq 1$ for all $i = 1, 2, \cdots, I$. Assume independent improper priors $p(\phi_i) \propto 1$, for $i = 1, \cdots, I$, and a proper prior $p(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ which is independent of $\boldsymbol{\phi}$. Then the posterior distribution for $\boldsymbol{\beta}$ derived from $L_M(\boldsymbol{\beta})$ is equivalent to that generated from $L_P(\boldsymbol{\phi}, \boldsymbol{\beta})$.

Proof: Let $\alpha_i = \exp(\phi_i)$, $i = 1, \cdots, I$. Then α_i has the prior $p(\alpha_i) \propto \alpha_i^{-1}$. The marginal posterior of $\boldsymbol{\beta}$ from $L_P(\boldsymbol{\phi}, \boldsymbol{\beta})$ is now given by

$$\begin{aligned}
\pi(\boldsymbol{\beta}|\mathbf{y}) &\propto p(\boldsymbol{\beta}) \prod_{i=1}^{I} \int_0^{\infty} \left\{ \alpha_i^{-1} \prod_{j \in J_i} \{\alpha_i g_{ij}(\boldsymbol{\beta})\}^{y_{ij}} \exp\{-\alpha_i g_{ij}(\boldsymbol{\beta})\} \right\} d\alpha_i \\
&= p(\boldsymbol{\beta}) \prod_{i=1}^{I} \left\{ \prod_{j \in J_i} \{g_{ij}(\boldsymbol{\beta})\}^{y_{ij}} \int_0^{\infty} \alpha_i^{\sum_{j \in J_i} y_{ij} - 1} \exp\{-\alpha_i G_i(\boldsymbol{\beta})\} d\alpha_i \right\} \\
&\propto p(\boldsymbol{\beta}) \prod_{i=1}^{I} \prod_{j \in J_i} \left\{ \frac{g_{ij}(\boldsymbol{\beta})}{G_i(\boldsymbol{\beta})} \right\}^{y_{ij}} \\
&= p(\boldsymbol{\beta}) L_M(\boldsymbol{\beta}),
\end{aligned}$$

which is obviously the same as the posterior distribution of $\boldsymbol{\beta}$ generated from $L_M(\boldsymbol{\beta})$.

The following theorem establishes the propriety of the above posterior under very mild conditions.

COROLLARY 1. If $\sum_{j \in J_i} y_{ij} \geq 1$ for all $i = 1, \cdots, I$, and $p(\boldsymbol{\beta})$ is proper, then $\pi(\boldsymbol{\beta}|\mathbf{y})$ is proper.

Proof. Let Ω denote the support of $\boldsymbol{\beta}$. Then by Theorem 1,

$$\int_{\boldsymbol{\beta}\in\Omega} \pi(\boldsymbol{\beta}|\mathbf{y})d\boldsymbol{\beta} \propto \int_{\boldsymbol{\beta}\in\Omega} p(\boldsymbol{\beta})L_M(\boldsymbol{\beta})d\boldsymbol{\beta}$$

Propriety of the posterior thus follows as an immediate consequence of the equivalence of the two analyses.

REMARK 1. If instead I use independent priors $p(\alpha_i) \propto \alpha_i^{a_i-1}$ ($a_i > 0$ for all $i = 1, \cdots, I$), then the assumption $\sum_{j\in J_i} y_{ij} \geq 1$ can be dropped to establish propriety of the resulting posterior for $\boldsymbol{\beta}$. But this posterior will no longer be proportional to $p(\boldsymbol{\beta})L_M(\boldsymbol{\beta})$ as $G_i(\boldsymbol{\beta})$ will then have the power $\sum_{j\in J_i} y_{ij} + a_i$ rather than $\sum_{j\in J_i} y_{ij}$.

Bayesian analogous of all the examples of Baker (1994) can now be handled from this general theorem. For brevity, I omit these examples, and proceed to the next section to show the equivalence of posteriors based on prospective and retrospective likelihoods in stratified case-control studies where some of the exposure variables could be missing completely at random (Little and Rubin, 2002).

2.3 Stratified Case-Control Studies with Missing Exposures

In this section, I prove that a Bayesian analysis of stratified case-control data with missing exposure that uses the prospective likelihood, and assumes *a uniform prior for the log odds that an individual with baseline exposure is diseased*, is exactly equivalent to an analysis that uses the retrospective likelihood and assumes *a uniform prior distribution for the exposure probabilities in the control group.* My analysis handles the case when some of the exposure variables are missing completely at random.

Suppose there are I strata where each stratum has s cases and t controls in a stratified case-control study. Let S_i denote the i-th stratum. Let D_{ij}(= 1 or 0) correspond to the presence or absence of a disease for the jth individual in ith stratum, and let $\boldsymbol{x}_{ij}$ denote the vector of discrete exposure variables for the jth observed subject in the ith stratum. I assume that each $\boldsymbol{x}_{ij}$ can take one of the K

possible values $\{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_K\}$. Suppose now

$$\begin{aligned} P(D_{ij} = 1 | \boldsymbol{X}_{ij} = \boldsymbol{z}_k, S_i) &= \frac{\alpha_i \exp(\boldsymbol{\beta}^T \boldsymbol{z}_k)}{1 + \alpha_i \exp(\boldsymbol{\beta}^T \boldsymbol{z}_k)}; \\ P(\boldsymbol{X}_{ij} = \boldsymbol{z}_k | D_{ij} = 0, S_i) &= \frac{\gamma_{ik}}{\sum_{l=1}^{K} \gamma_{il}}. \end{aligned} \tag{2–3}$$

The probability that individual j in stratum i has exposure value $\boldsymbol{z}_k$ given that the individual is a member of the control population is proportional to γ_{ik}. For each exposure value $\boldsymbol{z}_k$, these probabilities are assumed to be same for all controls in stratum i and do not depend on j.

Using (2–3) I can obtain the distribution of the exposure in the case population and write the prospective and retrospective models in the following form

$$\begin{aligned} P(D_{ij} = d | \boldsymbol{X}_{ij} = \boldsymbol{z}_k, S_i) &= \frac{\alpha_i^d \exp(d\boldsymbol{\beta}^T \mathbf{z}_k)}{\sum_{l=0}^{1} \alpha_i^l \exp(l\boldsymbol{\beta}^T \boldsymbol{z}_k)}; \\ P(\boldsymbol{X}_{ij} = \boldsymbol{z}_k | D_{ij} = d, S_i) &= \frac{\gamma_{ik} \exp(d\boldsymbol{\beta}^T \boldsymbol{z}_k)}{\sum_{l=1}^{K} \gamma_{il} \exp(d\boldsymbol{\beta}^T \boldsymbol{z}_l)}, \end{aligned} \tag{2–4}$$

where $d = 0, 1$.

Let Δ_{ij} denote the missingness indicator for the ith stratum (0 indicating missingness) with

$$P(\Delta_{ij} = 1 | S_i) = 1 - P(\Delta_{ij} = 0 | S_i) = \eta_i. \tag{2–5}$$

Let $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_I)^T$. With the missing completely at random assumption, η_i does not depend on the parameters γ_{ik}, α_i or $\boldsymbol{\beta}$.

Let $y_{idk} = \sum_{j=1}^{s+t} \{I[\boldsymbol{X}_{ij} = \boldsymbol{z}_k] I[D_{ij} = d] I[\Delta_{ij} = 1]\}$, $d = 0, 1$, i.e, y_{i0k} and y_{i1k} are the respective numbers of undiseased and diseased subjects having $\boldsymbol{X} = \boldsymbol{z}_k$ in the ith stratum, and I denotes the usual indicator function. Now, the prospective likelihood

is

$$\begin{aligned} L_P &= \prod_{i=1}^{I}\prod_{j=1}^{s+t}\Big[P(D_{ij}|\boldsymbol{x}_{ij},S_i)\Big]^{\Delta_{ij}} \\ &= \prod_{i=1}^{I}\prod_{d=0}^{1}\prod_{k=1}^{K}\Big[\frac{\alpha_i^d\exp(d\boldsymbol{\beta}^T\boldsymbol{z}_k)}{\sum_{l=0}^{1}\alpha_i^l\exp(l\boldsymbol{\beta}^T\boldsymbol{z}_k)}\Big]^{y_{idk}}, \end{aligned} \tag{2–6}$$

and the retrospective likelihood is

$$L_R = \prod_{i=1}^{I}\prod_{d=0}^{1}\prod_{k=1}^{K}\Big[\frac{\gamma_{ik}\exp(d\boldsymbol{\beta}^T\boldsymbol{z}_k)}{\sum_{l=1}^{K}\gamma_{il}\exp(d\boldsymbol{\beta}^T\boldsymbol{z}_l)}\Big]^{y_{idk}}. \tag{2–7}$$

I now have the following equivalence theorem.

THEOREM 2. Suppose $\sum_{k=1}^{K} y_{i1k} \geq 1$ and $\sum_{k=1}^{K} y_{i0k} \geq 1$, for all $i = 1, \cdots, I$. Assume mutually independent priors for the α_i, γ_{ik}, $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$, where $p(\alpha_i) \propto \alpha_i^{-1}$, $p(\gamma_{ik}) \propto \gamma_{ik}^{-1}$, while $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ have proper priors $\pi_1(\boldsymbol{\eta})$ and $\pi_2(\boldsymbol{\beta})$. Then the posterior distribution of $\boldsymbol{\beta}$ derived from the prospective likelihood is approximately equivalent to that from the retrospective likelihood.

Proof: Suppose that random variables Y_{idk} are independently distributed as $Y_{idk} \sim$ Poisson(λ_{idk}), where

$$\log\lambda_{idk} = \log\eta_i + \log\gamma_{ik} + d(\log\alpha_i + \boldsymbol{\beta}^T\boldsymbol{z}_k). \tag{2–8}$$

Then writing $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_I)^T$, and $\boldsymbol{\gamma} = (\gamma_{11}, \cdots, \gamma_{1K}, \cdots, \gamma_{I1}, \cdots, \gamma_{IK})^T$, the joint prior is

$$\pi(\boldsymbol{\eta},\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta}) \propto \pi_1(\boldsymbol{\eta})\pi_2(\boldsymbol{\beta})\Big\{\prod_{i=1}^{I}\Big(\alpha_i^{-1}\big(\prod_{k=1}^{K}\gamma_{ik}^{-1}\big)\Big)\Big\}.$$

The joint posterior is now given by

$$\begin{aligned} \pi(\boldsymbol{\eta},\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta}|\mathbf{y}) &\propto \prod_{i=1}^{I}\prod_{d=0}^{1}\prod_{k=1}^{K}\Big\{\exp(-\lambda_{idk})\frac{\lambda_{idk}^{y_{idk}}}{y_{idk}!}\Big\}\pi_1(\boldsymbol{\eta})\pi_2(\boldsymbol{\beta})\Big\{\prod_{i=1}^{I}\Big(\alpha_i^{-1}\big(\prod_{k=1}^{K}\gamma_{ik}^{-1}\big)\Big)\Big\} \\ &= \prod_{i=1}^{I}\prod_{d=0}^{1}\prod_{k=1}^{K}\Big\{\Big[\exp\{-\gamma_{ik}\eta_i\alpha_i^d\exp(d\boldsymbol{\beta}^T\boldsymbol{z}_k)\}\frac{\big(\gamma_{ik}\eta_i\alpha_i^d\exp(d\boldsymbol{\beta}^T\boldsymbol{z}_k)\big)^{y_{idk}}}{y_{idk}!}\Big]\Big\} \\ &\times \Big\{\prod_{i=1}^{I}\big(\prod_{k=1}^{K}\gamma_{ik}^{-1}\big)\Big\}\pi_1(\boldsymbol{\eta})\pi_2(\boldsymbol{\beta})\Big\{\prod_{i=1}^{I}\alpha_i^{-1}\Big\}. \end{aligned} \tag{2–9}$$

First note that

$$\begin{aligned}&\int_0^\infty \exp\big\{-\eta_i\gamma_{ik}[1+\alpha_i\exp(\boldsymbol{\beta}^T\boldsymbol{z}_k)]\big\}\gamma_{ik}^{y_{i0k}+y_{i1k}-1}d\gamma_{ik}\\ \propto\ &\eta_i^{-(y_{i0k}+y_{i1k})}[1+\alpha_i\exp(\boldsymbol{\beta}^T\boldsymbol{z}_k)]^{-(y_{i0k}+y_{i1k})}.\end{aligned}$$

Thus the joint posterior of $\boldsymbol{\eta}, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ is given by

$$\begin{aligned}\pi(\boldsymbol{\eta},\boldsymbol{\alpha},\boldsymbol{\beta}|\mathbf{y}) &\propto \prod_{i=1}^{I}\prod_{d=0}^{1}\prod_{k=1}^{K}\Big[\frac{\alpha_i^d\exp(d\boldsymbol{\beta}^T\boldsymbol{z}_k)}{\sum_{l=0}^{1}\alpha_i^l\exp(l\boldsymbol{\beta}^T\boldsymbol{z}_k)}\Big]^{y_{idk}}\pi_1(\boldsymbol{\eta})\pi_2(\boldsymbol{\beta})\Big\{\prod_{i=1}^{I}\alpha_i^{-1}\Big\}\\ &= L_P\pi_1(\boldsymbol{\eta})\pi_2(\boldsymbol{\beta})\Big\{\prod_{i=1}^{I}\alpha_i^{-1}\Big\}.\end{aligned}$$

Next integrating out $\boldsymbol{\eta}$, the joint posterior of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is

$$\pi(\boldsymbol{\alpha},\boldsymbol{\beta}|\mathbf{y}) \propto L_P\pi_2(\boldsymbol{\beta})\prod_{i=1}^{I}\alpha_i^{-1}. \tag{2–10}$$

Let $\theta_{ik}=\gamma_{ik}/\sum_{l=1}^{K}\gamma_{il}$ and $\varphi_i=\sum_{l=1}^{K}\gamma_{il}$; thus $\gamma_{ik}=\varphi_i\theta_{ik}$. The Jacobian of this transformation is

$$\Big|\frac{\partial(\gamma_{i1},\cdots,\gamma_{iK})}{\partial(\theta_{i1},\cdots,\theta_{iK-1},\varphi_i)}\Big| = \varphi_i^{K-1}.$$

Thus, the prior structure on γ_{ik} implies the following prior structure for $\boldsymbol{\varphi}=(\phi_1,\cdots,\phi_I)$, and $\boldsymbol{\theta}=(\theta_{11},\cdots,\theta_{1K},\cdots,\theta_{I1},\cdots,\theta_{IK})$:

$$p(\boldsymbol{\varphi},\boldsymbol{\theta}) = \prod_{i=1}^{I}\varphi_i^{K-1}\prod_{i=1}^{I}\prod_{k=1}^{K}(\varphi_i\theta_{ik})^{-1} = \prod_{i=1}^{I}\varphi_i^{-1}\prod_{i=1}^{I}\prod_{k=1}^{K}(\theta_{ik})^{-1}$$

Now, the joint posterior given in (2–9) can be written as

$$\begin{aligned}\pi(\boldsymbol{\eta},\boldsymbol{\alpha},\boldsymbol{\varphi},\boldsymbol{\theta},\boldsymbol{\beta}|\mathbf{y}) &\propto \prod_{i=1}^{I}\prod_{k=1}^{K}\Big\{\Big[\exp\{-\varphi_i\theta_{ik}\eta_i\}\{\varphi_i\theta_{ik}\eta_i\}^{y_{i0k}}\Big]\Big\}\\ &\times \prod_{i=1}^{I}\prod_{k=1}^{K}\Big\{\Big[\exp\{-\varphi_i\theta_{ik}\eta_i\alpha_i\exp(\boldsymbol{\beta}^T\boldsymbol{z}_k)\}\{\varphi_i\theta_{ik}\eta_i\alpha_i\exp(\boldsymbol{\beta}^T\boldsymbol{z}_k)\}^{y_{i1k}}\Big]\Big\}\\ &\times \Big\{\prod_{i=1}^{I}(\varphi_i^{-1}\prod_{k=1}^{K}\theta_{ik}^{-1})\Big\}\pi_1(\boldsymbol{\eta})\pi_2(\boldsymbol{\beta})\Big\{\prod_{i=1}^{I}\alpha_i^{-1}\Big\}.\end{aligned} \tag{2–11}$$

Again, note that

$$\begin{aligned}
&\int_0^\infty \exp\Big\{ -\eta_i\varphi_i\alpha_i\big[\sum_{k=1}^K \theta_{ik}\exp(\boldsymbol{\beta}^T\boldsymbol{z}_k)\big]\Big\}\alpha_i^{\sum_{k=1}^K y_{i1k}-1}d\alpha_i \\
\propto\ & (\eta_i\varphi_i)^{-\sum_{k=1}^K y_{i1k}}\Big\{\sum_{k=1}^K \theta_{ik}\exp(\boldsymbol{\beta}^T\boldsymbol{z}_k)\Big\}^{-\sum_{k=1}^K y_{i1k}}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\pi(\boldsymbol{\eta},\boldsymbol{\varphi},\boldsymbol{\theta},\boldsymbol{\beta}|\mathbf{y}) \ \propto\ & \prod_{i=1}^I\prod_{k=1}^K\Big\{\Big[\exp\{-\varphi_i\theta_{ik}\eta_i\}\{\varphi_i\theta_{ik}\eta_i\}^{y_{i0k}}\Big]\Big\} \\
\times\ & \prod_{i=1}^I\Big\{\Big[\sum_{k=1}^K \theta_{ik}\exp(\boldsymbol{\beta}^T\boldsymbol{z}_k)\Big]^{-\sum_{k=1}^K y_{i1k}}\prod_{k=1}^K\{\theta_{ik}\exp(\boldsymbol{\beta}^T\boldsymbol{z}_k)\}^{y_{i1k}}\Big\} \\
\times\ & \Big\{\prod_{i=1}^I(\varphi_i^{-1}\prod_{k=1}^K\theta_{ik}^{-1})\Big\}\pi_1(\boldsymbol{\eta})\pi_2(\boldsymbol{\beta}).
\end{aligned}$$

Integrating with respect to $\boldsymbol{\varphi}$, I have,

$$\begin{aligned}
\pi(\boldsymbol{\eta},\boldsymbol{\theta},\boldsymbol{\beta}|\mathbf{y}) \ \propto\ & \prod_{i=1}^I\Big[\Big(\sum_{k=1}^K\theta_{ik}\Big)^{-\sum_{k=1}^K y_{i0k}}\prod_{k=1}^K\theta_{ik}^{y_{i0k}}\prod_{k=1}^K\Big\{\frac{\theta_{ik}\exp(\boldsymbol{\beta}^T\boldsymbol{z}_k)}{\sum_{l=1}^K\theta_{il}\exp(\boldsymbol{\beta}^T\boldsymbol{z}_l)}\Big\}^{y_{i1k}}\Big] \\
\times\ & \pi_1(\boldsymbol{\eta})\pi_2(\boldsymbol{\beta})\Big\{\prod_{i=1}^I(\prod_{k=1}^K\theta_{ik}^{-1})\Big\}.
\end{aligned}$$

Then integrating with respect to $\boldsymbol{\eta}$, and rewriting L_R as given in (2–7) as

$$L_R=\prod_{i=1}^I\Big[\prod_{k=1}^K{\theta_{ik}}^{y_{i0k}}\prod_{k=1}^K\Big\{\frac{\theta_{ik}\exp(\boldsymbol{\beta}^T\boldsymbol{z}_k)}{\sum_{l=1}^K\theta_{il}\exp(\boldsymbol{\beta}^T\boldsymbol{z}_l)}\Big\}^{y_{i1k}}\Big],$$

and noting that $\sum_{k=1}^K\theta_{ik}=1$, I have,

$$\pi(\boldsymbol{\theta},\boldsymbol{\beta}|\mathbf{y}) \ \propto\ L_R\pi_2(\boldsymbol{\beta})\Big\{\prod_{i=1}^I\prod_{k=1}^K\theta_{ik}^{-1}\Big\}. \tag{2–12}$$

Since the order of integration of the joint posterior does not matter as long as the posterior is proper, comparing (2–10) and (2–12), it follows that after integrating the nuisance parameters, $\boldsymbol{\alpha}$ or $\boldsymbol{\theta}$, the posterior for $\boldsymbol{\beta}$ generated from L_P or L_R remains the same.

REMARK 2. Theorem 2 indicates that the marginal posterior distributions of $\boldsymbol{\beta}$ from either (2–6) or (2–7) are the same. Thus, in the presence of exposures missing completely at random, one may fit either the prospective or the retrospective model to stratified case-control data.

REMARK 3. A stratified case-control study without missing exposures is a special case, where $P(\Delta_{ij} = 1|S_i) = 1-P(\Delta_{ij} = 0|S_i) = 1$. A $1 : M$ individually matched case-control study is a special case of stratified case-control study with $s = 1$ and $t = M$ where M is a positive integer and the strata are defined as the matched sets. Note that we could very well assume that there are s_i cases and t_i controls in each stratum, and the proof will still carry through.

REMARK 4. It is interesting to note that the posterior $\pi(\boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\beta}|\mathbf{y})$ is non-identifiable in $\boldsymbol{\eta}$, in the sense of Dawid (1979), since $\pi(\boldsymbol{\eta}|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{y}) = \pi_1(\boldsymbol{\eta})$ which does not depend on $\mathbf{y}$. This, however, does not impede the propriety of the joint posterior as shown in the following theorem. For a general result relating non-identifiability with propriety or impropriety of posteriors, I refer to Ghosh *et al.* (2000).

The next theorem proves the propriety of the posterior under the assumed model under certain conditions. I need the notations $n_{dk} = \sum_{i=1}^{I} y_{idk}$, $d = 0, 1$.

THEOREM 3. Assume (i) $\sum_{k=1}^{K} y_{idk} \geq 1$ and (ii) $E[\exp\{(2d-1)(\boldsymbol{\beta}^T \sum_{k=1}^{K} \boldsymbol{z}_k n_{dk}\}] < \infty$ for $d = 0$ and $d = 1$. Here E denotes expectation with respect to the prior distribution on $\boldsymbol{\beta}$, namely, $\pi_2(\boldsymbol{\beta})$. Then $\pi(\boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\mathbf{y})$ is proper.

Proof. It suffices to show that $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{y})$ is proper. This amounts to showing $\int \prod_{i=1}^{I} I_i(\boldsymbol{\beta})\pi(\boldsymbol{\beta})d\boldsymbol{\beta} < \infty$, where

$$I_i(\boldsymbol{\beta}) = \int_{-\infty}^{\infty} \alpha_i^{-1} \prod_{k=1}^{K} \prod_{d=0}^{1} \Big\{ \frac{\alpha_i^d \exp(d\boldsymbol{\beta}^T \boldsymbol{z}_k)}{\sum_{l=0}^{1} \alpha_i^l \exp(l\boldsymbol{\beta}^T \boldsymbol{z}_k)} \Big\} d\alpha_i, \quad i = 1, \cdots, I.$$

Let $\omega_i = \exp(\alpha_i)$, $i = 1, \cdots, I$. Then

$$\begin{aligned} I_i(\boldsymbol{\beta}) &= \int_{-\infty}^{\infty} \prod_{k=1}^{K} \Big\{ \frac{1}{1+\exp(\omega_i + \boldsymbol{\beta}^T \boldsymbol{z}_k)} \Big\}^{y_{i0k}} \times \Big\{ \frac{\exp(\omega_i + \boldsymbol{\beta}^T \boldsymbol{z}_k)}{1+\exp(\omega_i + \boldsymbol{\beta}^T \boldsymbol{z}_k)} \Big\}^{y_{i1k}} d\omega_i \\ &< \Big\{ \int_0^{\infty} \exp(-\omega_i \sum_{k=1}^{K} y_{i0k}) d\omega_i \Big\} \exp(-\boldsymbol{\beta}^T \sum_{k=1}^{K} \boldsymbol{z}_k y_{i0k}) \\ &+ \Big\{ \int_{-\infty}^{0} \exp(\omega_i \sum_{k=1}^{K} y_{i1k}) d\omega_i \Big\} \exp(\boldsymbol{\beta}^T \sum_{k=1}^{K} \boldsymbol{z}_k y_{i1k}) \\ &< \exp(-\boldsymbol{\beta}^T \sum_{k=1}^{K} \boldsymbol{z}_k y_{i0k}) + \exp(\boldsymbol{\beta}^T \sum_{k=1}^{K} \boldsymbol{z}_k y_{i1k}), \end{aligned}$$

where the final step follows from assumption (i). Hence, by the inequality $(a+b)^{1+\delta} \leq 2^{\delta}(a^{1+\delta} + b^{1+\delta})$ for $\delta > 0$,

$$\begin{aligned} \prod_{i=1}^{I} I_i(\boldsymbol{\beta}) &< 2^{I-1} \Big\{ \exp(-\boldsymbol{\beta}^T \sum_{i=1}^{I} \sum_{k=1}^{K} \boldsymbol{z}_k y_{i0k}) + \exp(\boldsymbol{\beta}^T \sum_{i=1}^{I} \sum_{k=1}^{K} \boldsymbol{z}_k y_{i1k}) \Big\} \\ &= 2^{I-1} \Big\{ \exp(-\boldsymbol{\beta}^T \sum_{k=1}^{K} \boldsymbol{z}_k n_{0k}) + \exp(\boldsymbol{\beta}^T \sum_{k=1}^{K} \boldsymbol{z}_k n_{1k}) \Big\}. \end{aligned}$$

The proof is now completed by assumption (ii), which essentially requires the finiteness of the moment generating function corresponding to the prior distribution $\pi_2(\boldsymbol{\beta})$.

2.4 Discussion

As we know that, the MP transformation can simplify maximization of multinomial likelihood by considering a Poisson likelihood with additional parameters. Introducing some specific priors to the latent parameters, and arbitrary priors to the parameters of interest in the Poisson likelihood, I show that the marginal posterior distribution of the parameters of interest is exactly equivalent to that generated from the multinomial likelihood. However, the MP transformation requires categorical covariates. If some of them are continuous, the current practice is either to discretize them or follow the Bayesian bootstrap as proposed in Gustafson *et al.* (2002). An important open question is extension of the present results to continuous exposures.

CHAPTER 3
BAYESIAN MODELING FOR GENETIC ASSOCIATION IN CASE-CONTROL STUDIES: ACCOUNTING FOR UNKNOWN POPULATION SUBSTRUCTURE

3.1 Introduction

The evaluation of the association between molecular markers and disease status can be used to study the genetic basis of common human diseases (Risch and Merikangas, 1996; Morton and Collins, 1998; Sullivan *et al.*, 2001). The basic principle for such so-called association studies arises from the dependence of allele frequencies at marker loci upon those of disease variants, that is, the linkage disequilibria between alleles from different genetic loci. A significant association detected between a marker and the disease can be considered as evidence for close physical linkage between the marker and a disease locus, given that the linkage disequilibrium between any two genes always decays exponentially with their genetic distance in a random mating idealized population (Lynch and Walsh, 1998).

In practice, however, there rarely exists an idealized population as a result of the action of various evolutionary forces (Lynch and Walsh, 1998). Evolutionary forces, such as population structure and population admixture, operating on a population can result in spurious associations between a phenotype and markers that are not linked to any causative loci. The presence of spurious association suggests that the detected statistical association does not necessarily imply the physical linkage between the disease phenotype and arbitrary markers that have no physical linkage to causative loci (Lander and Schork, 1994). A classic example of spurious association caused by population substructure is presented in Knowler *et al.* (1988). In this study, based on a sample of Native Americans of the Pima and Papago tribes, a very strong negative association between the Gm haplotype Gm3;5,13,14 and type 2 or

non-insulin-dependent diabetes mellitus was detected. One might conclude from this observation that the absence of this haplotype, or the presence of a closely linked gene is a causal risk factor for the disease. However Gm3;5,13,14 is a marker for Caucasian admixture, and it is most likely that the presence of Caucasian alleles and decrease in Indian alleles led to lower susceptibility to type 2 diabetes, rather than the direct action of the haplotype or of a closely linked locus. This study demonstrates the effects of confounding due to population substructure, and the importance of considering genetic admixture while investigating the association between a disease and genetic markers.

In order to overcome the problem of spurious associations, many different genetic strategies have been proposed. Spielman *et al.* (1993) used the transmission disequilibrium test (TDT) to measure the association between a candidate gene and disease status by incorporating the genotypes of parents of affected individuals. This test has been instrumental in genetic association studies of human diseases (Spielman and Evens, 1998), but it is often limited because of difficulties with DNA sampling. For this reason, a simple case-control design that uses affected individuals and unrelated controls has recently received increased attention (Freedman *et al.*, 2004; Marchini *et al.*, 2004). A number of approaches have been developed to avoid the generation of spurious associations in case-control studies of disease-gene association. For a comprehensive recent review of admixture mapping for complex traits see McKeigue (2005).

Pritchard and colleagues used multilocus genotype data to estimate population substructure. They proposed a model-based clustering method to identify the population structure by genotyping samples at additional unlinked markers (Pritchard *et al.*, 2000a). This method was then extended to allow for the linkage between different markers (Falush *et al.*, 2003). A software package, STRUCTURE, has been

written to implement their algorithms that consider both linked and unlinked markers. Pritchard *et al.* (2000b) proposed a two-stage procedure in which first the population structure is inferred by employing the method of Pritchard *et al.* (2000a), and then the tests of association within subpopulations are conducted conditional on the imputed substructure. However, this method does not develop a model for the probability of disease incidence and cannot be generalized easily to provide estimates of the odds ratio corresponding to the genetic risk factor. Hoggart *et al.* (2003, 2004) developed a combination of Bayesian and classical approaches for association studies based on the admixture between populations with different ancestries. Apart from STRUCTURE, two other softwares which employ Bayesian ideas for statistical modeling of genetic data from admixed population are ADMIXMAP (Hoggart *et al.*, 2003, 2004) and ANCESTRYMAP (Patterson *et al.*, 2004).

Different from the above treatments, Satten *et al.* (2001) provided a novel latent-class analysis to study the association between the disease and the candidate genes based on a series of additional markers that are in linkage equilibrium with each other and with the candidate genes within subpopulations. Based on the Akaike information criterion (AIC), their method can estimate the number of subpopulations. But by either assuming the disease to be rare, or collapsing multiple genotypes into various binary genotypes, their method has not fully capitalized on the information about the multiple-genotype inheritance of the candidate gene.

In this chapter, I provide an alternative parametric Bayesian model for inferring on disease-gene association after accounting for population substructure. As in Satten *et al.* (2001), I use the latent-class approach to estimate the association parameters, while I account for the population substructure in a way similar to that of Pritchard *et al.* (2000a). However, unlike Satten *et al.* (2001), our analysis does not require the rare disease assumption or analyzing multicategory genotypes by several analyses using various possible binary genotypes of the candidate gene. Our model

can also handle multi-allelic genotypes of the candidate genes, extending on earlier approaches for the genotypic analysis of only biallelic loci. The computational strategy followed in Satten *et al.* (2001) involved use of the E-M algorithm to estimate the parameters in the model, combined with a parametric bootstrap strategy to obtain standard error estimates. The MCMC strategy designed in this chapter simplifies the computational complexity, with posterior standard deviation estimates and credible intervals being obtained from the random observations generated from the full conditional distributions of the parameters.

I should emphasize that in our Bayesian analysis, inference on the disease-gene association is not carried out on the basis of the particular imputed structure as done in Pritchard *et al.* (2000a). Instead, through use of model averaging (see for example, Madigan and Raftery (1994)), the association parameters are estimated by incorporating the uncertainty in estimating the substructure. In particular, instead of assuming the number of subpopulations I to be fixed, I put a prior on I and obtain the posterior distribution of I. For each possible value of I with positive posterior probability, I then estimate the association parameters in the disease-gene risk model. Finally I take the weighted average of these estimates, the weights being proportional to the posterior probabilities of the different values of I. The explicit model averaging formulas are given in Section 3.3.2. Our analysis thus combines the substructure estimation ideas of Pritchard *et al.* (2000a) using Bayesian clustering, and the latent class disease risk models of Satten *et al.* (2001) posed in a purely frequentist framework, through a more general unified Bayesian approach. This chapter presents a novel two-stage model with a clustering algorithm for inferring on cryptic population structure, followed by a logistic model for disease incidence, tied together through the technique of Bayesian model averaging.

The outline of the chapter is as follows. Section 3.2 states both the statistical model and the genetic model, and briefly introduces the methods in Pritchard *et al.*

(2000a) to estimate the number of subpopulations. Section 3.3 derives the underlying likelihood. I also introduce in this section the appropriate priors for the model parameters and obtain their estimates based on the posteriors. The posteriors are analytically intractable. So the Bayesian procedure is implemented by the MCMC numerical integration technique. In Section 3.4, I state the simulation strategy and provide results on simulated case-control studies under both a rare disease and a common disease assumption. The simulation studies are conducted in the same setting as in Satten *et al.* (2001) and mimic an admixed Argentinean population as described in Sala *et al.* (1998, 1999). Under the rare disease assumption, I compare our results with those obtained in Satten *et al.* (2001). In Section 3.5, I apply our methods to real data collected in a genetic association study with obesity as the disease outcome and the β2-adrenergic receptor (β2AR) as the candidate gene under investigation. Some concluding remarks are made in Section 3.6.

3.2 Model and Notation

3.2.1 Statistical Model

Let the binary variable D denote disease and let G be a (possibly vector-valued) genetic risk factor. I assume that the overall population of size N is comprised of I subpopulations, each having different frequencies of G and D. By the unmeasured covariate Z, I indicate the subpopulation to which an individual belongs. Thus, $D_j(= 1 \text{ or } 0)$ corresponds to the presence or absence of a disease for the jth individual with a genetic risk factor G_j, $j = 1, \cdots, N$.

I assume G_j to be a univariate discrete random variable, taking $M + 1$ values $g_0(= 0)$, g_1, $\cdots$, g_M. I assume that the prospective conditional logistic distribution for the disease status is

$$\Pr(D_j = 1 | G_j = g_m, Z = i) = H\{\beta_{0i} + \beta_{1m}\}, \qquad m = 0, \cdots, M, \qquad (3\text{–}1)$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$. Here β_{0i} is a term representing the subpopulation effect on the probability of disease for individuals belonging to a particular subpopulation i, and β_{1m} is the coefficient corresponding to the genetic exposure variable in the above logistic regression model. For parameter identifiability, I set $\beta_{10} = 0$. The method can immediately be extended to a vector valued genetic risk factor $\boldsymbol{G}_j$ for individual j.

3.2.2 Genetic Model

Since different subpopulations may have different frequencies of other marker genes, I use a latent-class approach to infer about the population substructure by using information on those additional marker loci. Consider x_l^c as the allele at marker l on chromosome c =1, 2 (labeling of the two chromosomes in a given pair as 1 or 2 is arbitrary) and let $X = (x_1^1, x_1^2, \cdots, x_L^1, x_L^2)$, where L is the number of marker loci under consideration.

First, I assume that the genes at the additional marker loci are unrelated to disease, that is

$$\Pr(D_j = 1|G_j, X_j, Z = i) = \Pr(D_j = 1|G_j, Z = i). \tag{3–2}$$

In the analysis that follows, I assume that Hardy-Weinberg equilibrium holds for each subpopulation. Human populations rarely show much divergence from the Hardy-Weinberg equilibrium once population substructure has been accounted for (Report of Committee on DNA Forensic Science 1996, pp. 104 and references cited therein).

Further, by choosing additional marker loci on different chromosomes from the chromosome where G is found, I first assume that the additional mutually independent marker genes are in linkage equilibrium with the candidate gene G, so that

$$\Pr(G_j, X_j|Z = i) = \Pr(G_j|Z = i) \times \Pr(X_j|Z = i). \tag{3–3}$$

By Hardy-Weinberg equilibrium,

$$\Pr(X_j|Z=i)=\prod_{l=1}^{L}\prod_{c=1}^{2}p_{lix_l^c}, \tag{3–4}$$

where $p_{lix_l^c}$ is the proportion of persons in subpopulation i having allele x_l^c at marker loci l, L being the number of marker loci.

Suppose the candidate gene G has w alleles, e.g., $a_1,\cdots,a_w$, and the frequency of the allele a_u $(u=1,\cdots,w)$ in the ith subpopulation is

$$\rho_{iu}=\Pr[G_l^c=a_u|Z=i].$$

Then by Hardy-Weinberg equilibrium the probabilities of the genotypes of G (a_ua_v) $(u,v=1,\cdots,w)$ are given by:

$$\Pr[G=a_ua_v|Z=i]=\begin{cases}\rho_{iu}^2, & u=v;\\ 2\rho_{iu}\rho_{iv}, & u\neq v.\end{cases} \tag{3–5}$$

3.2.3 Inference on I for The Model with Admixture

I consider the situation where I have multilocus genotype data from individuals sampled from a population with possibly unknown structure. Pritchard *et al.* (2000a) used the genotypes of a sample of individuals to identify the presence of population structure which is difficult to detect using visible characters, but may be significant in genetic terms. As Pritchard *et al.* (2000a) pointed out, the problem of inferring on the number of unknown populations, I, present in a data set is a very difficult task. In a Bayesian paradigm, with a suitably chosen prior distribution on I, one can base inference for I on the posterior distribution:

$$P(I|\mathbf{X})\propto P(\mathbf{X}|I)P(I), \tag{3–6}$$

where $\mathbf{X}$ denotes the vector of genotypes of the sampled individuals including the candidate gene G. Let $\mathbf{Z}$ denote the unknown population of origin of the individuals,

$\mathbf{P}$ denote the unknown allele frequency vector in all populations, and $\mathbf{Q}$ denote the vector of admixture proportions for each individual. The harmonic mean estimator is one of the simplest ways of estimating $P(\mathbf{X}|I)$,

$$\frac{1}{P(\mathbf{X}|I)} = \int \frac{P(\mathbf{Z},\mathbf{P},\mathbf{Q}|\mathbf{X},I)}{P(\mathbf{X}|\mathbf{Z},\mathbf{P},\mathbf{Q},I)} d\mathbf{Z} d\mathbf{P} d\mathbf{Q} \approx \frac{1}{K}\sum_{k=1}^{K} \frac{1}{P(\mathbf{X}|\mathbf{Z}^{(k)},\mathbf{P}^{(k)},\mathbf{Q}^{(k)},I)}. \quad (3\text{–}7)$$

However this estimator is notoriously unstable, often having infinite variance, and thus poses severe computational challenges. Pritchard *et al.* (2000a) described an alternative approach which is a more *ad hoc* but effective approach based on the Bayesian deviance function

$$DV(\mathbf{Z},\mathbf{P},\mathbf{Q}) = -2\log P(\mathbf{X}|\mathbf{Z},\mathbf{P},\mathbf{Q}). \quad (3\text{–}8)$$

Let $k = 1, 2, \cdots$ denote the k-th iteration in the Markov chain. One estimates the conditional mean and variance of the deviance function DV given $\mathbf{X}$ as follows:

$$\begin{aligned} E(DV(\mathbf{Z},\mathbf{P},\mathbf{Q})|\mathbf{X}) &\approx \frac{1}{K}\sum_{k=1}^{K} -2\log P(\mathbf{X}|\mathbf{Z}^{(k)},\mathbf{P}^{(k)},\mathbf{Q}^{(k)}) = \hat{\mu}, \\ Var(DV(\mathbf{Z},\mathbf{P},\mathbf{Q})|\mathbf{X}) &\approx \frac{1}{K}\sum_{k=1}^{K} (-2\log P(\mathbf{X}|\mathbf{Z}^{(k)},\mathbf{P}^{(k)},\mathbf{Q}^{(k)}) - \hat{\mu})^2 = \hat{\sigma}^2. \end{aligned}$$

By assuming that the conditional distribution of the deviance function DV given $\mathbf{X}$ is normal, it follows from (3–7) that

$$-2\log P(\mathbf{X}|I) \approx \hat{\mu} + \hat{\sigma}^2/4. \quad (3\text{–}9)$$

An analytical explanation of this approximation is provided in Appendix A. An alternative interpretation of this method is that model selection is based on penalizing the mean of the Bayesian deviance by a quarter of its variance. Pritchard *et al.* (2000a) pointed out that replacing the assumption of normality with the assumption of the Bayesian deviance function being distributed as a Gamma random variable

may be asymptotically more justifiable, but makes little or no difference in terms of estimation accuracy in practical applications.

One may use (3–9) to estimate $P(\mathbf{X}|I)$ for each I and then substitute the estimate into (3–6) to obtain approximate estimates of $P(I|\mathbf{X})$ (see Pritchard *et al.* 2000a, for a detailed algorithm). One would then impute the estimated substructure while conducting tests for disease-gene association. I will essentially follow the same technique for estimating $P(I|\mathbf{X})$ and embed the derived information into a disease risk model as described in the following section.

3.3 Likelihood and Priors

In this section, I derive the likelihood function, state the prior distributions and derive the posteriors. The key aspect of the modeling is in how I develop algorithms for estimating the model parameters and at the same time account for the population structure in our framework.

3.3.1 Likelihood

Because different subpopulations may have different frequencies of other marker genes, I make inference based on the marginal joint distribution of D, G and X, summing over all possible values of Z, the latent variate. Let $\Pr(Z = i) = q_i$, which is the proportion of persons in subpopulation i. Note that for subject j, G_j takes one of the values g_m, $m = 0, 1, \cdots, M$. By (3–3) and (3–4), for given I, the full likelihood L_I is factorized as follows:

$$\begin{aligned} L_I &= \prod_{j=1}^{N} \Pr(D_j, G_j, X_j) \\ &= \prod_{j=1}^{N} \sum_{i=1}^{I} \left[\Pr(Z = i) \times \Pr(G_j, X_j | Z = i) \times \Pr(D_j | G_j, Z = i) \right] \\ &= \prod_{j=1}^{N} \sum_{i=1}^{I} \left[q_i \times \left\{ \prod_{l=1}^{L} \prod_{c=1}^{2} p_{lix_l^c} \right\} \times \Pr(G_j = g_m | Z = i) \right. \\ &\quad \left. \times \frac{\exp\{D_j \times (\beta_{0i} + \beta_{1m})\}}{1 + \exp\{\beta_{0i} + \beta_{1m}\}} \right], \end{aligned} \tag{3–10}$$

where $\Pr(G_j|Z=i)$ is a function of ρ_{iu} $(u=1,\cdots,w)$ as described in (3–5), and L is the number of marker loci which are in linkage equilibrium with G.

I use a marginal likelihood rather than a conditional likelihood approach. The likelihood involves parameters of interest β_{1m} $(m=1,\cdots,M)$, and the nuisance parameters β_{0i}, ρ_{iu}, q_i and $p_{lix}(i=1,\cdots,I;\ \forall\ \ l$ and $\forall\ \ x)$, which grow in direct proportion to the number of subpopulations. This gives rise to the well known Neyman-Scott phenomenon where MLEs turn out to be inconsistent if I grows with sample size. Typically I deal with I between 1 through 7, and handling nuisance parameters is not a difficult issue in such scenarios. However, the marginal model does contain a large number of parameters, and I carry out Bayesian inference by introducing appropriate prior distributions for these parameters.

3.3.2 Priors and Posteriors

The main problem is to estimate the regression parameters β_{1m} $, m=1,\cdots,M$; I consider the following mutually independent normal priors:

$$\beta_{0i} \sim \text{Normal}\ (\mu_{\beta_{0i}}, \sigma^2_{\beta_{0i}}), \quad i=1,\cdots,I;$$
$$\beta_{1m} \sim \text{Normal}\ (\mu_{\beta_{1m}}, \sigma^2_{\beta_{1m}}), \quad m=1,\cdots,M.$$

When inferring the number of subpopulations I, I consider a discrete uniform prior on the domain of I. The priors for $\mathbf{P}$ and $\mathbf{Q}$ correspondingly are the following:

$$\begin{aligned}(q_1,\cdots,q_I) &\sim \text{Dirichlet}(\alpha,\cdots,\alpha);\\ \rho_{iu} &\sim \text{Beta}(a_i,b_i);\\ (p_{li1},p_{li2},\cdots,p_{liX_l}) &\sim \text{Dirichlet}(\lambda_{p_{li1}},\lambda_{p_{li2}},\cdots,\lambda_{p_{liX_l}}).\end{aligned}$$

With the above model and prior specifications, one can obtain the full conditional distributions for the parameters β_{0i}, β_{1m}, ρ_{iu}, q_i and p_{lix}. None of the conditionals has a standard distributional form.

For each given value of I, the parameters of interest can be estimated by generating random observations from the full conditionals using a MCMC numerical integration scheme and then taking averages of the generated observations. Corresponding to each value of I, I also have associated posterior probabilities $P(I|\mathbf{X})$ as discussed in Section 3.2.3. Therefore, by setting $\theta = (\beta_{11}, \cdots, \beta_{1M})$, using a model-averaging technique, any generic parameter θ is estimated by the posterior mean

$$E(\theta|\mathbf{X}) = \sum_i E(\theta|\mathbf{X}, I = i) \Pr(I = i|\mathbf{X}) \tag{3–11}$$

with posterior variance

$$\begin{aligned} V(\theta|\mathbf{X}) &= \sum_i V(\theta|\mathbf{X}, I = i) \Pr(I = i|\mathbf{X}) \\ &+ \sum_i [E(\theta|\mathbf{X}, I = i)]^2 \Pr(I = i|\mathbf{X}) - \Big[\sum_i E(\theta|\mathbf{X}, I = i) \Pr(I = i|\mathbf{X})\Big]^2. \end{aligned} \tag{3–12}$$

Thus the posterior variance estimates for the parameters of interest account for uncertainty in the estimation of I. The final point estimates are not byproducts of a single model with a fixed value of I, but averaged over possible models with weights proportional to the posterior probabilities $P(I|\mathbf{X})$.

3.3.3 Computational Details

1. Estimation of association parameters

None of the conditional distributions of the parameters has a standard distributional form and thus generating observations from the posterior distributions or calculating the posterior estimates is not automatic. I adopted a componentwise Metropolis-Hastings algorithm for each of the parameters.

Let η stand for a generic parameter, i.e., any of the β_{0i}, β_{1m}, ρ_{iu}, q_i and p_{lix} ($m = 1, 2$; $i = 1, \cdots, I$; $\forall\ \ l, x$). Let $L(\eta|\cdot)$ denote the full likelihood as given in (3–10) as a function of η given the data and all the other parameters. Let $\pi(\eta)$ be the

prior distribution on η. In order to simulate observations from the full conditional distribution of η, namely $\pi(\eta|\cdot)$, I proceed as follows.

Step 1: Start with any reasonable initial value of η, say η_0. This is the current value of η.

Step 2: Generate a new value of η, say η^*, from a candidate density $g(\eta)$.

Step 3: Replace η_0 by η^* with probability min $\left\{1, \frac{\pi(\eta^*|\cdot)g(\eta_0)}{\pi(\eta_0|\cdot)g(\eta^*)}\right\}$. Retain the existing value of η_0 otherwise. Note that $\pi(\eta|\cdot) \propto \pi(\eta)L(\eta|\cdot)$. If the candidate density $\pi(\eta) = g(\eta)$, then the acceptance probability reduces to (after cancelation of the prior term with the identical candidate density term) min $\left\{1, \frac{L(\eta^*|\cdot)}{L(\eta_0|\cdot)}\right\}$.

2. Inference of the number of subpopulations I

The following algorithm (Pritchard *et al.*, 2000a) is used to sample from $\Pr(\mathbf{Z}, \mathbf{P}, \mathbf{Q})$. Starting with initial values of $\mathbf{Z}^{(0)}$, iterate the following steps for $k = 1, 2, \cdots$

Step 1. Sample $\mathbf{P}^{(k)}$ and $\mathbf{Q}^{(k)}$ from $\Pr(\mathbf{P}, \mathbf{Q}|\mathbf{X}, \mathbf{Z}^{(k-1)})$;

Step 2. Sample $\mathbf{Z}^{(k)}$ from $\Pr(\mathbf{Z}|\mathbf{X}, \mathbf{P}^{(k)}, \mathbf{Q}^{(k)})$;

Step 3. Update α using Metropolis-Hastings step (where I consider a uniform(0,10) prior to α).

Step 2 may be performed by simulating $z_l^{(j,c)}$ (population of origin of allele copy $x_l^{(j,c)}$), independently for each j, c and l from

$$\Pr(z_l^{(j,c)} = i|\mathbf{X}, \mathbf{P}) = \frac{q_i^{(j)} \Pr(x_l^{(j,c)}|\mathbf{P}, z_l^{(j,c)} = i)}{\sum_{i'=1}^{I} q_{i'}^{(j)} \Pr(x_l^{(j,c)}|\mathbf{P}, z_l^{(j,c)} = i')}, \tag{3–13}$$

where $\Pr(x_l^{(j,c)}|\mathbf{P}, z_l^{(j,c)} = i) = p_{il x_l^{(j,c)}}$.

3.4 Simulation

To illustrate our approach, I consider a scenario similar to the one in Satten *et al.* (2001) with an admixture of European and American Indian ancestry in Argentinean population. Sala *et al.* (1998, 1999) published allele frequency data on twelve short tandem repeat (STR) loci in Argentineans of European ancestry, as well as in three Argentinean American Indian aboriginal groups (Mapuche, Tehuelche,

and Wichi) (Table 3–1). The Metropolitan population of Buenos Aires was studied and the population did not exhibit any significant difference from Hardy-Weinberg equilibrium. However, the STR allele frequency distributions are characterized by significant differences within and also between different populations. I assume that Argentinean Europeans constituted 70% of a hypothetical target population and that each American Indian group constituted 10%.

I simulate a population such that all eleven additional mutually independent STR loci are in linkage equilibrium with the candidate gene for persons in the same subpopulation. Simulated data sets are constructed by using reasonable true values of the parameters. Specifically, by using the allele frequencies from Sala *et al.* (1999), I generate data on the candidate gene and other marker loci in a population that comprises four subpopulations. As in Satten *et al.* (2001), I select allele 3 of locus D6S366 as the disease-causing allele, with frequencies 0.277, 0.341, 0.446 and 0.557 in European, Mapuche, Tehuelche, and Wichi, respectively. Consider a biallelic candidate gene, i.e., a candidate gene with two alleles A (the disease-causing allele) and a (the non-disease-causing allele). The candidate gene G has 3 possible genotypes g_0, g_1 and g_2 corresponding to persons having zero (aa), one (Aa) and two (AA) copies of a disease-causing allele. If the frequency of the disease-causing allele in the ith subpopulation is

$$\rho_i = \Pr[G_l^c = A | Z = i] = 1 - \Pr[G_l^c = a | Z = i], \tag{3–14}$$

then by Hardy-Weinberg equilibrium, the probabilities of the genotypes of G are as the follows:

$$\begin{aligned} \Pr[G = g_0 | Z = i] &= (1-\rho_i)^2; \\ \Pr[G = g_1 | Z = i] &= 2(1-\rho_i)\rho_i; \\ \Pr[G = g_2 | Z = i] &= \rho_i^2. \end{aligned} \tag{3–15}$$

Finally, the disease status data that vary with changing frequencies of the disease-causing allele for each subpopulation are generated. As stated in Satten *et al.* (2001), persons who were homozygous for the disease-causing allele had an increased risk of disease corresponding to a log-odds ratio of 1.0 (relative risk = exp(1.0) = 2.72); and persons who were heterozygous for the disease-causing allele had no increase in risk. This implies, in our notation, $\beta_{11} = 0$ and $\beta_{12} = 1.0$. The log odds of the rare disease (which implies that the control population mimics the whole population, and $\Pr(G = g_m|D = 0, Z = i) \approx \Pr(G = g_m|Z = i)$) among persons with zero or one copy of the disease-causing allele was -5, -4, -3 and -3 in the European, Mapuche, Tehuelche, and Wichi populations, respectively. For the common disease with a higher prevalence rate, I assume that the log odds among persons with zero or one copy of the disease-causing allele was -2, -1.5, -1 and -1 in the European, Mapuche, Tehuelche, and Wichi populations, respectively.

The results I presented are based on a set of diffuse and mutually independent priors. I use $N(0, 9)$ prior on β_{0i} and β_{1m}, $Beta(0.5, 0.5)$ on ρ_i and a symmetric Dirichlet prior for the allele frequency parameters with all λ's being 0.5. For $(q_1, \cdots, q_I)$, I choose a Dirichlet$(\alpha, \cdots, \alpha)$ prior, with a U(0,10) hyperprior on α.

For each scenario, I generated 100 different data sets and obtained the parameter estimates by computing the model averaged posterior means for each simulated data set. In each replication of our simulation, I generated data for 125 (250) cases and 125 (250) controls from the above simulation strategy, followed by sampling the cases and controls from a larger random sample of subjects. For each replication, I ran multiple Markov chains, typically with $20000 - 30000$ iterations. The posterior means calculated for each replication were based on every tenth observation of the last 5000 observations in each chain, combined together to reduce auto-correlation. An estimate of the posterior variance was calculated based on the aggregate of the last 5000 values for each replication. I report average values for these quantities over

the 100 replications. I also calculated an estimate of the mean squared error (MSE) corresponding to the estimates of each of the parameters of interest (say θ in general) based on the 100 replications. I considered this MSE, i.e., the squared deviations of the estimates from the true parameter, averaged over the 100 replications as a measure of performance of our method.

$$\text{MSE} = \frac{1}{100}\sum_{r=1}^{100}(\text{Posterior mean of } \theta \text{ in } r\text{-th replication} - \text{True value of } \theta)^2.$$

To examine the effect of the number of STR loci on the estimators, I analyzed the datasets with 250 subjects (125 cases and 125 controls) by (i) using all the additional loci and (ii) only the first six additional loci. These two scenarios are labeled as X12 and X6 in Tables 3–2 and 3–4 respectively. By applying the methods stated in Section 3–2 (Pritchard *et al.* 2000a) and introducing a uniform prior for I ($I \in \{1, 2, 3, 4\}$), for each simulated dataset, first I obtain estimates of $P(I|\mathbf{X})$. For example, by (i), I obtain $P(I = 3|\mathbf{X}) = 0.2$ and $P(I = 4|\mathbf{X}) = 0.8$. Then the model averaged estimate of I is 0.2× 3+ 0.8× 4=3.8. The estimates of the association parameters are computed following (3–11) and (3–12). For the same dataset, the estimate of β_{12} is 1.09 for $I = 3$ and 1.02 for $I = 4$, thus the final model averaged estimate of β_{12} for that dataset is $1.09 \times 0.2 + 1.02 \times 0.8 = 1.034$. The results in Table 3–2 are obtained by averaging these estimates over the 100 simulated datasets, which shows that the posterior standard deviations of our model averaged estimates are typically smaller than the standard errors furnished by Satten *et al.* (2001) (I include the relevant numbers from Tables 3-2 and 3-3 of Satten *et al.* (2001) directly in Table 3–3 of the current chapter). I realize that though our simulation settings are the same as of Satten *et al.* (2001), the two sets of estimates may not be exactly comparable as the two methods are not implemented on identical datasets, but still this might serve as a precursor for comparison purposes. Satten *et al.* (2001) do not provide MSE for their estimates over the replications. As a result I cannot compare the two procedures

directly in terms of the MSE. As one might expect, when I increased the sample size to 500 (250 cases and 250 controls), adequate performance is achieved even with just the first six STR loci and the overall pattern of the results remain the same.

I also include the naive analysis completely ignoring additional multilocus information (denoted as X0 in Tables 3–2 and 3–3). One can note that the estimation results are much inferior if one ignores the genotypic information at a series of additional unlinked marker loci.

To show that the methods are not limited to the assumptions that either the disease is rare or the genotypes G are binary, I also analyzed a simulated dataset with 250 subjects (125 cases and 125 controls) and another with 500 subjects (250 cases and 250 controls) where the disease has a higher prevalence rate. The overall pattern of the results are fairly similar to the rare disease case. I note relatively smaller MSE's and posterior standard deviations for this common disease case as compared to the rare disease case. The results are presented in Table 3–4.

For analyzing the simulated data, I used the implicit prior belief that the source population may have 4 or less subpopulations, by putting a discrete uniform prior on $1, 2, 3, 4$ for I. However, I have also tried to put non-zero probability on a value of I greater than the true simulation value of 4, for instance, a discrete uniform prior on $1, \cdots, 8$. In this case, the estimates of the regression parameters β_{1m} appear to change very little even when I is estimated to be slightly greater than the true value used to generate the data (results are not provided). Pritchard *et al.* (2004a) note that for situations where several values of I give similar estimates of $\log \Pr(\mathbf{X}|I)$, it is often the case that the smallest of these is 'correct'. In the practical implementation, I adopt a model selection perspective and try to obtain the smallest value of I that captures the major structure in the data.

3.5 Application to A Real Dataset

To illustrate our method, I apply our approach to explore genetic association of obesity and the β2AR candidate gene (for details of the study, please see Lin *et al.*, 2005). The β-adrenergic receptors (βAR) are known to play an important role in cardiovascular function and in response to drug. I analyze complete data on 144 men and women who participated in this study and ignore the observations with missingness. Each of the participating subjects were genotyped for SNP markers at codon 16 within the β2AR gene, at codon 389 within the β1AR gene and at codon 492 within the α1A gene. The phenotypic information collected are weight and height of individuals, by which the body mass index (BMI) of each subject can be calculated. I define "obese", i.e, $D = 1$ when BMI ≥ 30.0, and $D = 0$ otherwise. This leads to 85 undiseased and 59 diseased subjects in the dataset I consider.

Previous studies have detected possible association between polymorphism in the β2AR gene and obesity, the focus being particularly on codon 16 and codon 27 substitutions, but no association has been detected within β1AR gene or α1A gene (Johnson and Terra 2002, Lin *et al.* 2005, Takami *et al.* 1999). Therefore, I consider the β2AR gene as the candidate gene, denoted by G and the β1AR gene and the α1A gene as two other genes unrelated with the disease, denoted by $\mathbf{X} = (X_1, X_2)$. Note that in this dataset, I only have the genotypic information regarding single polymorphisms in these three genes which have biallelic genotypes, generally expressed as $x = 0, 1, 2$. So the expression in (3–4) will be changed as $P(\mathbf{X}|Z = i) = \prod_{l=1}^{2} p_{lix}$, where p_{lix} is the proportion of persons in subpopulation i having genotype x ($x = 0, 1, 2$) corresponding to gene l.

I analyzed the data by considering genotypic information on all three genes (denoted by "X2+G") and by only the candidate gene (denoted by "X0+G"). Since in the real data, I do not know the true value of I, I should try to estimate the smallest value of I that captures the major substructure in the data, if any. To this end, I

introduce a discrete uniform prior on $1, 2, \cdots, 15$ for I. I consider $(p_{li1}, p_{li2}, \cdots, p_{liI}) \sim$ Dirichlet$(0.5, 0.5, \cdots, 0.5)$, and for $(q_1, \cdots, q_I)$, I choose a Dirichlet$(\alpha, \cdots, \alpha)$ prior with a uniform hyperprior on α with range from 0 to 10. By applying the methods stated in Section 3.2, I first obtain inference on I. The principal findings are that with the inclusion of the two other genes, I detect some evidence of substructure with an estimate of I, as $\hat{I} = 3$, with $P(I = 3|\mathbf{X}) = 1$, whereas without these two genes and by only using G, I obtain $P(I = 1|\mathbf{X}) = 1$, implying $\hat{I} = 1$, i.e., no population substructure can be detected in the source population. In fact, the data came from a North American population with diverse ethnic composition of blacks, whites and others, so one could expect some latent population substructure in this data. The results of our analysis are presented in Table 3–5. In all the methods of analysis, the genetic factor does not appear to be a statistically significant risk factor. The results suggest that codon 16 (Arg16Gly) polymorphisms of the β2AR gene is not a major contributing factor to obesity for this studied population. In fact, in Swedish Caucasians, Gln27Glu polymorphism at codon 27 of the β2AR gene was shown to be associated with obesity, but no such association was shown for Arg16Gly polymorphism at codon 16. None of the Gln27Glu and Arg16Gly polymorphisms of the β2AR gene were found to be a major contributing factor to obesity in Japanese men (Hayakawa *et al.* 2000). In the ordinary logistic regression model, with G as a categorical factor, I also find insignificance of G, (P-values 0.8591 and 0.1571 corresponding to G=1 and 2, respectively). Even after accounting for information in the other genes and population substructure, the effect of the candidate gene remains insignificant. Notice that the Bayesian HPD intervals are wider than the ordinary logistic model due to addition of layer of uncertainty on I.

3.6 Discussion

In this chapter, I present an alternative Bayesian model for accounting for population substructure in genetic association studies. As compared to previous approaches, our model is advantageous in terms of the following aspects. First, it can estimate the number of subpopulations (I) that comprise the overall population. Although Satten *et al.* (2001) can also provide such an estimate, their approach is based on the grid procedure in which multiple different I's are fitted and the optimal one is then determined in terms of the minimum AIC. On the other hand, Pritchard *et al.* (2000b) estimated substructure and then conducted tests based on the imputed substructure. Based on marker and candidate gene information, our model estimates the posterior probabilities of I, which is then used in forming the final estimates of the relative risk parameters through model averaging. An additional advantage is that, unlike Satten *et al.*'s (2001) approach, our model does not rely on the assumption of the rare disease or the collapsing of multiple genotypes into binary genotypes, thus offers more power to study the genetic architecture of any type of diseases.

A new feature of the Bayesian analysis is the use of model averaging to estimate the regression coefficients. Rather than relying on one particular model with a fixed number of strata I, I have put a prior on I, and have estimated the regression parameters as the weighted average of their estimates for different values of I. The weights are proportional to the posterior probabilities of the different values of I. Thus I embed the substructure estimation together with inference on the association parameters in a unified Bayesian framework. The standard error of the relative risk estimates does incorporate the uncertainty in the estimation of I as reflected in (3-14). This is unlike the method proposed in Pritchard *et al.* (2000b) where the substructure is estimated first and tests are conducted based on the imputed substructure. Table 3–2 shows that our methods are comparable to those of Satten *et al.* (2001); however, since our set-up is different from that of Pritchard *et al.* (2000b),

it is hard to compare the two methods directly in numerical sense. In principle, I do believe that combining inferences of the substructure and association modeling will lend one more power in detecting association.

It should be pointed out that fewer additional markers are needed when the sample size is large. When additional marker loci are involved, the number of nuisance parameters (the allele frequencies of those loci for each subpopulation) in the model would increase, requiring more data to estimate them properly.

There remains the problem of handling marker loci in linkage disequilibrium with the candidate gene in our framework. According to Falush *et al.* (2003), there are three sources of linkage disequilibria (LD), mixture LD, admixture LD and background LD. The mixture LD arises from variation in individuals' ancestry and it can be measured by unlinked markers. The admixture LD occurs because of the correlation in ancestry among an extended genomic region. The background LD decays on a short scale and, therefore, occurs within a fine chromosomal structure. Pritchard *et al.* (2000a) modeled the mixture LD for association studies. In their "linkage" model, Falush *et al.* (2003) incorporated the "admixture LD" into the inference of population structure. The incorporation of the background LD is an interesting open question.

In summary, I have derived flexible Bayesian estimation techniques for disease-gene association in case-control studies by accounting for population structure. First, I applied Pritchard *et al.*'s (2000a) methods to infer population structure (i.e. estimating $P(I|\mathbf{X})$ and I) by using the genotypes of sampled individuals at a series of unlinked markers. Second, I propose a latent variable approach to estimate the association parameters, and account for population substructure using additional marker loci information as in Satten *et al.* (2001). The final results are calculated by the model averaging technique (as described in (3–11) and (3–12)) which combines inferences from the above two steps. Estimation results based on a simulated admixed

population (mimicking the results presented in Sala *et al.* (1998)) show that the estimates of the relative risk parameters using additional mutilocus genetic information are superior to those when such information is not exploited. I also apply our method to a real dataset on obesity. This chapter illustrates how the modeling tool of Bayesian model averaging can be effectively used to conduct posterior inference in an interesting application in human genetics.

Table 3–1: Allele frequencies for Twelve STR loci in the four Argentinean subpopulations.

Locus	Argentinian Europeans	Mapuche	Tehuelche	Wichi
D6S366	0.082	0.091	0.143	0
	0.204	0.114	0.071	0
	0.277	0.341	0.446	0.557
	0.119	0.136	0.036	0.086
	0.091	0.125	0.036	0.029
	0.183	0.159	0.143	0.200
	0.028	0.011	0.018	0.071
	0.015	0.023	0.107	0.057
FABP	0.589	0.683	0.732	0.485
	0.110	0.058	0.107	0.162
	0.300	0.260	0.161	0.353
CSF1PO	0.330	0.266	0.339	0.226
	0.313	0.282	0.232	0.194
	0.298	0.367	0.411	0.581
	0.059	0.085	0.018	0
F13A	0.151	0.222	0.357	0.173
	0.060	0.122	0.125	0.077
	0.202	0.122	0.054	0.346
	0.209	0.178	0.143	0.115
	0.325	0.344	0.304	0.288
	0.053	0.011	0.017	0
FESFPS	0.260	0.170	0.143	0.257
	0.420	0.500	0.714	0.543
	0.247	0.284	0.107	0.043
	0.073	0.045	0.036	0.157
THO1	0.233	0.526	0.286	0.132
	0.250	0.298	0.429	0.721
	0.105	0.009	0.018	0
	0.185	0.026	0.089	0.015
	0.226	0.140	0.179	0.132
HPRTB	0.032	0	0	0
	0.179	0.032	0.091	0
	0.317	0.323	0.227	0.357
	0.285	0.403	0.591	0.167
	0.137	0.242	0.091	0.357
	0.050	0	0	0.119
VWA	0.063	0.0096	0.036	0.014
	0.099	0.077	0.054	0.014
	0.294	0.577	0.429	0.514
	0.297	0.125	0.214	0.343
	0.246	0.212	0.268	0.114
D13S317	0.090	0.020	0	0
	0.160	0.240	0.15	0.464
	0.060	0.070	0.05	0.179
	0.290	0.120	0.15	0.089
	0.250	0.260	0.3	0.089
	0.100	0.180	0.225	0.179
	0.040	0.110	0.125	0
D7S820	0.156	0.070	0.050	0
	0.115	0.050	0.050	0.070
	0.276	0.220	0.175	0.125
	0.245	0.420	0.525	0.450
	0.159	0.210	0.200	0.250
	0.046	0.030	0	0.105
D16S539	0.156	0.110	0.225	0.125
	0.100	0.130	0.075	0.232
	0.294	0.240	0.100	0.321
	0.252	0.370	0.550	0.250
	0.195	0.150	0.050	0.071
RENA4	0.772	0.728	0.881	0.690
	0.074	0.229	0.023	0
	0.153	0.041	0.095	0.310

Cited from Sala *et al.* (1998) and Satten *et al.* (2001).

Table 3–2: The results of simulated rare-disease data with marker loci in linkage equilibrium with the candidate gene D6S366. Ratio of the sample sizes of cases to controls is 125/125 and 250/250. X12 and X6, represent that the parameters were estimated by using the twelve and the first six additional marker loci, respectively. X0 is the analysis without using any additional marker loci. Mean and posterior standard deviation refer to the average of the Bayes estimates and posterior standard deviations obtained in 100 replications, whereas MSE is the estimated mean squared error based on 100 replications.

Sample size	Model		β_{11}	β_{12}	I
		True value	0.0000	1.0000	4
125/125	X12	Mean	-0.0475	1.1093	3.8178
		MSE	0.1497	0.0765	0.1802
		Post. std. dev.	0.3126	0.2638	0.3854
	X6	Mean	-0.1095	1.1028	3.6403
		MSE	0.2005	0.0986	0.3540
		Post. std. dev.	0.3277	0.3127	0.4763
	X0	Mean	-0.3380	0.8855	4.0000
		MSE	1.2277	0.4982	
		Post. std. dev.	1.5982	1.0677	
250/250	X12	Mean	0.0005	1.0966	3.7873
		MSE	0.0546	0.0551	0.2107
		Post. std. dev.	0.2704	0.1592	0.4089
	X6	Mean	0.0051	1.1035	3.5415
		MSE	0.0631	0.0582	0.4572
		Post. std. dev.	0.3127	0.1952	0.4994
	X0	Mean	-0.2766	0.9489	4.0000
		MSE	1.2603	0.4330	
		Post. std. dev.	1.4152	0.9236	

Table 3–3: The results of simulated rare-disease data with marker loci in linkage equilibrium with the candidate gene D6S366 which are analyzed by Satten *et al.* (2001). 125/125 and 250/250 denote ratio of the sample sizes of cases to controls. X12 and X6 represent that the parameters were estimated by using the twelve and the first six of the additional marker loci, respectively. Mean and standard error refer to the average of the estimates and standard errors obtained in 500 replications.

Sample Size	Model		β_{11}	β_{12}	I
		True value	0.000	1.000	4
125/125	X12	Mean	0.061	1.006	3.53
		Std. err.	0.293	0.453	0.76
	X6	Mean	0.023	0.883	3.32
		Std. err.	0.865	1.718	0.69
	Crude Analysis*	Mean	0.366	1.760	1.00
		Std. err.	0.285	0.370	
250/250	X6	Mean	0.023	0.962	3.37
		Std. err.	0.226	0.394	0.61

* Ignore stratification and analyze data without additional marker loci.

Table 3–4: The results of simulated common-disease data with marker loci in linkage equilibrium with the candidate gene D6S366. Ratio of the sample sizes of cases to controls is 125/125 and 250/250. X12 and X6, represent that the parameters were estimated by using the twelve and the first six additional marker loci, respectively. X0 is the analysis without using any additional marker loci. Mean and posterior standard deviation refer to the average of the Bayes estimates and posterior standard deviations obtained in 100 replications, whereas MSE is the estimated mean squared error based on 100 replications.

Sample size	Model		β_{11}	β_{12}	I
		True value	0.0000	1.0000	4
125/125	X12	Mean	-0.0062	1.1116	3.8492
		MSE	0.1106	0.1005	0.1456
		Post. std. dev.	0.3152	0.1607	0.3523
	X6	Mean	0.0017	1.1299	3.6279
		MSE	0.1173	0.1371	0.3634
		Post. std. dev.	0.3488	0.2766	0.4766
250/250	X12	Mean	0.0023	1.0928	3.9331
		MSE	0.0600	0.0551	0.0461
		Post. std. dev.	0.2165	0.1806	0.2412
	X6	Mean	0.0191	1.1051	3.6228
		MSE	0.0408	0.0470	0.3748
		Post. std. dev.	0.2627	0.1991	0.4846

Table 3–5: The results of real data analysis with the posterior mean (Estimate), posterior standard deviation and 95% highest posterior density (HPD) interval (MLE and confidence interval (CI) for the ordinary logistic regression model).

Model		β_{11}	β_{12}	I
X2+G	Estimate	-0.0895	0.7165	3
	Post std.dev.	0.3997	0.5201	$\cdot^*$
	HPD	(-0.8619,0.6831)	(-0.2996,1.7259)	
X0+G	Estimate	-0.1206	0.7433	1
	Post std.dev.	0.4515	0.5602	$\cdot^*$
	HPD	(-1.0028,0.7865)	(-0.3339,1.8303)	
Ordinary logistic regression	Estimate	-0.0668	0.7143	
with only G as covariate	Std.err.	0.3765	0.5048	
	CI	(-0.8047,0.6711)	(-0.2751,1.7037)	

*:All of the posterior probability concentrated on a single value of I, thus we are unable to obtain estimates of posterior variance.

CHAPTER 4
SEMIPARAMETRIC BAYESIAN ANALYSIS OF CASE-CONTROL DATA UNDER GENE-ENVIRONMENT INDEPENDENCE AND POPULATION STRATIFICATION

4.1 Introduction

Except for some rare diseases, such as Huntington or Tay Sachs disease which may be the result of a deficiency of a single gene product, most common human diseases have a multifactorial etiology involving complex interplay of many genetic and environmental factors. By identifying and characterizing such complicated gene-environment interactions, one has more opportunities to study etiology, diagnosis, prognosis and treatment of complex diseases.

The case-control study design, where sampling is conditional on the presence or absence of disease, is a powerful epidemiologic tool for studying potential risk factors of rare diseases. It has been established that prospective logistic regression analysis of case-control data is "efficient" in the modern semiparametric sense with respect to the underlying covariate density model (Breslow *et al.*, 2000). A special aspect of the gene-environment association problem is that it may often be reasonable to assume that a subject's genetic susceptibility is independent of the environmental exposure. Consequently, one may be able to obtain more efficient estimation techniques than the traditional logistic regression, by exploiting the additional gene-environment independence restriction instead of an unconstrained covariate density model.

Piegorsch *et al.* (1994) first observed that one can estimate multiplicative gene-environment interactions in logistic models with data from cases alone, provided that the environmental factor (E) and the genetic factor (G) are independent in the population and the disease is rare. The interaction parameter is obtained as the odds

ratio between G and E among cases only. They also noted that the estimate of the G-E interaction parameter from case-only data is more efficient than its counterpart obtained from case-control data using logistic regression.

However, methods that use G-E independence produce severely biased estimates if the assumption is violated (Schmidt and Schaid, 1999; Albert *et al.*, 2001). Non-independence is less likely to occur when the environmental exposure is external (pollution, pesticide or radio-active substance) or a randomized treatment in a clinical trial. One has to be much more cautious with the independence assumption when considering behavioral risk factors and metabolic polymorphisms which could alter an individual's behavior. Gatto *et al.* (2004) discuss several such potential sources of non-independence. In fact, genetic susceptibility factors and environmental exposures, though unlikely to be causally related at an individual level may be correlated at a population level due to their dependence on other variables that stratify the population, such as age, ethnicity, family history and alike. For example, a woman with a strong family history of breast cancer is more likely to carry BRCA1/2 (two major genes identified for breast and ovarian cancer) mutation and knowing her family history, less likely to use post-menopausal hormones. This may result in a negative association between BRCA1/2 mutation and hormone use. In such instances, G-E independence does not hold marginally, but may hold when conditioned on the stratification variables (for instance, family history). Modeling stratification effects can thus be viewed as a possible remedy to guard against resultant bias due to violation of the G-E independence assumption. One of the major goals of the current chapter is to develop techniques to model stratification effects in a flexible, data-adaptive way in an estimation framework which exploits *conditional* G-E independence.

The use of G-E independence through case-only studies has mainly been for estimating the gene-environment interaction parameter. Khoury and Flanders (1996)

noted that neither the genetic nor environmental exposure main effect can be estimated with case data only. Umbach and Weinberg (1997) showed that with data available on both cases and controls, one can estimate the main effects and interaction by fitting a suitably constrained log-linear model under a rare disease assumption. In a population based case-control study of ovarian cancer of Jewish women in Israel, Modan *et al.* (2001) argued that under gene-environment independence and the rare disease assumption, the disease odds ratio associated with E among subjects with genotype $G = g$ can be estimated by a logistic regression analysis that compares $P(E|D = 0)$ with $P(E|D = 1, G = g)$. However, the method proposed in Modan *et al.* (2001) also does not allow for the estimation of all main effects of interest. Most of the above methods consider very simple settings and it is not immediate how to exploit G-E independence in the presence of population stratification as a direct extension of these methods.

Chatterjee and Carroll (2005) (referred to as CC in rest of the text) propose a semiparametric maximum likelihood method of estimation of *all* the logistic regression parameters. They exploit the G-E independence assumption and use data from both cases and controls. Their method addresses many of the limitations of the existing methods as discussed above. CC derive a robust profile-likelihood based estimation technique which does not require the rare disease assumption. They also consider the issue of population stratification and propose a method when G-E independence assumption only holds conditional on the set of stratification variables (S). CC consider a logistic disease probability model for $P(D|G, E, S)$. They proceed to work with the joint retrospective likelihood of the form $P(G, E, S|D)$, factorized as,

$$P(G, E, S|D) = \frac{P(D|G, E, S)P(G|E, S)P(E, S)}{\sum_{\text{G,E, and S}} P(D|G, E, S)P(G|E, S)P(E, S)}.$$

Under the assumption of G-E independence conditional on S, the second factor on the right hand side reduces to $P(G|E, S) = P(G|S)$ and thus it remains to model

$P(E, S)$ and $P(G|S)$. CC leave the joint distribution of the environmental exposure and the stratification variables, $P(E, S)$ to be fully non-parametric. However, they model $P(G|S)$ in a parametric way, by assuming a logistic regression model with S as covariate. As we will note, the parametric logistic model for the $P(G|S)$ is often inadequate, especially for a genetic mutation which is rarely detected in healthy controls but commonly prevalent in the case population. In such circumstances, the estimation, especially of the main effect due to G, suffers in the method proposed by CC. To overcome this problem, I use a factorization of the partially retrospective likelihood $P(G, E|D, S)$ that allows us to model the genotype frequencies separately in the case and the control population. Moreover, for genetic mutations like the BRCA1/2, there are several genetic risk models (Antoniou *et al.*, 2004) as well as empirical data (Risch *et al.*, 2001; Couch *et al.*, 1997) which predict population mutation frequencies after adjusting for covariates like family history and ancestry. A flexible Bayesian model can incorporate this accumulated scientific evidence in the form of a prior distribution assigned to $P(G|S)$ and lead to more accurate estimation than a logistic model for carrier probabilities. To elicit this advantage of the Bayesian paradigm while estimating all the parameters in the *G-E* logistic regression model, and *not just G-E* interaction, remains another primary goal of this chapter. The dataset I use is a replica of the one that CC use, based on a case-control study on ovarian cancer patients in Israel (Modan *et al.*, 2001). I consider presence of mutation of BRCA1/2 as the genetic risk factor and number of years of oral contraceptive (OC) use and parity as the environmental exposures. The stratification variables I consider are age group, ethnicity, personal history of breast cancer (PHB) and family history of breast and ovarian cancer (FHBO). I model the control distribution of the continuous environmental exposures conditional on S as a Dirichlet process mixture of normals (DPM). The DPM model is appealing in this context as it provides a natural measure of the degree of stratification and is model-robust. I also present a

parametric Bayesian alternative for comparison purposes. An extensive simulation study providing an in-depth comparison of the proposed Bayesian methods with the powerful estimation techniques provided by CC, the case-only method and ordinary logistic regression is a very important feature of this chapter. The simulation explores several scenarios, with changing distributions for G and E as well as under violation of the G-E independence assumption even when conditioned on observable confounders.

It appears that under G-E independence, the proposed semiparametric Bayesian method has a real advantage over the competing methods under any of the following situations (i) the individual genotype frequencies in each stratum do not follow the logistic multiplicative odds model in terms of stratification variables, (ii) the genetic mutation is rare in the control population and is commonly prevalent in the case population. The gain is significant when the number of strata defined by S is relatively large. When the G-E independence assumption even when conditional on S fails, all the methods which use this assumption perform poorly, least so for the Bayesian semiparametric method, which is more robust to model changes.

The rest of this chapter is organized as follows. In Section 4.2 I present the model, likelihood, priors and posteriors. Section 4.3 contains analysis of the Israeli ovarian cancer data. Section 4.4 presents the details of our simulation study and the results. Section 4.5 contains concluding discussion, while proofs and computational details are relegated to Appendix B.

4.2 Model, Likelihood, Priors and Posteriors

Consider a case-control study with n subjects , n_1 cases and n_0 controls. Let D be the binary disease variable, i.e., $D_j = 1$ if the j^{th} subject is a case, and $D_j = 0$ if the subject is a control. The genetic risk factor G is essentially the genotype at a single locus within a candidate gene. I will consider G as a categorical variable with $M + 1$ levels, namely $g_0, \cdots, g_M$. In addition, the data are assumed to be stratified based on some other covariates, say $\boldsymbol{S}$. I consider the following logistic regression

function to model the disease probability in terms of G, E and $\boldsymbol{S}$,

$$P(D=1|G,E,\boldsymbol{S}) = H\{\beta_0(\boldsymbol{S}) + \sum_{m=0}^{M} I(G=g_m)\beta_{1m} + \beta_2 E + E\sum_{m=0}^{M}\beta_{3m} I(G=g_m)\}, \tag{4–1}$$

where $H(u) = \{1+\exp(-u)\}^{-1}$. The intercepts $\beta_0(\boldsymbol{S})$ capture stratification effects due to the covariates $\boldsymbol{S}$ on the risk of disease. Let $\boldsymbol{\beta}_1 = (\beta_{10}, \cdots, \beta_{1M})$, β_2, and $\boldsymbol{\beta}_3 = (\beta_{30}, \cdots, \beta_{3M})$ represent the main effect of the genetic factor, the main effect of the environmental factor, and their interaction effect respectively. For parameter identifiability, I set $\beta_{10} = 0$ and $\beta_{30} = 0$. For simplicity, I present my model with only one continuous environmental exposure. Extension to multiple continuous exposures E is straightforward and one such analysis is presented in Section 4.3. Extension of the methodology when E is a set of categorical exposures or a mixed set of continuous and categorical exposures is indicated later in this section.

As I continue to compare and contrast our methods with CC and traditional logistic regression, I would first like to point out that each method is based on a different likelihood, the CC method uses a fully retrospective likelihood, $P(G,E,S|D)$, the traditional logistic model uses a fully prospective likelihood, $P(D|G,E,S)$, whereas our method uses the following partially retrospective likelihood $P(G,E|D,S)$ factorized as

$$L_R = \prod_{j=1}^{n} P(G_j, E_j|\boldsymbol{S}_j, D_j) \;=\; \prod_{j=1}^{n}\left[P(G_j|E_j,\boldsymbol{S}_j,D_j)P(E_j|\boldsymbol{S}_j,D_j)\right]. \tag{4–2}$$

As illustrated in Prentice and Pyke (1979) and discussed again in Roeder *et al.* (1996) and Müller and Roeder (1997), the form of the retrospective likelihood considered here is compatible with the logistic form of the prospective likelihood. Evaluation of the likelihood function (4–2) requires the conditional distribution of $[G|E,\boldsymbol{S},D]$ and the conditional distribution of $[E|\boldsymbol{S},D]$. I will make the following assumption:

Assumption 1: Conditional on $\boldsymbol{S}$, G and E are independent in the control population, i.e., $P(G|D=0,E,\boldsymbol{S}) = P(G|D=0,\boldsymbol{S})$.

When the disease is rare in each stratum, and the control population mimics the entire population, the usual G-E independence assumption in source population, i.e., $P(G|E, \boldsymbol{S}) = P(G|\boldsymbol{S})$ is approximately equivalent to Assumption 1. The two assumptions of G-E independence in source population and rare disease are made by Piergorsch *et al.* (1994), Umbach and Weinberg (1997) and Modan *et al.* (2001), while CC do not need the rare disease assumption. Our analysis is exact under Assumption 1 which may hold even when the disease is not rare. As pointed out in Schmidt and Schaid (1999), the rare disease assumption is quite subtle and may not hold, for example in situations where the disease risk is much higher for the carriers of a particular gene mutation or for certain strata of the population. In the dataset I consider, the risk of ovarian cancer is known to be higher for BRCA1/2 carriers and for subjects with family history of breast or ovarian cancer. Fortunately, the bias due to the rare disease assumption has less impact when the overall disease prevalence $P(D = 1)$ is small, even with highly penetrant genes (Schmidt and Schaid, 1999).

I do recognize that directly verifying Assumption 1 empirically could be quite difficult based on the given study at hand, as tests of independence will have little power. Many researchers have considered this issue of verifying G-E independence in control population in the context of using this as a screening tool to validate the use of case-only analysis (Albert *et al.*, 2001). Sensitivity analysis shows that the G-E association pattern in controls reflect G-E association in source population when baseline disease risk is less than 0.1% (Gatto *et al.*, 2004). To address this issue, in the simulations, I do consider various departures from Assumption 1, and the performance of all the methods under violation of this assumption. I advocate that when substantial uncertainty remains on the validity of the independence assumption, statistically significant results based on the proposed methods should be treated as precursors for high priority investigations for future epidemiologic studies.

Assuming that the first n_0 observations are controls and the next $n - n_0$ observations are cases, under Assumption 1, the retrospective likelihood in (4–2) reduces to

$$\begin{aligned} L_R &= \prod_{j=1}^{n_0} \left[P(G_j|\boldsymbol{S}_j, D_j = 0)P(E_j|\boldsymbol{S}_j, D_j = 0)\right] \\ &\times \prod_{j=n_0+1}^{n} \left[P(G_j|E_j, \boldsymbol{S}_j, D_j = 1)P(E_j|\boldsymbol{S}_j, D_j = 1)\right]. \end{aligned}$$

Consequently, to evaluate the likelihood contributed from control data I will need to specify probability models for $P(G|\boldsymbol{S}, D = 0)$ and $P(E|\boldsymbol{S}, D = 0)$. Following the technique first suggested by Satten and Kupper (1993), I present the following Lemmas which will then furnish expressions for $P(G|\boldsymbol{S}, E, D = 1)$ and $P(E|\boldsymbol{S}, D = 1)$, once having the control distributions and the prospective model as in (4–1).

Lemma 1:

$$\frac{P(G = g_m|E, \boldsymbol{S}, D = 1)}{P(G = g_m|E, \boldsymbol{S}, D = 0)} = \frac{P(D = 1|G = g_m, E, \boldsymbol{S})/P(D = 0|G = g_m, E, \boldsymbol{S})}{P(D = 1|E, \boldsymbol{S})/P(D = 0|E, \boldsymbol{S})}.$$

Lemma 2:

$$\frac{P(D = 1|E, \boldsymbol{S})}{P(D = 0|E, \boldsymbol{S})} = \sum_{m=0}^{M} \frac{P(D = 1|G = g_m, E, \boldsymbol{S})}{P(D = 0|G = g_m, E, \boldsymbol{S})} P(G = g_m|D = 0, E, \boldsymbol{S}).$$

Lemma 3:

$$\frac{P(E|\boldsymbol{S}, D = 1)}{P(E|\boldsymbol{S}, D = 0)} = \frac{P(D = 1|E, \boldsymbol{S})/P(D = 0|E, \boldsymbol{S})}{\int \frac{P(D = 1|E, \boldsymbol{S})}{P(D = 0|E, \boldsymbol{S})} P(E|\boldsymbol{S}, D = 0)dE}.$$

The proofs of the Lemmas are collected in Appendix B.

Remark 1: With the likelihood conditional on $\boldsymbol{S}$, I do not intend to estimate the relative risks due to the stratification variables $\boldsymbol{S}$ and focus only on the parameter of interest $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \beta_2, \boldsymbol{\beta}_3)$. As I proceed, I note that under our formulation, I would tacitly avoid direct estimation of the stratum specific intercept parameters $\beta_0(\boldsymbol{S})$ which appears in the disease risk model (4–1).

Before describing the estimation theory, I first would like to address the identifiability of the parameters in the prospective model (4–1) and the retrospective likelihood L_R. As stated in Prentice and Pyke (1979), if there are no assumptions made on the covariate distribution $\mathcal{H}(g, e|\boldsymbol{s}) = P(G = g, E = e|\boldsymbol{S} = \boldsymbol{s})$ neither $\mathcal{H}(\cdot, \cdot|\boldsymbol{s})$, nor $\beta_0(\boldsymbol{S})$ is identifiable. But $\boldsymbol{\beta}$ is always identifiable under any choice of $\mathcal{H}$. Following Lemma 1 of Roeder *et al.* (1996) it can be easily shown that under Assumption 1 on the covariate density, $\boldsymbol{\beta}$ remains identifiable in our likelihood L_R.

Remark 2: I would like to point out that unlike the Prentice-Pyke result for general nonparametric covariate density case, with an additional independence restriction on $\mathcal{H}(\cdot, \cdot|\boldsymbol{s})$ in the *source* population (not *just* in *control* population as stated in Assumption 1), Lemma 1 of CC proves that both the intercept and the covariate distributions are identifiable given $\boldsymbol{S} = \boldsymbol{s}$. For a rare disease, Assumption 1 is approximately equivalent to independence in the source population. Thus, in the rare disease case, with our formulation, by Lemma 1 of CC, I do have identifiability of the entire likelihood, not just of $\boldsymbol{\beta}$.

I consider the stratification variables $\boldsymbol{S}$ as a vector of $q \geq 1$ categorical covariates, with the k^{th} variable having r_k categories or levels. Therefore, the level combinations of $\boldsymbol{S}$ defines $I = \prod_{k=1}^{q} r_k$ possible strata. For instance, in the Israeli ovarian cancer data I consider $q = 4$ stratification variables: (Age group, ethnicity, PHB, FHBO), the first three having two categories each and FHBO having three categories. Therefore $\boldsymbol{S}$ defines $I = 2 \times 2 \times 2 \times 3 = 24$ possible strata. For ease of notation, I will introduce Z, a single index variable with I possible values, each value representing a distinct stratum. So for subject j, Z_j can take exactly one of the values $1, \cdots, I$, completely determined by the observed values of the stratification variables for subject j, namely $\boldsymbol{S}_j$. I can now rewrite the likelihood L_R after replacing $\boldsymbol{S}_j$ by the stratum membership

indicator of subject j, namely Z_j.

$$L_R = \prod_{j=1}^{n_0} \Big[P(G_j|Z_j, D_j = 0) P(E_j|Z_j, D_j = 0) \Big] \times \prod_{j=n_0+1}^{n} \Big[P(G_j|E_j, Z_j, D_j = 1) P(E_j|Z_j, D_j = 1) \Big]. \quad (4\text{–}3)$$

I consider the following model for the control distribution of the genetic factor in stratum i,

$$\log \frac{P(G = g_m|Z = i, D = 0)}{P(G = g_0|Z = i, D = 0)} = \gamma_{im}.\ m = 1, \cdots, M. \quad (4\text{–}4)$$

Note that $\gamma_{i0} = 0$. The above model does not assume any stringent parametric form for $P(G|D = 0, \boldsymbol{S})$ in terms of $\boldsymbol{S}$ and simply treats the probabilities in each stratum to be the model parameters, allowing complete distributional flexibility.

Result 1: Using (4–1), (4–4) and Lemma 1, I obtain the case distribution of G as:

$$P(G = g_m|E, Z = i, D = 1) = \frac{\exp\{\beta_{1m} + \beta_{3m}E + \gamma_{im}\}}{1 + \sum_{k=1}^{M} \exp\{\beta_{1k} + \beta_{3k}E + \gamma_{ik}\}}, \quad m = 1, \cdots, M. \quad (4\text{–}5)$$

Proof of Result 1 is presented in Appendix B. Note that although in the control population by virtue of the independence assumption, $P(G|E, D = 0, Z = i) = P(G|D = 0, Z = i)$, in the case population $P(G|E, D = 1, Z = i)$ does depend on E.

Due to high dimensional nature of the stratification variables $\boldsymbol{S}$, it is often hard to model the effect of $\boldsymbol{S}$ on the distribution of the exposure variable (E) explicitly. I consider a flexible nonparametric Bayesian approach to model the distribution $\big[E|D = 0, Z = i\big]$ which allows for possible stratification effects on the distribution of E and does so in a data adaptive way. I consider the case when E is continuous, as in the data example. The Dirichlet process mixture model (DPM) with a normal

kernel can be expressed in the following hierarchical structure

$$\begin{aligned} \big[E|D=0,Z=i\big] &\sim N(\mu_i,\sigma_i^2) \\ \boldsymbol{\theta}_i=(\mu_i,\sigma_i^2)|\ \mathcal{P} &\sim \mathcal{P} \\ \mathcal{P} &\sim \mathrm{DP}(\alpha\mathcal{P}_0), \end{aligned} \tag{4–6}$$

where $\mathcal{P}$, serving as a prior on the $\boldsymbol{\theta}_i$, $i=1,\cdots,I$, is itself a *random* probability measure. I assume that $\mathcal{P}$ is realization of a Dirichlet process (DP) with scalar precision parameter $\alpha \geq 0$ and base measure (or base prior) $E[\mathcal{P}]=\mathcal{P}_0$ which is a bivariate CDF on $\mathcal{R}\times\mathcal{R}^+$. A property of the DP prior is that the random probability measure $\mathcal{P}$ is almost surely discrete, leading to the following properties which reinterpret the DPM model structure (see Antoniak, 1974 and Sethuraman, 1994 for details):

1. Any realization of $\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_I$ generated from $\mathcal{P}$ lies in a set of $K(\leq I)$ distinct values, denoted by $\boldsymbol{\omega}=\{\boldsymbol{\omega}_1,\cdots,\boldsymbol{\omega}_K\}$;
2. $\boldsymbol{\omega}_l, (l=1\cdots,K)$ are a random sample from the base prior $\mathcal{P}_0$;
3. $K(\leq I)$ is drawn from an implicitly determined prior distribution depending on the precision parameter α and I;
4. Given $K\leq I$, the I values are selected from the set $\boldsymbol{\omega}$ according to a uniform multinomial distribution.

The above discussion is conditional on α and the hyperparameters which determine $\mathcal{P}_0$.

With this hierarchical mixture prior structure for the control distribution of E and the prospective logistic model (4–1) it now remains to investigate the nature of the case-distribution of E. The following result provides an answer.

Result 2: Assume that the $\boldsymbol{\theta}_i$ take values $\boldsymbol{\omega}_l$ from the set $\boldsymbol{\omega}$ as described in 1. Then

$$\big[E|Z=i,D=1,\boldsymbol{\theta}_i=\boldsymbol{\omega}_l\big]=\sum_{m=0}^{M} p_{ilm}\phi(E;\boldsymbol{\omega}^*_{lm}), \tag{4–7}$$

where $\phi(\cdot;\boldsymbol{\theta})$ denotes the normal density with parameter vector $\boldsymbol{\theta}$, $\boldsymbol{\omega}_l = (\mu_l, \sigma_l^2)$, say, and $\boldsymbol{\omega}^*_{lm} = (\mu_l + \beta_2\sigma_l^2 + \beta_{3m}\sigma_l^2, \sigma_l^2)$ and $p_{ilm} = \exp(\beta_{1m} + (\mu_l + \beta_2\sigma_l^2 + \beta_{3m}\sigma_l^2)^2/(2\sigma_l^2) + \gamma_{im})/\sum_{k=0}^{M}\exp(\beta_{1k} + (\mu_l + \beta_2\sigma_l^2 + \beta_{3k}\sigma_l^2)^2/(2\sigma_l^2) + \gamma_{ik})$. Hence, the distribution of E among the case population conditional on all other parameters is again a DP mixture but not with a normal kernel but with a mixture kernel given by (4–7).

The exact expression of the likelihood (4–3) and proof of Result 2 is deferred to Appendix B. I will refer to this model for E as EDPM for future references.

Prior Structure: The likelihood (4–3) involves the association parameters $\boldsymbol{\beta}_1$, β_2, $\boldsymbol{\beta}_3$, and $\gamma_{i1}, \cdots, \gamma_{iM}$, and $\boldsymbol{\theta}_i = (\mu_i, \sigma_i^2)$, $i = 1, \cdots, I$. I use independent normal priors for all the association parameters and also on γ_{im}'s, $m = 1, \cdots, M$. I will note in the real data example (with only two possible values of G, so that $m = 0, 1$) that if we know a priori that the mutation is rare in the control population, and have an established genetic risk model for $P(G|\boldsymbol{S})$, we should select an informative prior on γ_{i1}, so that the effective range of the carrier probabilities in the control/case population for each stratum reflects the scientific guesses for these values.

It now remains to describe the hierarchical prior structure involved in the DPM model. Note that the mean of the random probability measure $\mathcal{P}$ is $\mathcal{P}_0$ which is a bivariate distribution, and I consider the following standard normal-inverse Gamma structure, namely, under $\mathcal{P}_0$, $\mu_i|\ \sigma_i^2 \sim N(m_0, \tau\sigma_i^2)$, $(\sigma_i^2) \sim \text{IG}(s/2, S/2)$. For computation, I used a $N(\mu_{m_0}, \sigma^2_{m_0})$ prior on m_0, which adds an extra layer of uncertainty in $\mathcal{P}_0$. I use Inverse Gamma (IG)$(a_\tau/2, b_\tau/2)$ prior on τ. Lastly, following Escobar and West (1995), I assume a Gamma(a_α, b_α) prior on the precision parameter α. I choose the prior parameters (a_α, b_α) in such a way that the mean of the prior distribution of K is reasonably large (compared to I) and the variance is modest. Choosing such a "further" prior is suggested in West *et al.* (1994).

None of the full conditional distributions follows a standard distributional form and posterior inference is made by using the MCMC numerical integration technique.

Conditional on $\boldsymbol{\theta}_i$, drawing random numbers from the respective conditional distributions is straightforward application of the Metropolis-Hastings algorithm. To update $\boldsymbol{\theta}_i$, I use the no gaps algorithm prescribed by MacEachern and Müller (1998). I describe the computational details of the algorithm in Appendix B.

Remark 3: An interesting feature of the EDPM model is that it selects K, the number of distinct values in I realizations from $\mathcal{P}$ or the cardinality of the set $\boldsymbol{\omega}$ in a data adaptive way depending on the extent of stratification on the distribution of the environmental exposure. In the presence of strong stratification effects, all of the $\boldsymbol{\omega}_l$ could be distinct, i.e. $K = I$; in the complete absence of stratification effects, $K = 1$. Typically K will lie somewhere in between. The posterior mode of K thus serves as an indicator of the degree of stratification effects on the control distribution of E.

In the above discussion, I assumed $\boldsymbol{S}$ to be a set of categorical stratification variables which is most often the case. If any of the stratification variables is continuous, I recommend categorizing them for implementing the EDPM model.

Remark 4: Since I assume the distribution of E to be a Dirichlet Mixture of normals, this model applies only to continuous environmental exposures. The model can be easily extended to multiple continuous exposure, simply by taking a DPM model with multivariate normal kernel (as used by Müller and Roeder (1997)). I illustrate this multivariate extension in one segment of the real data analysis. For categorical exposures, the models could be adapted as shown in Seaman and Richardson (2001) by using a Dirichlet distribution as prior on the probabilities for each category. For a mixture of discrete and continuous environmental exposures one could either categorize the continuous exposure into classes or adapt the Bayesian bootstrap ideas as described in Gustafson *et al.* (2002). The main theme is common between all three methods, trying to model the distribution of the environmental exposure in a non-parametric way to guard against violations of model assumptions.

Remark 5: Note that, as indicated in Remark 1, via the above formulation, the nuisance parameter $\beta_0(\boldsymbol{S})$ does not present itself in the case distributions of G and E as presented, respectively. $\beta_0(\boldsymbol{S})$ appears as a common term in both the numerator and denominator of (4–5) and (4–7) and thus gets canceled in the ratio. Hence, the retrospective likelihood does not involve $\beta_0(\boldsymbol{S})$.

Remark 6: One could naturally think of the following parametric logistic model for modeling the distribution of G, instead of using the more flexible model as given in (4–4)

$$\log \frac{P(G_j = g_m | D_j = 0, \boldsymbol{S}_j)}{P(G_j = g_0 | D_j = 0, \boldsymbol{S}_j)} = \nu_0 + \boldsymbol{\nu}_m^T \boldsymbol{S}_j, \ m = 1, \cdots, M, \tag{4–8}$$

where $\boldsymbol{\nu}_m$ is a vector of regression parameters capturing the effect of stratification variables on the incidence of the genetic susceptibility factor in the control population. CC assume a similar logistic model for $P(G|S)$ for their real data analysis, though they recognize that it is hard to predict BRCA1/2 carrier probabilities using this logistic structure. Indeed, when I based my inference using the model in (4–8) with normal priors on ν_0 and $\boldsymbol{\nu}_m$, the estimates of the parameters of interest $\boldsymbol{\beta}$ were less accurate when compared to the ones using (4–4). Thus, for the sake of brevity I only include results where I used model (4–4) for carrier probabilities.

4.3 The Israeli Ovarian Cancer Data

In this section, I apply the proposed methodology to the data from a population-based case-control study on all ovarian cancer patients identified in Israel between March 1, 1994 and June 30, 1999 (Modan *et al.*, 2001). Blood samples were collected from the cases and the controls in order to test for the presence of mutation in the two major breast and ovarian cancer susceptibility genes BRCA1 and BRCA2. In addition, the subjects were interviewed to collect data on reproductive/gynecological history such as parity, number of years of OC use and gynecological surgery. The main goal of the study was to examine the interplay of the BRCA1/2 genes and known reproductive/gynecological risk factors of ovarian cancer. Since the actual

data had confidentiality issues, a replica was generated by replacing only the original genetic susceptibility factor by a simulated binary genetic risk factor, retaining all the features as in the original dataset. The dataset I used contained 832 cases and 747 controls.

This is a real example where OC use and BRCA1/2 mutation may appear to be correlated simply because both could be related to the stratification variables $\boldsymbol{S}$ like age and family history, and it is more realistic to assume independence between these two genetic and gynecological risk factor conditional on $\boldsymbol{S}$. However, it is hard to verify Assumption 1 based on this single dataset as only 7 out of the 747 controls were BRCA1/2 carriers. I ran a logistic regression of G on the exposures of interest E in the controls in each stratum, and though the tests of association were insignificant, the sparsity of the data makes the results of these tests for association unstable and less reliable. However, Modan *et al.* (2001, page 236) and Chatterjee and Carroll (2005) both indicate that it is reasonable to assume that carrier status is independent of the exposures under consideration, namely parity and number of years of OC use, and I also employ this assumption in the analysis.

It is known that the risk of ovarian cancer is higher for certain strata (for example for the subgroup with family history of both breast and ovarian cancer) as well as for BRCA1/2 carriers so the rare disease assumption may not hold for all levels of the genetic factor or for certain subgroups. However, Modan *et al.* (2001) reported only 1326 cases of epithelial ovarian cancer during the five-year study period with a baseline population of approximately 1.5 million, suggesting an empirical estimate of disease prevalence $P(D = 1) = 8.7 \times 10^{-4}$, suggesting that the odds-ratio estimates obtained through the analysis under Assumption 1 will provide adequate approximations to the ones obtained via exact analysis using G-E independence in source population.

All analyses are carried out conditional on four stratification variables: $\boldsymbol{S}$=(Age group (=0 if age $<$ 50 years and =1 if age $\geq$ 50 years), ethnicity (=1 for Ashkenazi

Jews and 0 otherwise), presence or absence of a personal history of breast cancer PHB (= 1 if present and 0 if absent) and a family history of breast or ovarian cancer FHBO (= 0 if no history, 1 if one breast cancer case in family and 2 if ovarian cancer or two or more breast cancer cases in the family)). So the total number of strata defined by the level combinations of $\boldsymbol{S}$ is $I = 24$.

I analyze the data using the EDPM method as described in the previous section. For modeling the distribution of the genetic factor, I use (4–4). The genetic factor G is binary with $G = 0$ for absence of any BRCA1/2 mutation and $G = 1$ for carrying at least one BRCA1/2 mutation. It is well known that BRCA1/2 mutations are very rare among ovarian cancer controls, and as Modan *et al.* (2001) pointed out, traditional logistic regression analysis would yield imprecise estimates of parameters of interest. Compounding to the sparsity is the fact that I do have a relatively large number of strata defined by $\boldsymbol{S}$ and as a result, estimation of genotype frequencies individually in each stratum would be imprecise in a classical set-up. CC adopt a parametric logistic model for $P(G|\boldsymbol{S})$ to circumvent this problem which is also not satisfactory. In a Bayesian paradigm I effectively use the prior knowledge on BRCA1/2 carrier probabilities in different age groups, ancestry and with varying levels of family history based on genetic algorithms (BRCAPRO: Parmigiani *et al.*, 1998; BOADICEA: Antoniou *et al.* 2004) and empirical data (Couch *et al.*, 1997, Risch *et al.*, 2001). I allow uncertainty in these predictions by allowing the informative prior on γ_{i1} to vary around the scientific guesses and in this process relax the stringent logistic assumption. The effective range of prior probabilities for $P(G = 1|\boldsymbol{S}, D = 0)$ typically varied from 10^{-1} to 10^{-4} across different strata.

I present two analyses, the first with OC use as the only environmental exposure (E) as a direct illustration of the methods formulated in Section 4.2. With a binary G, there are three parameters of interest involved in the disease risk model (4–1):

$\beta_{11} = \beta_1$, β_2, and $\beta_{31} = \beta_3$.

$$\text{logit}P(D = 1|G, E, \boldsymbol{S}) = \beta_0(\boldsymbol{S}) + \beta_1 I[G = 1] + \beta_2 OC + \beta_3 I[G = 1] * OC$$

For each of β_1, β_2, β_3, I use $N(0, 16)$ prior. Since scientific theory suggests high positive value of β_1, one could also select a sharper prior for β_1. For the EDPM model as described in (4–6), under the base-measure $\mathcal{P}_0$, I assume that the variance component $\sigma^2 \sim IG(2, 1)$ and $\mu|\sigma^2 \sim N(m_0, \tau\sigma^2)$. The exposure variable, number of years of OC use typically ranges from 0 to 20 years. I chose a diffuse prior on m_0, namely, $m_0 \sim N(3, 9)$. I use $IG(3, 1)$ prior on τ. Choosing priors for α is a challenging task as α has the dual role of capturing the degree of faith in the base measure, as well as determining the number of distinct values of $\boldsymbol{\theta}$. As prescribed by Escobar and West (1995) I choose a Gamma prior on α which allows for prior probabilities for larger values of $K \leq I = 24$. I experimented with various choices of the shape and scale parameters of the Gamma prior, and the results are presented for Gamma$(4, 1)$ prior on α. Detailed algorithm for resampling from the full conditionals is collected in the appendix.

For comparison purposes, I also analyzed this data with a parametric model, largely targeted towards this dataset. As the data contain 832 cases and 747 controls of which 678 cases and 586 controls did not use oral-contraceptives at all, I used a zero-inflated model (EZIM) for the control distribution of OC use. For individual j, I consider p_j as the probability of non-exposure ($E_j = 0$), and with probability $(1-p_j)$, the exposure values follow $N(\mu_j, \sigma^2)$, where $\mu_j = \delta_0 + \boldsymbol{\delta}_1^T \boldsymbol{S}_j$. The mixing probabilities are also modeled through the four observed stratification factors, $\text{logit}(p_j) = \eta_0 + \boldsymbol{\eta}_1^T \boldsymbol{S}_j$. The case distributions can be obtained as mixture distributions via Lemmas 1-3. For the EZIM model, I consider mutually independent N$(0, 16)$ prior for the regression parameters, β_1, β_2 and β_3, as well as on δ_0, η_0 and each component of $\boldsymbol{\delta}_1$

and $\boldsymbol{\eta}_1$. For the scale parameter σ^2, I use $IG(2,1)$ prior. Posterior inference is again based on MCMC samples from the full conditional distribution of the parameters.

I analyzed this data through the method proposed by CC and the case-only method after adjusting for the covariates $\boldsymbol{S}$. The case-only method only furnishes estimate of the BRCA1/2*OC interaction parameter β_3. The results are presented in Table 4–1. There is little in the way of differences for estimation of β_2 and β_3 by all the four methods which use G-E independence. But for estimating the main effect of G as measured by β_1, the Bayesian methods have much smaller posterior standard deviation and narrower HPD interval compared to the standard error and the CI for the estimate of β_1 in the CC method. The results indicate that standard logistic assumption is less likely to hold for $P(G|\boldsymbol{S})$ in this dataset, and the more flexible model for G as given in (4–4), boosted with the scientifically validated priors, and adapting itself more naturally to the features of the data. Interestingly, the non-parametric EDPM model for OC use performs quite comparably with the parametric zero-inflated model which is designed specifically to capture the distribution of OC use.

I also analyzed the data by ordinary logistic regression analysis which does not exploit G-E independence in any manner. The wider confidence intervals, especially for the interaction parameter indicates that any method using G-E independence is able to estimate the interaction parameter more precisely. Whereas all the other four methods declare G-E interaction to be statistically significant, the ordinary logistic model cannot detect significance.

In summarizing the results, I first observe that for women who never used OC (E=0), there is an almost astronomic increase in risk of ovarian cancer for a BRCA1/2 mutation carrier. The estimated odds ratio by the EDPM method is $\exp(3.75) = 42.52$. On the other hand, among non-carriers, longer use of OC is related to decrease in disease risk with associated odds ratio $\exp(-0.0748) = 0.92$. However, the estimate

of the interaction parameter β_3 suggests that among BRCA1/2 carriers, the risk of ovarian cancer increases slightly with OC use, with an odds ratio $\exp(-0.0748) \times \exp(0.1091) = 1.03$. The precision estimates and the credible intervals all indicate that the main effect of BRCA1/2 and the BRCA-OC interaction are statistically significant whereas the main effect of OC use is only marginally significant.

Figures 4–1 and 4–2 present plots of posterior distribution for β_1, β_2, β_3 and also for α and K for the EDPM method. To explore the degree of stratification, I also present a plot of $var(\mu_i)$ and $var(\sigma_i)$ in the EDPM model ($i = 1, \cdots, 24$). I notice that the μ_i's and σ_i's do reflect variation in the values, the variability in σ_i being greater. The posterior mode of K is at 5 with posterior mean around 5.76, suggesting that though there are 24 possible strata, not all of them have distinct effects on the distribution of number of years of OC use.

I present another analysis with OC and parity both considered as environmental exposures, I omit the details corresponding to this analysis and only collect the results in Table 4–2. I note that for women with parity=0, and OC=0, BRCA1/2 mutation is associated with a huge increase in risk of ovarian cancer. Among BRCA1/2 non-carriers, higher parity is associated with decreased risk of ovarian cancer. The parity*BRCA1/2 interaction estimate suggests that the decrease in risk of ovarian cancer associated with increased parity is modestly larger for carriers than for non-carriers, but this difference is not statistically significant.

Since for a real dataset, the true state of the parameters is unknown and it is not really possible to compare the methods, I conduct an extensive simulation study to assess the performances of the methods over a range of different scenarios and provide recommendations for the practitioner.

4.4 Simulation

In order to simulate a dataset for comparing the Bayesian methods along with the method proposed by CC, case-only analysis and ordinary logistic regression, I used

the ovarian cancer data as a prototype to elicit realistic true values of the parameters. I set the true values close to the results I obtained in the analysis of real data by EDPM method in Table 4–1, i.e., $\beta_1 = 3$, $\beta_2 = -0.07$, and $\beta_3 = 0.12$. I generated 1500 observations following the scheme as below:

1. I started with generating the $\mathbf{S}$ = (Age group, Ethnicity, PHB, FHBO) from a multinomial distribution, where the stratum probabilities are consistent with the real study.

2. Given $\boldsymbol{S}$, I generated a binary variable D representing the disease status, with probabilities $P(D = 1|\mathbf{S})$ in agreement with the ovarian cancer study, the marginal disease probability in the generated population being around 0.1%. I also experimented with several other choices of $P(D = 1)$ for which the results are not included.

3. I generated the environmental exposure E from two distributions:

(i) A zero-inflated model exactly mimicking the exposure OC use as in the real dataset. The true values of all associated parameters were chosen as the estimates obtained from the real data when analyzed by the EZIM model.

(ii) Mixture of two normal distributions: To deviate from the exact pattern of real data and to put the nonparametric and parametric methods to test, I considered the case when $[E|D = 0, Z = i]$ comes from the following mixture: $0.5 \times N(2, 1) + 0.5 \times N(5, 1)$.

4. Finally, I generated a binary variable G standing for BRCA1/2 mutation status using the probability structure $P(G|D, E, Z)$ as given in (4–4) and (4–5). I select the true values for γ_{i1} in such a way that $\Pr(G = 1|D = 0) \approx 3.3\%$ and $\Pr(G = 1|D = 0) \approx 46.9\%$ to represent the two situations with a moderately rare and a common genetic mutation respectively. I also provide one set of simulations when G was generated from the parametric logistic regression model as in (4–8) (Table 4–4).

Apart from the above set-up which assumes G-E independence, to test the robustness of our model, under violations of this assumption, I simulate G using the model

$$\log\left\{\frac{\Pr(G=1|Z=i,E,D=0)}{\Pr(G=0|Z=i,E,D=0)}\right\}=\gamma_{i1}+\gamma_E E. \tag{4–9}$$

To vary the degree of dependence I consider two choices $\gamma_E = 0.1$ and $\gamma_E = 0.25$, that is, the odds of having $G = 1$ with one unit increase in E increases by a factor of 1.105 and 1.284 respectively. Results for only $\gamma_E = 0.25$ are included in the text (Tables 4–3, 4–4 and 4–5). The strategies I followed for choosing priors for the Bayesian methods in the simulation study are essentially same as discussed in the real data analysis. I replicated the simulation 100 times and calculated MSE based on these 100 estimates. The results are given in Tables 4–3, 4–4 and 4–5.

The simulation results are fairly clear. If interest lies in estimating the main effect of the genetic factor β_1, the Bayesian EDPM model performs the best for any choice of distributions of G and E. The fully parametric Bayesian EZIM model suffers when E is originated from any other model, for example the mixture of two normal distributions (Tables 4–4 and 4–5). When the parametric logistic assumption for $P(G|\boldsymbol{S})$ does not hold, there is a clear dominance of the Bayesian methods over the CC method for estimating β_1. Even when the data is generated from an exactly logistic model for $P(G|\boldsymbol{S})$ (Table 4–4), the Bayesian methods perform quite comparably with the CC method. The efficiency gain (for estimating β_1) in Bayesian methods is larger when the genetic mutation rarely occurs in the control population (Tables 4–3 and 4–5), which could be due to the flexibility of the likelihood in modeling the control distributions separately in the Bayesian methods, whereas CC model the marginal distribution of $G|\boldsymbol{S}$. If interest lies in estimating the main effect of E, both the CC method and the EDPM method are comparable, with CC method having a slight edge in some cases. One may note that the MSE corresponding to

β_2 for the EDPM model is often larger than the other methods as with the DPM structure I'm adding another level of model uncertainity. Indeed, the advantage of the DPM is not in terms of gain in efficiency for estimating β_2 across all scenarios, but because of its robustness. One may note that instead of modeling $P(E|\boldsymbol{S})$, CC model $P(E, \boldsymbol{S})$ nonparametrically. Their profile likelihood technique works extremely well across many different data generating mechanisms for E. On a minor note, in a small proportion of times, there does appear to be a problem with the convergence of their estimation algorithm which appears to be related to the choice of starting values. I excluded those runs when presenting the final tables for our simulation. For estimating the G-E interaction β_3, one could choose either case-only, EDPM or the CC method. When simultaneous estimation of all three parameters is considered, and Assumption 1 is fairly reasonable, the EDPM model appears to be a superior choice. Under violation of the independence assumption, performance of all the methods worsen (Tables 4–3, 4–4 and 4–5), but the loss of efficiency appears to be the least for EDPM model. The ordinary logistic regression model which is least efficient under G-E independence, especially for the interaction parameter, does not lose much efficiency under violation of G-E independence as it does not impose any restrictions on the G-E distribution.

4.5 Discussion

Epidemiologists have long grappled with this issue on how to measure interaction in a biologically meaningful way and there is still no consensus in the literature (Botto and Khury, 2001). One must recall that the statistical interaction parameter β_3 as in this chapter has a very specialized meaning which is related to the general notion of "interaction" in the scientific community only in a vague way (Cox, 1984). "No statistical interaction" in our model means constant multiplicative effect of genotype on the disease odds across all levels of the environmental exposure. A biologist might define "interaction" in a broader mechanistic sense that interaction

exists if the genetic factor and environmental exposure work on the same pathway (Brennan, 2002, Clayton and McKeigue, 2001). Assessing the joint effects of genetic and environmental factors within strata defined by other variables may provide useful insight into disease etiology and help to determine effective public health intervention strategies. The article by CC is thus a major breakthrough which emphasizes that case-control studies of gene-environment "interaction" go well beyond estimating the statistical interaction parameter β_3 and any design or analysis strategy should allow one to estimate other different parameters of interest, and should not only be targeted towards estimation of β_3. However, as emphasized throughout this chapter, I recommend extremely cautious use of the independence assumption. Scientific and empirical validation of this assumption is of utmost importance while using the proposed methods.

To conclude, I would like to highlight some of the new features of the chapter. In this chapter I proposed a fully flexible, robust Bayesian semiparametric model for estimating not only the interaction parameter, but the main effects under gene-environment independence in a stratified control population. The method outperforms the existing methods in many instances and performs comparably in others. With genetic mutation which has unequal frequencies in case and control population, the ability to model them separately through the proposed likelihood has a natural justification. When the G-E independence assumption does not hold, the method performs better when compared to other contenders. This chapter not only addresses an important problem in modern epidemiology, it also introduces some interesting statistical techniques especially for handling the high-dimensional stratum effects on the genetic and environmental exposure distribution in a data adaptive way. The use of the DPM model as illustrated in Result 2 in conjunction with transition from control to case distribution is a nice application of the theory on DP. Using prior

biological information on the frequencies of the genetic mutation reiterates the fundamental advantage of following a Bayesian paradigm. The simulation study is an additional asset of this chapter, comparing the Bayesian methods with the commonly used frequentist methods and the recently proposed method by CC.

How to handle misclassification of G and measurement error in E will be discussed in the next chapter. The ascertain bias due to different control selection mechanisms in the above framework remains topics for future research.

Table 4–1: Analysis of Israeli ovarian cancer data by all five methods, considering OC use as the only environmental exposure, with 95% HPD and confidence intervals

Model		β_1	β_2	β_3	α	K
EZIM	Estimate	3.7832	-0.0527	0.0910		
	post. stdev	0.1317	0.0243	0.0326		
	HPD	(3.4641, 3.9764)	(-0.1265, -0.0140)	(0.0270, 0.1482)		
EDPM	Estimate	3.7537	-0.0748	0.1091	14.7455	5.7630
	post. stdev	0.1294	0.0303	0.0352	5.8361	1.8836
	HPD	(3.4358,3.9310)	(-0.1409,-0.0151)	(0.0364,0.1791)	(5.8913,28.5666)	(2,10)
CC	Estimate	3.6323	-0.0624	0.1110		
	std. error	0.3999	0.0266	0.0341		
	CI	(2.8485, 4.4161)	(-0.1145, -0.0103)	(0.0442, 0.1778)		
Ordinary Logistic	Estimate	3.7710	-0.0642	0.0476		
	std. error	0.4407	0.0268	0.0999		
	CI	(2.9072, 4.6348)	(-0.1167, -0.0117)	(-0.1482, 0.2434)		
Case-Only	Estimate			0.0924		
	std. error			0.0329		
	CI			(0.0279, 0.1569)		

Table 4–2: Analysis of Israeli ovarian cancer data by all five methods, considering both OC use and parity as environmental exposures, with 95% HPD and confidence intervals

Model		β_G	β_{OC}	β_{Parity}	β_{OC*G}	$\beta_{Parity*G}$
EZIM	Estimate	3.7877	-0.0829	-0.0369	0.1566	-0.0781
	post. stdev	0.1573	0.0272	0.0304	0.0366	0.0427
	HPD	(3.4937, 4.0909)	(-0.1491, -0.0379)	(-0.0864, 0.0119)	(0.0765,0.2230)	(-0.1686,0.0026)
EDPM	Estimate	3.8808	-0.0631	-0.0404	0.1360	-0.1072
	post. stdev	0.1566	0.0202	0.0311	0.0331	0.0501
	HPD	(3.5748,4.1713)	(-0.1034,-0.0224)	(-0.0947,0.0260)	(0.0823,0.2123)	(-0.2207,0.0031)
CC	Estimate	3.8961	-0.0620	-0.0599	0.1128	-0.1041
	std. error	0.4297	0.0267	0.0320	0.0344	0.0599
	CI	(3.0539,4.7383)	(-0.1143,-0.0097)	(-0.1227,0.0029)	(0.0454,0.1802)	(-0.2214,0.0133)
Ordinary Logistic	Estimate	4.7321	-0.0582	-0.0388	0.0292	-0.3869
	std. error	0.7411	0.0263	0.0317	0.1080	0.1481
	CI	(3.2795,6.1847)	(-0.1097,-0.0067)	(-0.1009,0.0233)	(-0.1825,0.2409)	(-0.6772,-0.0966)
Case-Only	Estimate				0.0931	-0.0565
	std. error				0.0331	0.0591
	CI				(0.0283,0.1579)	(-0.1724,0.0594)

Table 4–3: Simulation scenarios: E is Zero-Inflated ; G :rare or common; G-E independence assumption holds ($\gamma_E = 0$) or does not hold ($\gamma_E = 0.25$). Mean denotes the mean estimate based on 100 replications, whereas MSE is the estimated mean squared error based on 100 replications.

G	γ_E	Model		β_1	β_2	β_3
			True	3.0000	-0.0700	0.1200
Rare	0	EZIM	Mean	2.9777	-0.0623	0.1132
			MSE	0.0196	0.0006	0.0015
		EDPM	Mean	2.9315	-0.0615	0.1140
			MSE	0.0242	0.0011	0.0016
		CC	Mean	2.9013	-0.0630	0.1152
			MSE	0.2633	0.0006	0.0015
		Ordinary	Mean	2.9053	-0.0632	0.1877
			MSE	0.3743	0.0006	0.0619
		Case-Only	Mean			0.1196
			MSE			0.0017
Common	0	EZIM	Mean	2.9838	-0.0811	0.1282
			MSE	0.0099	0.0016	0.0014
		EDPM	Mean	2.9722	-0.0834	0.1260
			MSE	0.0108	0.0018	0.0016
		CC	Mean	2.7742	-0.0787	0.1239
			MSE	0.0805	0.0015	0.0012
		Ordinary	Mean	2.8179	-0.0775	0.1214
			MSE	0.0663	0.0016	0.0026
		Case-Only	Mean			0.1296
			MSE			0.0014
Common	0.25	EZIM	Mean	2.8643	-0.2752	0.3368
			MSE	0.0313	0.0489	0.0536
		EDPM	Mean	2.8960	-0.1465	0.2119
			MSE	0.0218	0.0064	0.0090
		CC	Mean	2.4190	-0.3116	0.3723
			MSE	0.3580	0.0645	0.0695
		Ordinary	Mean	2.8006	-0.2586	0.1786
			MSE	0.0637	0.0426	0.0109
		Case-Only	Mean			0.3950
			MSE			0.0822

Table 4–4: Simulation scenarios: E: Mixture of two normals; G: with parametric logistic in terms of $\boldsymbol{S}$ as in (4–8) or commonly prevalent as in (4–4); G-E independence holds ($\gamma_E = 0$) or does not hold ($\gamma_E = 0.25$). Mean denotes the mean estimate based on 100 replications, whereas MSE is the estimated mean squared error based on 100 replications.

G model	γ_E	Model		β_1	β_2	β_3
			True	3.0000	-0.0700	0.1200
Generated by (4–8)	0	EZIM	Mean	3.0000	-0.0447	0.1323
			MSE	0.0685	0.0798	0.0092
		EDPM	Mean	2.9880	-0.0710	0.1290
			MSE	0.0588	0.0022	0.0022
		CC	Mean	3.0185	-0.0782	0.1284
			MSE	0.0563	0.0015	0.0019
		Ordinary	Mean	3.0157	-0.0793	0.1323
			MSE	0.1282	0.0017	0.0080
		Case-Only	Mean			0.1289
			MSE			0.0018
Generated by (4–4)	0	EZIM	Mean	2.8213	-0.1227	0.1352
			MSE	0.0363	0.0886	0.0076
		EDPM	Mean	2.9870	-0.0695	0.1189
			MSE	0.0303	0.0023	0.0021
		CC	Mean	2.7542	-0.0713	0.1249
			MSE	0.1098	0.0017	0.0023
		Ordinary	Mean	2.7683	-0.0722	0.1328
			MSE	0.1856	0.0019	0.0087
		Case-Only	Mean			0.1251
			MSE			0.0023
Generated by (4–4)	0.25	EZIM	Mean	2.8581	-0.3114	0.3778
			MSE	0.1568	0.2127	0.1886
		EDPM	Mean	2.8858	-0.2692	0.3473
			MSE	0.0449	0.0442	0.0549
		CC	Mean	1.9287	-0.2984	0.3682
			MSE	0.8899	0.0563	0.0652
		Ordinary	Mean	2.7678	-0.2509	0.1519
			MSE	0.1368	0.0374	0.0076
		Case-Only	Mean			0.3212
			MSE			0.0885

Table 4–5: Simulation scenarios: E: Mixture of two normals; G: rarely prevalent; G-E independence holds ($\gamma_E = 0$) or does not hold ($\gamma_E = 0.25$).Mean denotes the mean estimate based on 100 replications, whereas MSE is the estimated mean squared error based on 100 replications.

G	γ_E	$E\|D=0,\mathbf{Z}$		β_1	β_2	β_3
			True value	3.0000	-0.0700	0.1200
Rare	0	EZIM	Mean	2.9169	-0.0437	0.1587
			MSE	0.0822	0.0311	0.0209
		EDPM	Mean	2.9294	-0.0674	0.1296
			MSE	0.0670	0.0016	0.0039
		CC	Mean	2.8525	-0.0725	0.1335
			MSE	0.2223	0.0009	0.0047
		Ordinary	Mean	2.8732	-0.0715	0.1450
			MSE	0.8280	0.0010	0.0605
		Case-Only	Mean			0.1340
			MSE			0.0045
Rare	0.25	EZIM	Mean	3.0717	-0.2041	0.3076
			MSE	0.1688	0.1934	0.0481
		EDPM	Mean	3.1533	-0.1324	0.2990
			MSE	0.0965	0.0056	0.0341
		CC	Mean	2.0735	-0.1444	0.3143
			MSE	0.8476	0.0066	0.0398
		Ordinary	Mean	3.1541	-0.1393	0.0452
			MSE	0.7112	0.0059	0.0432
		Case-Only	Mean			0.3505
			MSE			0.0559

Figure 4–1: Real data analyzed with EDPM model by considering OC use as an environmental exposure: Histogram of last 5000 MCMC values for the main effects and interaction parameter with overlayed smoothed kernel density.

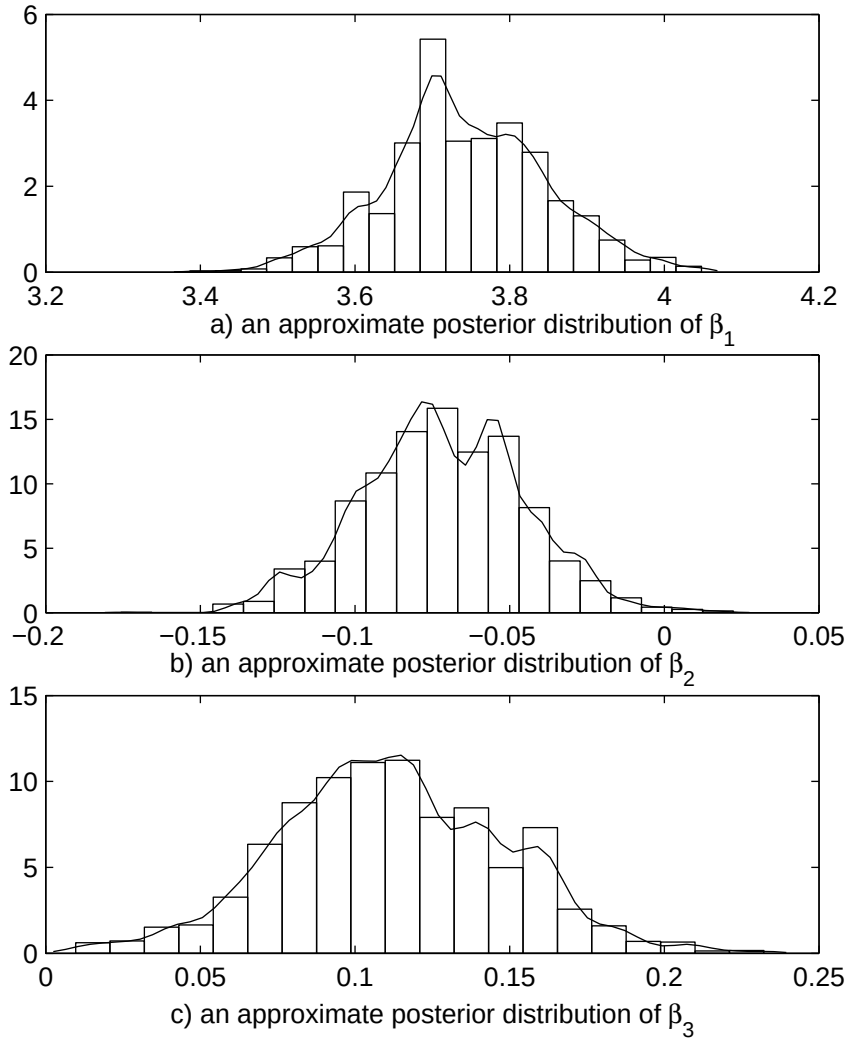
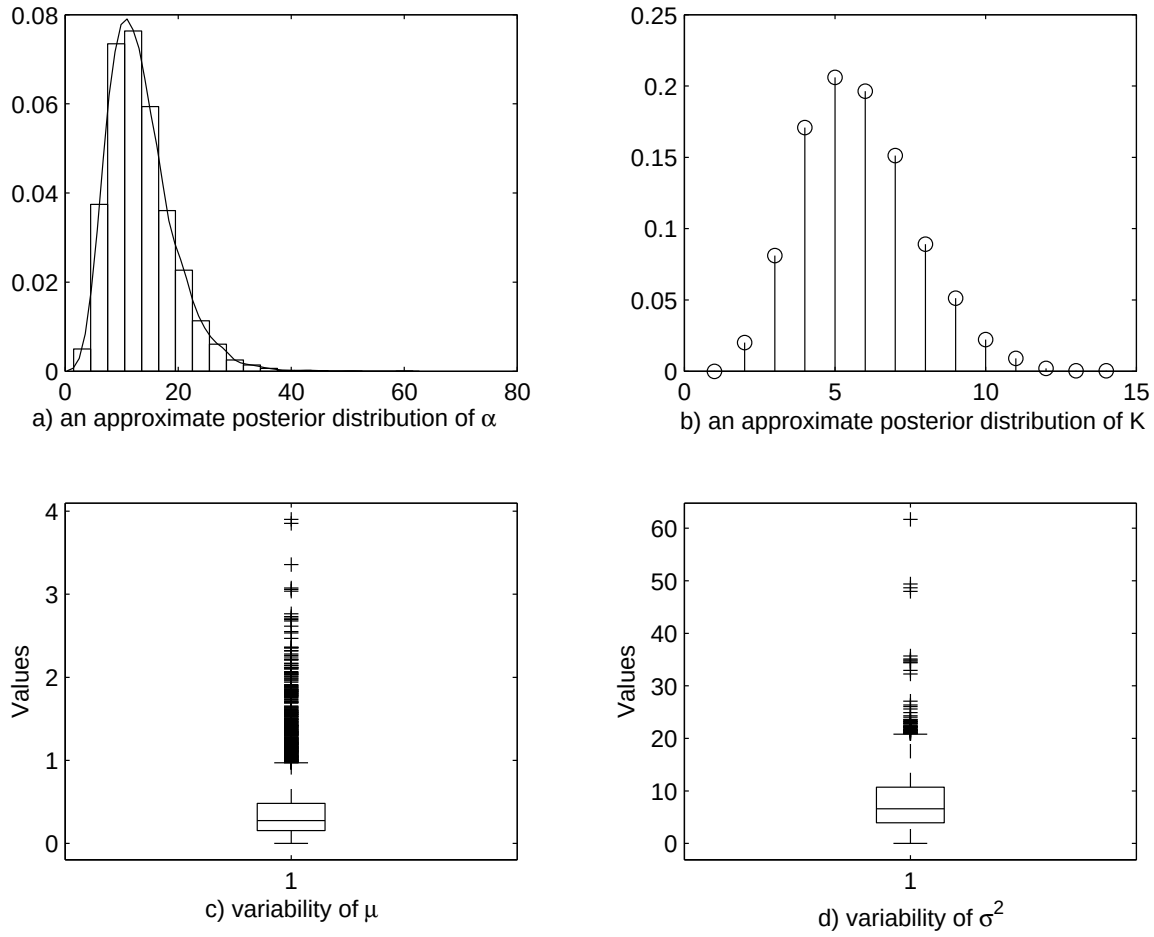


Figure 4–2: Details of DPM model by considering OC use as an environmental exposure: Histogram corresponding to approximate posterior distribution of α and K in the DPM model. Also plotted are histograms of variances of the μ_i's and σ_i $i = 1, \cdots, 24$, calculated for each of the last 5000 MCMC runs.



CHAPTER 5
ACCOUNTING FOR ERROR DUE TO MISCLASSIFICATION OF EXPOSURES IN CASE-CONTROL STUDIES OF GENE-ENVIRONMENT INTERACTION

5.1 Introduction

Measurement error in exposure assessment is one of the major sources of bias in epidemiological studies. When ignored, these errors bias point and interval estimates of effect, and invalidate p-values of hypotheses tests. Often, although not always, the bias is towards the null value, underestimating the true exposure-disease relationship, and there can be a substantial loss of power in hypothesis tests. The pervasiveness and extensiveness of these exposure measurement and misclassification errors in epidemiologic research may explain much of the inconsistent and inconclusive results currently reported in the literature.

Bashir and Duffy (1997) provided a general review of epidemiological methods for dealing specifically with measurement error and misclassification. Greenland and Kleinbaum (1983) proposed a simple two-stage procedure for estimating the odds ratio in matched pairs, with a corrected variance estimator developed later by Greenland (1989). Rice and Holmans (2003) obtained analytic formulae for estimates of genotypic relative risks in terms of the genotyping error probability in analysis of unmatched case-control studies with a single binary genetic factor as exposure. Later, Rice (2003) proposed a fulllikelihood-based approach to obtain estimates and confidence intervals for the parameters of interest in the presence of misclassification of a binary exposure in matched case-control studies. However, much of the discussion on the effects of misclassification of exposure have just focused on the impact on the relative risk and/or sample size in studies of a single factor. In contrast, less attention

has been given to the influence of misclassification on the assessment of interactions between two or more factors.

As already indicated in Chapter 4, one of the major goals in many recent epidemiological studies has been to investigate the effect of genes on a disease, in combination with environmental exposures. In case-control studies of gene-environment association with disease, when genetic and environmental exposures can be assumed to be independent in the underlying population, one may exploit the independence in order to derive more efficient estimation techniques than the traditional logistic regression analysis (Chatterjee and Carroll, 2005). Garcia-Closas *et al.* (1998) showed that under a set of conditions often satisfied in studies of gene-environment interactions, both differential and nondifferential misclassification of binary environmental factors bias a multiplicative interaction effect toward the null value. Garcia-Closas *et al.* (1999) proposed a simple approach to assess the impact of misclassification on bias in the estimation of multiplicative or additive interactions and on sample size requirements. They pointed out that under misclassification of exposures, increased sample size is needed to attain the same power to detect the attenuated interaction. The focus of Garcia-Closas *et al.* (1999) was primarily on the study design issues under misclassification, and did not propose corrected estimates of the parameters of interest, or inferential adjustments, if in fact, misclassification is present in the data.

The current chapter describes a relatively simple approach to adjust the estimation of the parameters of interest in case-control studies of gene-environment interaction in the presence of misclassification. The proposed method exploits G-E independence assumption, and obtains corrected parameter estimates of all the parameters of interest, and not just the interaction odds ratio. I consider a unmatched case-control set-up, adapt and extend the work of Rice and Holmans (2003) to the

situation when one has a binary genetic risk factor (G), a binary environmental exposure (E), and both are potentially subject to misclassification, where I use the traditional formulation in terms of odds ratios in a 2×4 table.

In Section 5.2, I focus on an unmatched case-control sampling design. I formulate the maximum likelihood estimation (MLE) problem under the G-E independence assumption and obtain parameter estimates under this additional restriction. I first make a rare disease assumption, in which case, I can obtain closed-form expression for parameter estimates. I point out that the estimate of the gene-environment interaction parameter obtained by this approach, as expected, is exactly identical to the estimate obtained by the popular case-only approach (Piergorsch *et al.*, 1994). However, with knowledge of the marginal prevalence of the disease in the population ($P(D = 1)$), I can relax the rare disease assumption. In the latter situation, I can also obtain constrained MLEs. Though the corresponding score equations do not have explicit closed form solutions, numerical evaluation is extremely straightforward.

After this preliminary formulation with a perfectly measured dataset, I delve into the issue of adjustments for the presence of misclassification. In the presence of misclassification, I adjust the MLEs based on the sensitivity and specificity of the measuring instruments for genetic and environmental factors. Test statistics and confidence intervals are formulated as in any standard likelihood-based inference using the asymptotic distribution of the MLE, once the adjustments are made. In fact, as misclassification error rates go to zero, the estimates reduce to the standard MLEs for data recorded without misclassification. I also provide sample size calculation and power comparisons for the proposed methods. I assume throughout that I have been given fixed values of sensitivity and specificity of the measurement devices and make no attempt to estimate them in this chapter. Rice and Holmans (2003) proposed to use an external validation study to estimate the sensitivity and specificity parameters which could also be adapted in this case. It may be noted that in contrast to the

central theme of this dissertation, I adopt purely classical maximum likelihood based techniques in this particular chapter instead of a Bayesian route, as the maximum likelihood computations are very efficient and compact in this particular unmatched framework with two binary exposures.

Many genetic and environmental factors in epidemiological studies are liable to mismeasurement, leading to incorrect findings. Efforts to improve the accuracy in measuring both genetic and environmental factors are critical for the valid assessment of G-E interactions as well as their main effects in case-control studies. Simulation studies show that my corrected inference gain substantially in terms of efficiency, and reduce bias when compared to the unadjusted ones. Also, my proposed adjusted approaches increase the power for testing the interaction effect, and thus decrease the sample size required to attain a designated power level.

The rest of the chapter is organized as follows. In Section 5.2 I discuss unmatched case-control studies, where I first present the models, likelihoods and assumptions in the absence of misclassification and discuss maximum likelihood estimation. I then introduce adjustments for the true parameters in the presence of misclassification. I point out the equivalence of case-only method and my method in terms of estimating the interaction parameter, and the subsequent adjustment due to the misclassified exposure. Section 5.3 contains supporting numerical evidence including simulation studies with sample size determination and power calculation for unmatched case-control studies. Section 5.4 contains concluding discussion. Some proofs and detailed calculations are relegated to Appendix C.

5.2 Unmatched Case-Control Studies of Gene Environment Interaction

I consider unmatched case-control studies with a binary genetic factor G and a binary environmental exposure E which take values 1 for susceptible (exposed in the case of E) and 0 for nonsusceptible (unexposed in the case of E) subjects. Let D denote the disease status, where $D = 1$ denotes affected, and $D = 0$ denotes

Table 5–1: Data for a unmatched case-control study with a binary genetic factor and a binary environmental exposure.

	$G=0$		$G=1$		
	$E=0$	$E=1$	$E=0$	$E=1$	total
j	1	2	3	4	
$D=0$	r_{01}	r_{02}	r_{03}	r_{04}	n_0
$D=1$	r_{11}	r_{12}	r_{13}	r_{14}	n_1

unaffected individuals. Using the same notation as in Garcia-Closas *et al.* (1999), the odds ratio OR_{eg} measures the association between disease and the environmental and genetic factors. Relative to subjects not exposed to the environmental or genetic factor, I define the following odds ratios: OR_{10} denotes the odds ratio for nonsusceptible subjects exposed to the environmental factor; OR_{01} denotes the odds ratio for susceptible subjects not exposed to the environmental factor; and OR_{11} denotes the odds ratio for susceptible subjects exposed to the environmental factor. Therefore, $\psi = OR_{11}/(OR_{10}OR_{01})$ is the multiplicative interaction parameter.

5.2.1 Maximum Likelihood Estimation under G-E Independence Assumption

Table 5–1 shows a general form of data under an unmatched case-control design. In the absence of misclassification, I can assume that data from control population and case population, each forms a multinomial distribution, namely,

$$\begin{aligned} \boldsymbol{r}_0 &\sim Mn(n_0, \boldsymbol{p}_0) \\ \boldsymbol{r}_1 &\sim Mn(n_1, \boldsymbol{p}_1), \end{aligned} \tag{5–1}$$

where n_0 and n_1 are fixed, and

$$\begin{aligned} \boldsymbol{r}_0 &= (r_{01}, r_{02}, r_{03}, r_{04}) \\ \boldsymbol{r}_1 &= (r_{11}, r_{12}, r_{13}, r_{14}) \\ \boldsymbol{p}_0 &= (p_{01}, p_{02}, p_{03}, p_{04} = 1 - p_{01} - p_{02} - p_{03}) \\ \boldsymbol{p}_1 &= (p_{11}, p_{12}, p_{13}, p_{14} = 1 - p_{11} - p_{12} - p_{13}). \end{aligned}$$

By the definition of the odds ratio,

$$\begin{aligned} OR_{10} &= \frac{p_{01}p_{12}}{p_{02}p_{11}} \\ OR_{01} &= \frac{p_{01}p_{13}}{p_{03}p_{11}} \\ OR_{11} &= \frac{p_{01}p_{14}}{p_{04}p_{11}} \\ \psi &= \frac{p_{02}p_{03}p_{11}p_{14}}{p_{01}p_{04}p_{12}p_{13}}. \end{aligned}$$

Thus, I obtain the case-distribution as,

$$\begin{aligned} p_{11} &= \frac{p_{01}}{p} \\ p_{12} &= \frac{p_{02}}{p} \cdot OR_{10} \\ p_{13} &= \frac{p_{03}}{p} \cdot OR_{01} \\ p_{14} &= \frac{p_{04}}{p} \cdot OR_{10} \cdot OR_{01} \cdot \psi, \end{aligned} \tag{5–2}$$

where $p = p_{01} + p_{02} \cdot OR_{10} + p_{03} \cdot OR_{01} + p_{04} \cdot OR_{10} \cdot OR_{01} \cdot \psi$. The corresponding multinomial likelihood, when parameterized in terms of the control probabilities and the odds ratios is

$$L_1 = L(OR_{10}, OR_{01}, \psi, p_{01}, p_{02}, p_{03} | \boldsymbol{r}_0, \boldsymbol{r}_1) = \prod_{d=0}^{1} \prod_{j=1}^{4} p_{dj}^{r_{dj}}. \tag{5–3}$$

Note that the parametrization of p_{01}, p_{02} and p_{03} imposes no other restrictions except for they lead to valid probability distributions (they are all positive and sum to less than 1). Similarly, the odds ratios are just expected to be positive. I can always fit the multinomial model to this data and obtain the maximum likelihood estimates (MLEs) of the parameters of interest and their estimated asymptotic variance ($\widehat{AVAR}$) as in Table 5–2 under the column of *traditional model*. The MLEs of the cell probabilities are simply given by $\hat{p}_{dj} = r_{dj}/n_d$, $d = 0, 1$, $j = 1, \cdots, 4$.

Now let me describe how the estimation changes with the additional assumption of *G-E* independence holding in the source population. I first investigate the estimates

Table 5–2: In the absence of misclassification, the MLEs of the odds ratios and their estimated asymptotic variances in terms of observed counts r_{dj} for both traditional model and the model under G-E independence and rare disease.

Parameters		Traditional model	G-E independence and rare disease
OR_{10}	MLE	$\widehat{OR}_{10} = \frac{r_{01}r_{12}}{r_{02}r_{11}}$	$\widehat{OR}_{10}^{IR} = \frac{r_{12}(r_{01}+r_{03})}{r_{11}(r_{02}+r_{04})}$
	$\widehat{AVAR}$	$\widehat{OR}_{10}^2(\frac{1}{r_{01}} + \frac{1}{r_{02}} + \frac{1}{r_{11}} + \frac{1}{r_{12}})$	$\left(\widehat{OR}_{10}^{IR}\right)^2(\frac{1}{r_{01}+r_{03}} + \frac{1}{r_{02}+r_{04}} + \frac{1}{r_{11}} + \frac{1}{r_{12}})$
OR_{01}	MLE	$\widehat{OR}_{01} = \frac{r_{01}r_{13}}{r_{03}r_{11}}$	$\widehat{OR}_{01}^{IR} = \frac{r_{13}(r_{01}+r_{02})}{r_{11}(r_{03}+r_{04})}$
	$\widehat{AVAR}$	$\widehat{OR}_{01}^2(\frac{1}{r_{01}} + \frac{1}{r_{03}} + \frac{1}{r_{11}} + \frac{1}{r_{13}})$	$\left(\widehat{OR}_{01}^{IR}\right)^2(\frac{1}{r_{01}+r_{02}} + \frac{1}{r_{03}+r_{04}} + \frac{1}{r_{11}} + \frac{1}{r_{13}})$
ψ	MLE	$\hat{\psi} = \frac{r_{02}r_{03}r_{11}r_{14}}{r_{01}r_{04}r_{12}r_{13}}$	$\hat{\psi}^{IR} = \frac{r_{11}r_{14}}{r_{12}r_{13}}$
	$\widehat{AVAR}$	$\hat{\psi}^2(\sum_{d=0}^{1}\sum_{j=1}^{4}\frac{1}{r_{dj}})$	$\left(\hat{\psi}^{IR}\right)^2(\sum_{j=1}^{4}\frac{1}{r_{1j}})$

under a rare disease assumption, an assumption which is routinely made in epidemiological studies. The assumption of G-E independence in the source population, $P(G,E) = P(G)P(E)$ in conjunction with the rare disease assumption, implies that G-E independence holds in the control population, i.e., $P(G,E|D=0) = P(G|D=0)P(E=0)$. This adds an additional restriction on p_{01}, p_{02} and p_{03}:

$$p_{01}(1 - p_{01} - p_{02} - p_{03}) = p_{02}p_{03}. \tag{5–4}$$

With this additional restriction, maximizing the likelihood (5–3) will not provide the same estimates as in the previous model. The MLEs and their $\widehat{AVAR}$ in this restricted parameter space are presented in Table 5–2 under the column of G-E *independence and rare disease*. The constrained ML equations and their solutions which lead to this column in Table 5–2 are presented in Appendix C.1.

If the disease prevalence $P(D=1) = \pi$ in the source population is known, I can relax the rare disease assumption by expressing the G-E independence as the following:

$$\begin{aligned} & P(G=g)P(E=e) = P(G=g, E=e) \\ = \; & P(G=g, E=e|D=0)P(D=0) + P(G=g, E=e|D=1)P(D=1) \end{aligned} \tag{5–5}$$

where $g, e = 0, 1$. Therefore, instead of the restriction as in (5–4), we have the following restriction on p_{01}, p_{02} and p_{03}:

$$\begin{aligned} f &= (1-\pi)p_{04} + \pi OR_{10}OR_{01}\psi p_{04}/p \\ &- [(1-\pi)(p_{02}+p_{04}) + \pi(OR_{10}p_{02} + OR_{10}OR_{01}\psi p_{04})/p] \\ &\times [(1-\pi)(p_{03}+p_{04}) + \pi(OR_{01}p_{03} + OR_{10}OR_{01}\psi p_{04})/p] = 0. \end{aligned} \quad (5\text{–}6)$$

The details of obtaining (5–6) is delegated to Appendix C.2. With this additional restriction, maximizing the likelihood (5–3) will not provide the same estimates as under the rare disease assumption. In fact the solutions to the ML equations can not be written in closed forms. However, we can obtain these restricted MLEs by the usual Newton-Raphson algorithm, and obtain the estimated asymptotic variance-covariance matrix by the inverse of the observed information matrix. The observed information matrix is constructed by taking the second derivative of the log-likelihood with respect the parameters, and evaluating them at the MLEs of the parameters.

Once we obtain the MLEs of the parameters of interest, in this case the odds ratios, one would like to conduct tests of association or interaction as well as construct large sample confidence intervals for the parameters of interest. Because of the skewness in the sampling distribution of the estimated odds ratios, statistical inference for the parameters of interest (denoted by a generic symbol θ) uses an alternative but equivalent measure: its natural logarithm, $\log(\hat{\theta})$, which has a less skewed sampling distribution that is closer to normal. By a simple use of delta method, the large-sample distribution of $\log(\hat{\theta})$ is approximately normal, i.e.

$$\log(\hat{\theta}) \sim N(\log(\theta), AVAR(\log\hat{\theta})),$$

where $\hat{\theta}$ is the MLE of θ, and the estimated variance of $\log(\hat{\theta})$ is given by, $\widehat{AVAR}(\log\hat{\theta}) = \hat{\theta}^{-2}\widehat{Var}(\hat{\theta})$. Z-tests and confidence intervals for the log-scale parameters are constructed based on the above large sample distribution.

REMARK 1: It is well-known that in a multinomial set-up, the expected cell counts, namely, $E_{\boldsymbol{p}_d}[\boldsymbol{r}_d]=n_d\,\boldsymbol{p}_d$, where $E_{\boldsymbol{p}_d}[\boldsymbol{r}_d]$ represents the row vector of expected cell counts corresponding to $D=d, d=0,1$, and $\boldsymbol{p}_d$ denotes the probability vector. Then the vector of *estimated* expected cell frequencies, denoted by $\tilde{\boldsymbol{r}}_d$ is given by $E_{\boldsymbol{p}_d}[\boldsymbol{r}_d]|_{\boldsymbol{p}_d=\hat{\boldsymbol{p}}_d} = n_d\hat{\boldsymbol{p}}_d$, (i.e., the expected frequencies evaluated at the MLE's of the model parameters). For example, for the usual multinomial model, without any restrictions on the exposure space, the vector of estimated expected cell frequencies matches exactly with the observed frequencies, that is, $\tilde{\boldsymbol{r}}_d = \boldsymbol{r}_d$, (as $\hat{\boldsymbol{p}}_d = \boldsymbol{r}_d/n_d$), where $\boldsymbol{r}_d$ is the vector of observed frequencies.

Under the independence and rare disease assumptions, (which I denote by the superscript IR below to distinguish from the other models), from Appendix C.1, I note that the MLEs for $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ are,

$$\begin{aligned}
\hat{p}_{01}^{IR} &= \frac{(r_{01}+r_{03})(r_{01}+r_{02})}{n_0^2} \\
\hat{p}_{02}^{IR} &= \frac{(r_{01}+r_{02})(r_{02}+r_{04})}{n_0^2} \\
\hat{p}_{03}^{IR} &= \frac{(r_{01}+r_{03})(r_{03}+r_{04})}{n_0^2} \\
\hat{p}_{04}^{IR} &= \frac{(r_{02}+r_{04})(r_{03}+r_{04})}{n_0^2} \\
\hat{p}_{1j}^{IR} &= \frac{r_{1j}}{n_1}, \quad j=1,2,3,4, \qquad (5\text{–}7)
\end{aligned}$$

and thus the estimated expected frequencies are obtained simply by $\tilde{\boldsymbol{r}}_d^{IR} = n_d\,\hat{\boldsymbol{p}}_d^{IR}$.

REMARK 2: One can obtain the estimates of OR_{10}, OR_{01} and the interaction effect ψ as well as the cell probabilities in the control population $\boldsymbol{p}_0$. Define the genetic and environmental marginal odds ratios OR_G and OR_E as the following,

$$\begin{aligned}
OR_E &= \frac{P(D=1|E=1)P(D=0|E=0)}{P(D=0|E=1)P(D=1|E=0)} \\
OR_G &= \frac{P(D=1|G=1)P(D=0|G=0)}{P(D=0|G=1)P(D=1|G=0)}.
\end{aligned}$$

Thus one also can estimate OR_G and OR_E by using the following identities,

$$\begin{aligned} OR_E &= \left\{\frac{p_{01}+p_{03}}{p_{02}+p_{04}}\right\}\left\{\frac{p_{02}OR_{10}+p_{04}OR_{10}OR_{01}\psi}{p_{01}+p_{03}OR_{01}}\right\} \\ OR_G &= \left\{\frac{p_{01}+p_{02}}{p_{03}+p_{04}}\right\}\left\{\frac{p_{03}OR_{01}+p_{04}OR_{10}OR_{01}\psi}{p_{01}+p_{02}OR_{10}}\right\}. \end{aligned}$$

Under the G-E independence and rare disease assumptions, we have $p_{01}p_{04} = p_{02}p_{03}$, furthermore, $P(E|D=0) \approx P(E)$ and $P(G|D=0) \approx P(G)$, thus one can estimate OR_G and OR_E by using

$$\begin{aligned} OR_E &= \frac{(1-P(G=1))OR_{10}+P(G=1)OR_{10}OR_{01}\psi}{(1-P(G=1))+P(G=1)OR_{01}} \\ OR_G &= \frac{(1-P(E=1))OR_{01}+P(E=1)OR_{10}OR_{01}\psi}{(1-P(E=1))+P(E=1)OR_{10}}. \end{aligned}$$

Note that $P(E=1) \approx p_{02}+p_{04}$ and $P(G=1) \approx p_{03}+p_{04}$.

5.2.2 Maximum Likelihood Estimation in The Presence of Misclassification

Now, I introduce the effects of misclassification in the present framework. Our model for misclassified data is based on the assumption that some perfectly classified "true" case-control data exist, where the "true" parameters of interest and the true underlying cell probabilities follow the same distribution as $\boldsymbol{p}_d$ as discussed above. Following the "star" notation of Rice (2003), I let the superscript asterisk denote the true parameters for the true data model as well as the true variables. Let sp_{dG} (se_{dG}) and sp_{dE} (se_{dE}) denote specificity (sensitivity) of G and E with disease status d, respectively, where sensitivity = P(observed exposed | truly exposed) and specificity = P(observed unexposed | truly unexposed), so $se_{dG} = P(G=1|G^*=1, D=d)$, $se_{dE} = P(E=1|E^*=1, D=d)$, $sp_{dG} = P(G=0|G^*=0, D=d)$ and $sp_{dE} = P(G=0|G^*=0, D=d)$. Applying a classical error structure, all subjects are assumed to have the same probability of the observed exposure, conditional on their case/control status and true exposure. I then have the following two results.

RESULT 1 Assuming that given the disease status $d(=0,1)$ and the true exposure statuses of G and E, the observed exposure statuses of G and E are independent, then

$$\begin{pmatrix} p_{d1} & p_{d2} \\ p_{d3} & p_{d4} \end{pmatrix} = \boldsymbol{A} \begin{pmatrix} p^*_{d1} & p^*_{d2} \\ p^*_{d3} & p^*_{d4} \end{pmatrix} \boldsymbol{B}, \tag{5–8}$$

where

$$\boldsymbol{A} = \begin{pmatrix} sp_{dG} & 1-se_{dG} \\ 1-sp_{dG} & se_{dG} \end{pmatrix} \quad \text{and} \quad \boldsymbol{B} = \begin{pmatrix} sp_{dE} & 1-sp_{dE} \\ 1-se_{dE} & se_{dE} \end{pmatrix}.$$

PROOF:

$$\begin{aligned} & P(G,E|D=d) \\ = & \sum_{g=0}^{1}\sum_{e=0}^{1} P(G,E|D=d,G^*=g,E^*=e)P(G^*=g,E^*=e|D=d) \\ = & \sum_{g=0}^{1}\sum_{e=0}^{1} P(G|D=d,G^*=g,E^*=e)P(E|D=d,G^*=g,E^*=e)P(G^*=g,E^*=e|D=d) \\ = & \sum_{g=0}^{1}\sum_{e=0}^{1} P(G|D=d,G^*=g)P(E|D,E^*=e)P(G^*=g,E^*=e|D=d). \end{aligned}$$

Note p_{dj} as defined in Table 5–1 denotes the cell probabilities of jth (G_d, E_d) configuration given disease status d, where values of j are as stated in Table 5–1. Result 1 holds for all three situations discussed in the previous section. Therefore, if the observed data comes from a common multinomial distribution with cell probabilities p_{dj}, then I can write down the likelihood (5–3) in terms of the true, "starred" parameters. I simply write the p_{dj}'s in terms of a linear function of the true parameters p^*_{dj} as defined by Result 1 and maximize the following multinomial likelihood in terms of the underlying true or starred parameters.

$$L_2 = L(OR^*_{10}, OR^*_{01}, \psi^*, p^*_{01}, p^*_{02}, p^*_{03}|\boldsymbol{r}_0, \boldsymbol{r}_1) = \prod_{d=0}^{1}\prod_{j=1}^{4}\{p_{dj}(\boldsymbol{p}^*_d)\}^{r_{dj}}, \tag{5–9}$$

where $p_{dj}(\boldsymbol{p}_d^*)$ denotes the linear transformation defined in (5–8), essentially I'm replacing the p_{dj} in the original likelihood by a function of the underlying true parameters as described in Result 1. Thus, by maximizing the likelihood (5–9) which now includes the effect of misclassification through the linear transformation on the parameters with the correction matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, I can now obtain the MLEs of the starred parameters, denoted by $\hat{\boldsymbol{p}}_d^*$.

As indicated in Remark 1, the vector of estimated expected cell counts under the multinomial model is given by $\tilde{\boldsymbol{r}}_d = n_d\, \hat{\boldsymbol{p}}_d$. Thus for the estimation with the starred parameters, the vector of estimated expected cell counts under the true data model is $\boldsymbol{r}_d^* = n_d\, \hat{\boldsymbol{p}}_d^*$. Note that by invariance property of the MLE, Result 1 holds when the parameters $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ are replaced with the MLEs for the perfectly classified data model and the misclassified data model. Thus by inverting Result 1 as in (5–8), replacing the parameters with the MLE's, I have

$$\begin{pmatrix} \hat{p}_{d1}^* & \hat{p}_{d2}^* \\ \hat{p}_{d3}^* & \hat{p}_{d4}^* \end{pmatrix} = \boldsymbol{A}^{-1} \begin{pmatrix} \hat{p}_{d1} & \hat{p}_{d2} \\ \hat{p}_{d3} & \hat{p}_{d4} \end{pmatrix} \boldsymbol{B}^{-1} = \frac{1}{n_d} \boldsymbol{A}^{-1} \begin{pmatrix} \tilde{r}_{d1} & \tilde{r}_{d2} \\ \tilde{r}_{d3} & \tilde{r}_{d4} \end{pmatrix} \boldsymbol{B}^{-1}.$$

This immediately leads to the following relationship between estimated expected cell counts for the true data and the misclassified data,

RESULT 2

$$\begin{pmatrix} r_{d1}^* & r_{d2}^* \\ r_{d3}^* & r_{d4}^* \end{pmatrix} = \boldsymbol{A}^{-1} \begin{pmatrix} \tilde{r}_{d1} & \tilde{r}_{d2} \\ \tilde{r}_{d3} & \tilde{r}_{d4} \end{pmatrix} \boldsymbol{B}^{-1}. \tag{5–10}$$

In fact, the result is true for the vector of expected cell counts involving the unknown parameters, not only the estimates, as is obvious from the above discussion.

Thus, for the traditional multinomial model and the model under the *G-E* independence and rare disease assumptions, the MLEs of the true starred parameters of interest have closed-form expression in terms of the estimated starred expected cell counts $\boldsymbol{r}_d^*$, which are shown in Table 5–3. To obtain $\boldsymbol{r}_d^*$, I simply obtain the MLEs

$\hat{\boldsymbol{p}}_d^*$ under different models and multiply by n_d. Note that the MLEs $\hat{\boldsymbol{p}}_d^*$ are also easily obtained by using the transformation in Result 1 and the ML estimation of $\boldsymbol{p}_d$ as discussed in Section 5.2.1 under different model assumptions. The MLEs $\hat{\boldsymbol{p}}^*$ turn out to be different functions of the observed cell counts $\boldsymbol{r}_d$, sensitivity and specificity parameters, the form of the function depending on the model assumptions. Therefore, $\boldsymbol{r}_d^*$ under different assumptions or constraints on the parameters might be different (I denote by the superscript IR under G-E independence and rare disease assumptions to distinguish from the other models) as the MLEs $\hat{\boldsymbol{p}}_d^*$ (and $\hat{\boldsymbol{p}}_d$) are different across the models with different assumptions (I refer to the discussion comparing the usual multinomial model, and the model with rare disease and G-E independence in Section 5.2.1). This simply means that I can apply the corrected counts instead of the observed counts $\boldsymbol{r}_d$, to the estimates obtained in Table 5–2 and that will lead to the exactly same estimates as described in Table 5–3. I emphasize that these estimators in Table 5–3 are only strictly valid as MLEs when they lie within the constrained parameter space. When the positivity constraints on the OR_{eg}^*'s or the probability constraints on the $\boldsymbol{p}_d^*$'s are violated (e.g. when very small values of sensitivity or specificity are used, corresponding to huge misclassification rates) then the constrained MLEs would be on the boundary of the parameter space, we should maximize the likelihood (5–9) with respect to the true parameters subject to the constraints, instead of transforming the observed MLE. However, as such estimates are indicative of extreme misclassification, or too small a sample, we might well treat these estimates with some caution.

We can also see the behavior of the estimators from Table 5–3 as the misclassification error rates go to 0 by Taylor series expansions of these estimators. Define the errors as $\varepsilon_G^p = 1-sp_{1G}$, $\varepsilon_G^e = 1-se_{1G}$, $\varepsilon_E^p = 1-sp_{dE}$ and $\varepsilon_E^e = 1-se_{1E}$. Expanding the log-scale estimator $\hat{\psi^*}^{IR}$ of the interaction parameter around $\varepsilon_G^p = \varepsilon_G^e = \varepsilon_E^p = \varepsilon_E^e = 0$,

Table 5–3: In the presence of misclassification, the MLEs of the true odds ratios in terms of estimated starred expected counts r^*_{dj} for the traditional model (Model 1) and $r^{*\,IR}_{dj}$ for the model under G-E independence and rare disease assumptions (Model 2).

Parameters		Model 1	Model 2
OR^*_{10}	MLE	$\frac{r^*_{01}r^*_{12}}{r^*_{02}r^*_{11}}$	$\frac{r^{*\,IR}_{01}r^{*\,IR}_{12}}{r^{*\,IR}_{02}r^{*\,IR}_{11}}$
OR^*_{01}	MLE	$\frac{r^*_{01}r^*_{13}}{r^*_{03}r^*_{11}}$	$\frac{r^{*\,IR}_{01}r^{*\,IR}_{13}}{r^{*\,IR}_{03}r^{*\,IR}_{11}}$
ψ^*	MLE	$\frac{r^*_{02}r^*_{03}r^*_{11}r^*_{14}}{r^*_{01}r^*_{04}r^*_{12}r^*_{13}}$	$\frac{r^{*\,IR}_{11}r^{*\,IR}_{14}}{r^{*\,IR}_{12}r^{*\,IR}_{13}}$

we see that

$$\begin{aligned}
&\log(\hat{\psi^*}^{IR}) = \log\left(\frac{r^{*\,IR}_{11}r^{*\,IR}_{14}}{r^{*\,IR}_{12}r^{*\,IR}_{13}}\right)\\
= \;&\log\left(\frac{r_{11}r_{14}}{r_{12}r_{13}}\right) + \frac{(r_{11}r_{14}-r_{12}r_{13})(r_{13}r_{14}\varepsilon^p_G + r_{12}r_{14}\varepsilon^p_E + r_{11}r_{12}\varepsilon^e_G + r_{11}r_{13}\varepsilon^e_E)}{r_{11}r_{12}r_{13}r_{14}}\\
&+\text{higher order terms.}
\end{aligned} \tag{5–11}$$

To a first-order approximation, the estimator reduces to the normal, perfect-data as in Table 5–2 as the errors reduce to 0. The first-order terms suggest that using a good approximation to errors may give better estimates than simply ignoring misclassification, i.e., setting the errors equal to 0.

Construction of the confidence intervals follows in exactly the same way as for the perfectly classified data with the standard error estimates obtained from the inverse of the information matrix of L_2 evaluated at the MLEs.

5.2.3 Case-only Method with Possible Misclassification

As discussed in detail in Chapter 4, the case-only method (Piegorsch *et al.* 1994) is a popular method to estimate the multiplicative gene-environment interaction parameter ψ, where under the rare disease and gene-environment independence assumptions, the odds ratio of G for exposed versus unexposed subjects among the cases only provides an efficient estimate of the interaction parameter ψ. The data used is as shown in the second raw of Table 5–1, ignoring the control data on the first row.

In the absence of misclassification, data from the case population forms a multinomial distribution, $\boldsymbol{r}_1 \sim Mn(n_1, \boldsymbol{p}_1)$, where n_1 is fixed. The interaction parameter (here denoted as ψ_{CO}) is obtained as the odds ratio between G and E among the case population, i.e.,

$$\psi_{CO} = \frac{p_{14}/p_{12}}{p_{13}/p_{11}}.$$

Together with $\sum_{j=1}^{4} p_{1j} = 1$, I have the following restrictions for $\boldsymbol{p}_1$,

$$\begin{aligned} p_{13} &= \frac{p_{11}(1 - p_{11} - p_{12})}{p_{11} + p_{12}\psi_{CO}} \\ p_{14} &= \frac{p_{12}\psi_{CO}(1 - p_{11} - p_{12})}{p_{11} + p_{12}\psi_{CO}}. \end{aligned}$$

The corresponding likelihood for the case-only method is thus

$$L_{CO} = L(\psi_{CO}, p_{11}, p_{12}|\boldsymbol{r}_1) = \prod_{j=1}^{4} p_{1j}^{r_{1j}},$$

and the MLE of the interaction parameter ψ_{CO} is

$$\hat{\psi}_{CO} = \frac{r_{11}r_{14}}{r_{12}r_{13}}$$

with variance $\hat{\psi}_{CO}^2(\sum_{j=1}^{4} 1/r_{1j})$.

Both Results 1 and 2 hold for $d = 1$ as well, therefore, estimating the true parameters in the presence of misclassification is straightforward by writing the likelihood in terms of the true parameters

$$L^*_{CO} = L(\psi^*_{CO}, p^*_{11}, p^*_{12}|\boldsymbol{r}_1) = \prod_{j=1}^{4} \{p_{1j}(\boldsymbol{p}^*_1)\}^{r_{1j}}.$$

Note $\hat{p}_{1j} = r_{1j}/n_1$, so the MLE of the "true" parameter ψ^*_{CO} in terms of $\boldsymbol{r}_1^{*CO}$ (the superscript CO is to distinguish from the other models) is

$$\hat{\psi}^*_{CO} = r_{11}^{*\,CO} r_{14}^{*\,CO} / (r_{12}^{*\,CO} r_{13}^{*\,CO}),$$

and ${r_{1j}^*}^{CO}$ can be obtained following Result 2, with $\tilde{r}_{1j} = n_1\hat{p}_{1j} = r_{1j}$. The variance estimators can be again estimated from the inverse of the information matrix of L_{CO}^* evaluated at the MLEs or by the technique as stated in Appendix C.1.

REMARK 3: Note that, the MLE of the interaction parameter and its variance obtained by the case-only method are exactly the same as those obtained in Section 5.2.1, where I also assume G-E independence and rare disease assumptions, but use both case and control data. This is true whether in the absence of misclassification, or in the presence of misclassification (unadjusted or adjusted). This establishes yet another proof of the fact that the under G-E independence and rare disease assumptions, the interaction odds ratio is exactly equal to the odds ratio of E on G for cases alone.

The details of the simple equivalence result appears in Appendix C.3. Remark 3 shows that our model with the G-E independence and rare disease assumptions can also obtain a highly efficient estimate of the interaction parameter ψ as in the case-only method. Moreover, our model is also able to estimate the main effects of genetic and environmental factors, which the case-only method cannot estimate. As Clayton and McKeigue (2001) noted, studies of gene-environment association with disease need to go beyond the estimation of the statistical interaction parameter ψ, and our study can estimate auxiliary parameters of interest without compromising on efficiency of the estimate of the interaction parameter. I only point this out as a side observation, as the case-only method does not have any advantage compared to the proposed method.

5.3 Simulation Studies

In this section, I present numerical evidence in the form of simulation studies to illustrate the advantage of the proposed methods in unmatched case-control studies. Generally, I assume that the genetic variant of interest is a bi-allelic locus with the wild and variant type alleles. I consider a dominant model for the effect of the gene

variant. I also assume a binary environmental exposure and consider it as a commonly prevalent exposure. Specifically, I follow the similar simulation design as mentioned in Chatterjee *et al.* (2005).

I first generate the parental genotype data for each individual. Given the genotypes of the parents, I generate the genotype for one offspring based on a standard Mendelian mode of inheritance. I independently generate the environmental exposures for this offspring based on the marginal probability of exposure ($E = 1$) for the underlying population. Given the information of genetic and environmental factors, I generate the disease outcome for each individual, independent of other, using the logistic regression model

$$\log\left\{\frac{P(D=1|G,E,\beta_0)}{P(D=0|G,E,\beta_0)}\right\} = \exp\left\{\beta_0 + \beta_E E + \beta_G G + \beta_{GE} G \times E\right\}. \qquad (5\text{–}12)$$

I choose the main effect parameters of $\beta_E = \log(OR_{10}) = \log(2)$ and $\beta_G = \log(OR_{01}) = \log(2)$ and consider a multiplicative interaction between G and E, fixing $\beta_{GE} = \log(\psi) = \log(2)$. I select the value of β_0 so that the marginal probability of the disease in the population, $P(D = 1) \approx 0.01$ (in fact $P(D = 1|G = 0, E = 0) \approx 0.001$). Following this scheme, I first generate data for a large number of randomly samples, which I treat as the underlying population, I then select 1000 diseased individuals and 1000 non-diseased individuals.

I only retain the appropriate disease, genotype and environmental exposure information and discard the rest of the information. Following the definition of sensitivity and specificity, I randomly misclassify the genotype and environmental exposure information, independent of one another, but keep the disease information unchanged. In the simulation, I let $sp_{0G} = sp_{1G} = sp_{0E} = sp_{1E} = 1$ and in the sampled dataset consider the following settings: 1) $se_{0G} = se_{1G} = 0.95$ and $se_{0E} = se_{1E} = 0.9$; 2) $se_{0G} = se_{1G} = 0.9$ and $se_{0E} = se_{1E} = 0.8$.

For each scenario, I simulate 500 datasets and analyze the data by implementing the adjusted formulation. To illustrate the efficiency of the proposed method, I compare the results for both in the absence and in the presence of misclassification (unadjusted and adjusted). To exploit the G-E independence assumption, I also apply the formulation under all three model assumptions as discussed in Section 5.2. Tables 5–4, 5–5, 5–6 and 5–7 show the results of analyzing unmatched case-control data with different sample sizes (1000/1000 and 750/750) for different set-up of misclassification errors.

To summarize, in the presence of misclassification, the estimates without adjusting show high bias and have significantly large mean square of errors (MSE), but the standard errors are not necessarily bigger if comparing with the estimates in the absence of misclassification. In fact, I notice that the estimates of OR_{10} and ψ without adjustment are biased towards null. The adjusted estimates which are obtained through the proposed formulation are quite close to the true parameters, except with relative large standard errors, and the power to detect the interaction is also a significant improvement over the others.

Under each same scenario, I notice that the estimators obtained by the traditional model suffer more from misclassification than those obtained under the G-E independence assumption. Even in the absence of misclassification, the models under the independence assumption provide more precise estimates, i.e., smaller standard errors and MSE. As I expected, larger sample size improves the power of testing the interaction effect as well as the precision of the parameter estimates.

I also present the sample size calculations to achieve the designed power, following the approach described by Lubin and Gail (1990) and discussed by Garcia-Closas and Lubin (1999). These calculations are performed using the program POWER. Table 5–8 and 5–9 present the impact of reducing sensitivity/specificity of the environmental factor assessment from 1.0 to 0.90 and 0.80, both in the absence and in the presence

of reduced sensitivity/specificity in the assessment of the genetic factor (from 1.0 to 0.95 and 0.90). In Figure 5–1, I explore in more detail the effects of misclassification on sample size. The solid lines in Figure 5–1 represent the sample size required to detect the specified 2-fold interaction in the absence of misclassification as a function of the true prevalence of the environmental factor for 0.2 prevalence of the genetic factor. The other lines in Figure 5–1 illustrate the impact of misclassification of the environmental factor on sample size for selected values of sensitivity and specificity of exposure assessment.

5.4 Conclusion

I describe relatively simple analytic formulation for accounting for misclassification of exposures in studies of gene-environment interaction based on sensitivity and specificity of the measurement instrument for genetic and environmental factors in unmatched case-control studies. As illustrated in the simulations, even relatively small degrees of error (i.e., sensitivity or specificity quite close to 1), the estimates of parameters of interest have relative large biases. The corrected estimates minimize the biases and are found closer to true parameters, although the standard errors are slightly large when the sample size is relative small.

I also consider different model assumptions to exploit the G-E independence assumption. According to results of the simulation, I do suggest using the formulation under the G-E independence assumption if the assumption holds in the source population, since the traditional model suffers more from misclassification. Cautions while using this independence assumption as discussed in Chapter 4 should be exercised while using these methods.

Improvements of the accuracy of exposure assessment for both the environmental and genetic factors can greatly reduce sample size requirements to study interactions

and are critical for accurate assessment of gene-environment interactions in case-control studies. How to exploit G-E independence assumption for a matched case-control study and adjust for misclassification error is part of my future work. Rice (2003) has proposed an ingenious full-likelihood based method of adjusting for misclassification in any $n_1 : n_2$ matched case-control study with a single binary exposure which uses the mixing distributions characterized in Rice (2004). I would consider the novel conditional likelihood proposed by Chatterjee *et al.* (2005) for matched case-control studies, with which one can exploit the G-E independence assumption. Then I will attempt to establish equivalence of this new conditional likelihood with the integrated full-likelihood for a special class of invariant marginal distributions in the spirit of Rice (2004), based on which I then could consider a full-likelihood based approach to account for misclassification errors. The interpretation of the novel conditional likelihood in terms of a random effects on Bayesian viewpoint, would open up a new strategy for computing the point and interval estimates and make it easier to adjust for error due to misclassification. How to handle continuous exposures, non-ignorable missingness in exposure remains a topic of future research.

Table 5–4: Results of unmatched case-control data (750/750), where specificity for both genetic and environmental factor =1.0, $se_{0G} = se_{1G} = 0.95$ and $se_{0E} = se_{1E} = 0.9$. $P(D = 1) \approx 0.01$, $P(E = 1) \approx 0.5$ and $P(G = 1) \approx 0.2$

Assumptions	Misclassification		OR_{10} 2.0000	OR_{01} 2.0000	ψ 2.0000	Power $H_0 : \psi = 1$
No	No	MLE	2.0079	2.0206	2.0048	0.796
		s.e.	0.2729	0.3857	0.4893	
		MSE	0.0743	0.1467	0.2509	
	Yes & unadjusted	MLE	1.8761	2.2789	1.7075	0.588
		s.e.	0.2459	0.4008	0.4084	
		MSE	0.0727	0.2476	0.2475	
	Yes & adjusted	MLE	2.0077	2.0262	2.0453	0.817
		s.e.	0.2732	0.3871	0.5000	
		MSE	0.0957	0.2197	0.3517	
$G - E$ independence and rare disease	No	MLE	2.0112	2.0211	1.9488	0.763
		s.e.	0.2617	0.3424	0.3296	
		MSE	0.0656	0.1178	0.1128	
	Yes & unadjusted	MLE	1.8787	2.2755	1.6640	0.547
		s.e.	0.2355	0.3550	0.2614	
		MSE	0.0675	0.2056	0.1803	
	Yes & adjusted	MLE	2.0109	2.0194	1.9866	0.786
		s.e.	0.3032	0.3991	0.4270	
		MSE	0.0882	0.1678	0.1806	
$G - E$ independence and $P(D = 1)$ known	No	MLE	2.0020	2.0005	1.9900	0.788
		s.e.	0.2604	0.3388	0.3402	
		MSE	0.0649	0.1150	0.1175	
	Yes & unadjusted	MLE	1.8713	2.2567	1.6957	0.577
		s.e.	0.2345	0.3521	0.2699	
		MSE	0.0690	0.1936	0.1643	
	Yes & adjusted	MLE	1.9998	1.9951	2.0371	0.813
		s.e.	0.3016	0.3950	0.4432	
		MSE	0.0872	0.1641	0.1954	

Table 5–5: Results of unmatched case-control data (1000/1000), where specificity for both genetic and environmental factor =1.0, $se_{0G} = se_{1G} = 0.95$ and $se_{0E} = se_{1E} = 0.9$. $P(D = 1) \approx 0.01$, $P(E = 1) \approx 0.5$ and $P(G = 1) \approx 0.2$

Assumptions	Misclassification		OR_{10} 2.0000	OR_{01} 2.0000	ψ 2.0000	Power $H_0 : \psi = 1$
No	No	MLE	1.9912	2.0055	1.9828	0.887
		s.e.	0.2340	0.3306	0.4181	
		MSE	0.0539	0.1271	0.1931	
	Yes & unadjusted	MLE	1.8717	2.2592	1.6897	0.696
		s.e.	0.2122	0.3435	0.3495	
		MSE	0.0598	0.1836	0.2134	
	Yes & adjusted	MLE	2.0042	2.0105	2.0088	0.898
		s.e.	0.2360	0.3318	0.4243	
		MSE	0.0725	0.1463	0.2523	
$G - E$ independence and rare disease	No	MLE	1.9953	2.0070	1.9329	0.864
		s.e.	0.2244	0.2937	0.2823	
		MSE	0.0480	0.0959	0.0923	
	Yes & unadjusted	MLE	1.8797	2.2712	1.6308	0.636
		s.e.	0.2038	0.3067	0.2215	
		MSE	0.0551	0.1578	0.1841	
	Yes & adjusted	MLE	2.0141	2.0212	1.9295	0.862
		s.e.	0.2628	0.3454	0.3573	
		MSE	0.0685	0.1069	0.1307	
$G - E$ independence and $P(D = 1)$ known	No	MLE	1.9863	1.9876	1.9728	0.883
		s.e.	0.2234	0.2908	0.2913	
		MSE	0.0477	0.0944	0.0943	
	Yes & unadjusted	MLE	1.8724	2.2527	1.6613	0.667
		s.e.	0.2030	0.3042	0.2287	
		MSE	0.0566	0.1467	0.1657	
	Yes & adjusted	MLE	2.0041	1.9995	1.9765	0.884
		s.e.	0.2616	0.3423	0.3705	
		MSE	0.0677	0.1062	0.1365	

Table 5–6: Results of unmatched case-control data (750/750), where specificity for both genetic and environmental factor =1.0, $se_{0G} = se_{1G} = 0.9$ and $se_{0E} = se_{1E} = 0.8$. $P(D = 1) \approx 0.01$, $P(E = 1) \approx 0.5$ and $P(G = 1) \approx 0.2$

Assumptions	Misclassification		OR_{10} 2.0000	OR_{01} 2.0000	ψ 2.0000	Power $H_0 : \psi = 0$
No	No	MLE	2.0079	2.0206	2.0048	0.796
		s.e.	0.2729	0.3857	0.4893	
		MSE	0.0743	0.1467	0.2509	
	Yes & unadjusted	MLE	1.7950	2.4692	1.5001	0.386
		s.e.	0.2305	0.4084	0.3613	
		MSE	0.0954	0.3965	0.4081	
	Yes & adjusted	MLE	2.0462	2.1126	2.0163	0.801
		s.e.	0.2788	0.4021	0.4943	
		MSE	0.1405	0.3561	0.6661	
$G - E$ independence and rare disease	No	MLE	2.0112	2.0211	1.9488	0.763
		s.e.	0.2617	0.3424	0.3296	
		MSE	0.0656	0.1178	0.1128	
	Yes & unadjusted	MLE	1.7949	2.4612	1.4584	0.343
		s.e.	0.2203	0.3631	0.2203	
		MSE	0.0891	0.3460	0.3486	
	Yes & adjusted	MLE	2.0456	2.0927	1.9369	0.755
		s.e.	0.3581	0.4797	0.5194	
		MSE	0.1253	0.2646	0.3227	
$G - E$ independence and $P(D = 1)$ known	No	MLE	2.0020	2.0005	1.9900	0.788
		s.e.	0.2604	0.3388	0.3402	
		MSE	0.0649	0.1150	0.1175	
	Yes & unadjusted	MLE	1.7890	2.4454	1.4830	0.369
		s.e.	0.2195	0.3608	0.2275	
		MSE	0.0912	0.3303	0.3265	
	Yes & adjusted	MLE	2.0328	2.0646	1.9934	0.789
		s.e.	0.3560	0.4748	0.5418	
		MSE	0.1232	0.2564	0.3492	

Table 5–7: Results of unmatched case-control data (1000/1000), where specificity for both genetic and environmental factor =1.0, $se_{0G} = se_{1G} = 0.9$ and $se_{0E} = se_{1E} = 0.8$. $P(D = 1) \approx 0.01$, $P(E = 1) \approx 0.5$ and $P(G = 1) \approx 0.2$

Assumptions	Misclassification		OR_{10} 2.0000	OR_{01} 2.0000	ψ 2.0000	Power $H_0 : \psi = 0$
No	No	MLE	1.9912	2.0055	1.9828	0.887
		s.e.	0.2340	0.3306	0.4181	
		MSE	0.0539	0.1271	0.1931	
	Yes & unadjusted	MLE	1.7887	2.4583	1.4780	0.460
		s.e.	0.1988	0.3524	0.3074	
		MSE	0.0830	0.3674	0.3774	
	Yes & adjusted	MLE	2.0307	2.0873	1.9669	0.880
		s.e.	0.2391	0.3445	0.4161	
		MSE	0.1053	0.2834	0.4169	
$G - E$ independence and rare disease	No	MLE	1.9953	2.0070	1.9329	0.864
		s.e.	0.2244	0.2937	0.2823	
		MSE	0.0480	0.0959	0.0923	
	Yes & unadjusted	MLE	1.7830	2.4338	1.4744	0.455
		s.e.	0.1894	0.3111	0.1928	
		MSE	0.0828	0.3014	0.3155	
	Yes & adjusted	MLE	2.0218	2.0470	1.9525	0.873
		s.e.	0.3054	0.4074	0.4493	
		MSE	0.0985	0.1968	0.2135	
$G - E$ independence and $P(D = 1)$ known	No	MLE	1.9863	1.9876	1.9728	0.883
		s.e.	0.2234	0.2908	0.2913	
		MSE	0.0477	0.0944	0.0943	
	Yes & unadjusted	MLE	1.7770	2.4179	1.4990	0.487
		s.e.	0.1887	0.3091	0.1991	
		MSE	0.0852	0.2863	0.2927	
	Yes & adjusted	MLE	2.0092	2.0195	2.0086	0.898
		s.e.	0.3037	0.4033	0.4686	
		MSE	0.0971	0.1916	0.2302	

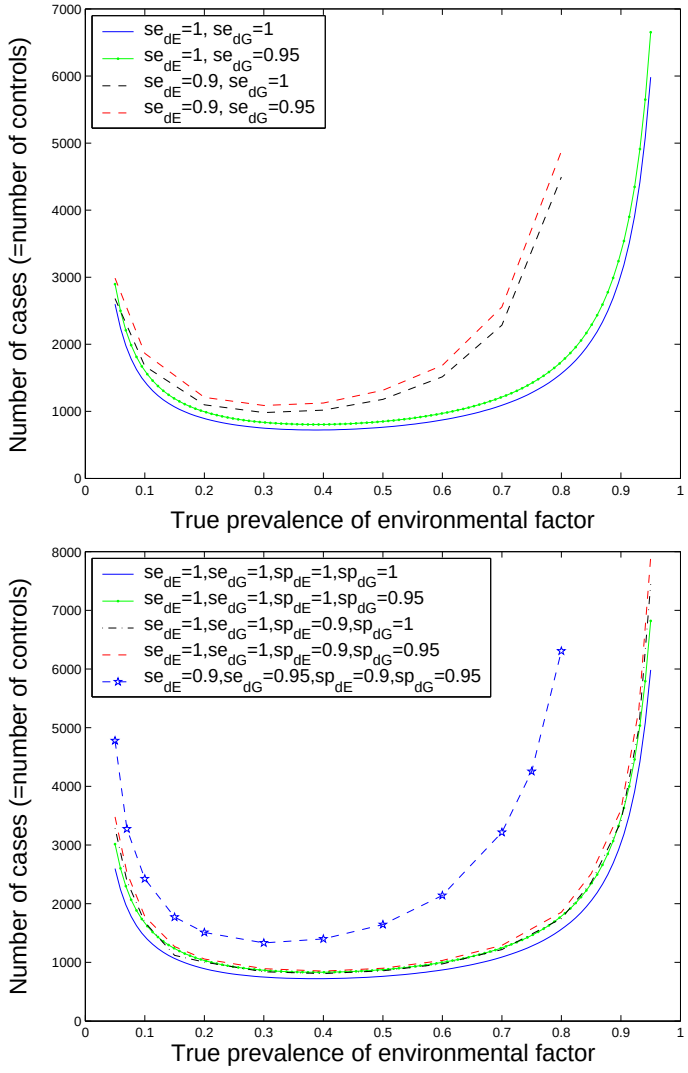
Table 5–8: Minimum number of cases (case:control ratio=1) required to detect a 2-fold multiplicative interaction ($OR_{10} = OR_{01} = 2$ and $OR_{11} = 8$) with 80% power for different levels of sensitivities and specificities of the environmental and genetic factors, where $P(E = 1) = 0.5$ and $P(G = 1) = 0.2$.

se_{dE}	se_{dG}	sp_{dE}	sp_{dG}	No. of cases
1	1	1	1	763
0.9	1	1	1	1178
1	0.95	1	1	847
0.9	0.95	1	1	1314
0.8	0.9	1	1	2077
1	1	0.9	1	857
1	1	1	0.95	802
1	1	0.9	0.95	902
1	1	0.8	0.9	1096
0.9	0.95	0.9	0.95	1639
0.8	0.9	0.9	0.95	2830

Table 5–9: Minimum number of cases (case:control ratio=1) required to detect a 3-fold multiplicative interaction ($OR_{10} = 1.3$, $OR_{01} = 7$ and $OR_{11} = 3$) with 80% power for different levels of sensitivities and specificities of the environmental and genetic factors, where $P(E = 1) = 0.2$ and $P(G = 1) = 0.01$.

se_{dE}	se_{dG}	sp_{dE}	sp_{dG}	No. of cases
1	1	1	1	3771
0.9	1	1	1	4563
1	0.95	1	1	3828
0.9	0.95	1	1	4478
0.8	0.9	1	1	5728
1	1	0.9	1	4543
1	1	1	0.95	1793
1	1	0.9	0.95	2235
1	1	0.8	0.9	3242
0.9	0.95	0.9	0.95	2944
0.8	0.9	0.9	0.95	3869

Figure 5–1: Minimum number of cases (case:control ratio = 1) required to detect a 2-fold interaction ($OR_{10} = 2$, $OR_{01} = 2$, and $OR_{11} = 8$) with 80% power as a function of the true prevalence of the environmental factor, P(E=1), for the prevalence of the genetic factor being 0.2, and for selected values of sensitivity and specificity of the exposure assessment.



CHAPTER 6
FUTURE WORK AND CONCLUSION

I plan to extend the work related to this dissertation in two principal directions. The first one is for estimating covariate effects in family-based case-control studies. Genetic epidemiology is a relatively new field which applies the conventional epidemiologic designs and methods to explore the role genetic factors play in determining a disease. Both theoretical and empirical studies have shown that traditional linkage studies may be inferior in power compared to studies directly utilizing allele status. As an alternative, population-based case-control association studies are subject to bias due to population stratification. As a compromise between linkage studies and population-based case-control studies, family-based association designs have received great attention recently due to their potentially higher power to identify complex disease genes and their robustness in the presence of population substructure (Zhao 2000).

A common phenomenon in genetic epidemiologic research is that sampled families are not representative of the targeted population as they are ascertained through probands with known phenotypic values. It is well known in the literature that statistical inference without proper ascertainment corrections would lead to biased estimation of key parameters. One simple remedy is to condition on the observed phenotypic values of the probands (case or control). In family-based case-control studies, a natural approach to account for family effect will be to conduct a matched case-control analysis with controls selected from the same family, and to use conditional logistic regression conditional on the number of cases in the family (there could be more than one person affected by the disease in a family). It has become increasingly clear that both genetic and environmental factors contribute to the aetiology of many

common diseases. If environmental exposures or other covariates are also important, they should be incorporated into these genetic analyses to control for confounding and increase statistical power.

Diggle *et al.* (1994, Ch. 9) present generalized linear mixed models for exponential families with canonical links. Witte *et al.* (1999) used a fixed effects logistic model in which the familial genetic effects were measured, rather than random effects, to analyze family studies. Pfeiffer *et al.* (2001) proposed the following two-level mixed effects model to account for common familial effects and for different genetic correlations among family members. Considering a binary disease variable D_{ij} for the jth member of the ith family and a covariate variable X_{ij}, for $j = 1, \cdots, n_i$ and $i = 1, \cdots, N$, let

$$\log\left\{\frac{p_{ij}}{1-p_{ij}}\right\} = \log\left\{\frac{\Pr(D_{ij}=1|a_i, g_{ij}, X_{ij})}{\Pr(D_{ij}=0|a_i, g_{ij}, X_{ij})}\right\} = \mu + a_i + g_{ij} + \beta X_{ij}. \qquad (6\text{–}1)$$

Here a_i denotes the random familial effect, which affects all family members equally and g_{ij} denotes an individual random genetic effect. They based their analysis on the marginal likelihood after integrating with respect to the joint random effects distribution. They adjusted for ascertainment by conditioning on the number of cases, $k_i(\geq 2)$, in the family and performed conditional maximum likelihood analysis based on this mixed effects model. The conditional distribution for family i can be written as

$$\begin{aligned} & \Pr(D_{i1}, \cdots, D_{in_i} | X_{i1}, \cdots, X_{in_i}, \sum_{j=1}^{n_i} D_{ij} = k_i) \\ = \; & \frac{\Pr(D_{i1}, \cdots, D_{in_i}, \sum_{j=1}^{n_i} D_{ij} = k_i | X_{i1}, \cdots, X_{in_i})}{\Pr(\sum_{j=1}^{n_i} D_{ij} = k_i | X_{i1}, \cdots, X_{in_i})}. \end{aligned} \qquad (6\text{–}2)$$

This approach allows to estimate environmental effects while accounting for varying genetic correlations among family members. This approach took into account of unmeasured familial and genetic effects that induce correlated responses and yielded

consistent estimators of covariate effects when the covariate has no effect on disease status even with a misspecified random effects distribution.

The approach presented in Pfeiffer *et al.* (2001) has flexibility but comes with the drawback of computational complexity as they have to approximate the integrated likelihood on a dense grid of points. Although the Monte Carlo approach worked well in their examples, larger Monte Carlo samples or other methods may be needed for larger pedigrees. I want to propose a full Bayesian alternative to build in a hierarchical pedigree structure and assume priors on the random effects which offers a more unified and computationally appealing alternative.

As Pfeiffer *et al.* (2001) pointed out that more fundamentally, often one does not know the precise nature of the genetic influences and hence the distribution of the familial individual genetic effects. The estimated random effect for each individual family will change by changing the distributions of the random effects. This point is important because there are many applications in which an estimate of the random effect itself is desired. For example, when the true distributions of the random effects are mixtures of normals, assuming that it is a single unmixed normal can lead to poor estimates of the random effects. In such applications, unbiased estimation of the random effects is crucial and the assumption of normality may introduce bias. From a Baysian perspective, inferential interest focuses on the posterior distribution of the fixed effects as well as the random effects. Allowing distributions other than the normal for the random effects may more accurately model our prior beliefs. It is also important to accurately model the distribution of the random effects when prediction for a future observation from a given subject is desired. I will provide a general framework for Bayesian analysis for the random effects model where a nonparametric Dirichlet Process (DP) prior is specified for the random effects.

Pfeiffer *et al.* (2001) modeled the covariance matrix of the familial genetic random effects as a function of the degree of kinship between members in each family

by assuming no dominance component of the genetic variance (Fisher 1918). Their inference about the random effects was reduced to only a scalar common variance σ_g^2. I would instead consider a more complicated covariance structure with the parental variance, siblings variance, the interclass (parent-sib) correlation and the intraclass (sib-sib) correlation as usually used for multivariate familial data (Srivastava 1984, Srivastava *et al.* 1988). Therefore, I would make the inference on covariance parameters related to the family data. This parametric covariance has a flexible degree of kinship and also can be easily extended to families of larger pedigrees.

I will extend my work in Chapter 5 to the case of matched case-control set-up, in particular to family-based case-control studies. Conditional logistic regression, the traditional method of analysis of the family-based case-control data fails to exploit the assumption that genetic susceptibility and environmental exposures are independently distributed of each other within families in the source population, and hence the traditional method can be inefficient. Alternatively, Chatterjee *et al.* (2005) proposed novel methods for the analysis of family-based case-control studies under the *G*-*E* independence assumption within families in the source population. The assumption of *G*-*E* independence within families in the source population is relatively weaker than our assumption in Chapter 4, as this assumption is not affected by spurious association between genotype and exposure status that may be created due to population substructure (Umbach and Weinberg, 2000). This approach leads to a simple and yet highly efficient methods of estimating interaction and various other risk parameters of scientific interest.

Chatterjee *et al.* (2005) leave the question of misclassification errors unaddressed, while they develop the classical asymptotic efficiency theory. Rice (2003) has proposed an ingenious full-likelihood based method of adjusting for misclassification in any $n_1 : n_2$ matched case-control study with a single binary exposure which uses the mixing distributions characterized in Rice (2004). I would first extend the result

of Rice (2004) for data with no misclassification and characterize a class of random effects mixing distributions for the nuisance parameters involved in the full likelihood such that when they are integrated out, the integrated likelihood is identical to the novel conditional likelihood proposed by Chatterjee *et al.* (2005) under a rare disease assumption. The established equivalence is a significant and exciting finding in itself, which renders a Bayesian interpretation to the novel conditioning paradigm proposed by Chatterjee *et al.* (2005), exactly as Rice (2004) provides a Bayesian interpretation to the traditional conditional likelihood. The estimation of parameters can then be carried out using MCMC techniques as suggested by Rice (2004). Characterizing the class of random effects distributions will provide a method to adjust for misclassification error based on the full likelihood which will be designed to include misclassification effects. I would obtain estimates and confidence intervals for the misclassified case which reduce back to ones obtained via Chatterjee *et al.*'s conditional likelihood as misclassification error rates go to zero.

To conclude, in this dissertation I address the important foundational issue of equivalence of prospective and retrospective analysis in a Bayesian framework. I consider some new problems in the domain of case-control methodology which are emerging with modern advances in genetic technology. To popularize the use of Bayesian methods in this area, one needs wider dissemination of user-friendly software codes. A broader goal of my future work will be to make these codes available online in a usable form.

APPENDIX A
APPENDIX TO CHAPTER 3

This is an explanation to the approximation in equation (3–9). From equation (3–7), by the strong law of large numbers,

$$\frac{1}{P(\boldsymbol{X}|I)} \overset{a.s.}{\to} E\Big(\frac{1}{Y}\Big),$$

where $Y = P(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{P}, \boldsymbol{Q})$. This implies that,

$$-2\log P(\boldsymbol{X}|I) \overset{a.s.}{\to} 2\log E\Big(\frac{1}{Y}\Big), \tag{A–1}$$

Let $W = -2\log Y$=$[DV|\boldsymbol{X}]$, then

$$E\Big(\frac{1}{Y}\Big) = E_Y(\exp(-\log Y)) = M_W(\frac{1}{2}), \tag{A–2}$$

where $M_W(t)$ denotes the moment generating function of the distribution of W. By assuming that the deviance function $[DV|\boldsymbol{X}]$ is normal, i.e., $W \sim N(\mu, \sigma^2)$, by (A–2), I have,

$$E\Big(\frac{1}{Y}\Big) = \exp\big(\mu/2 + \sigma^2/8\big).$$

Hence by (A–1), and the fact that $\hat{\mu}$ and $\hat{\sigma}^2$ are consistent estimates of μ and σ^2, I have the approximation in (3–9).

Remark: Suppose I assume instead of normality of the deviance function, that $[DV|\mathbf{X}]$=$W = -2\log Y \sim Gamma(\mu^2/\sigma^2, \sigma^2/\mu)$, where μ and σ^2 are the mean and variance of W, and Gamma(a,b) denotes a Gamma distribution with shape parameter a and scale parameter b. Then by steps similar as above, one will obtain an analogue of (3–9) in the Gamma case as,

$$-2\log P(\boldsymbol{X}|I) \approx -2\hat{\mu}^2/\hat{\sigma}^2 \log\Big(1 - \frac{\hat{\sigma}^2}{2\hat{\mu}}\Big), \quad \text{for } \frac{\hat{\sigma}^2}{\hat{\mu}} < 2.$$

APPENDIX B
APPENDIX TO CHAPTER 4

B.1 Proof of Lemmas and Results

Proof of Lemma 1

$$\begin{aligned}
& \frac{P(G=g_m|E,\boldsymbol{S},D=1)}{P(G=g_m|E,\boldsymbol{S},D=0)} \\
= & \frac{P(D=1|G=g_m,E,\boldsymbol{S})P(G=g_m|E,\boldsymbol{S})/P(D=1|E,\boldsymbol{S})}{P(D=0|G=g_m,E,\boldsymbol{S})P(G=g_m|E,\boldsymbol{S})/P(D=0|E,\boldsymbol{S})} \\
= & \frac{P(D=1|G=g_m,E,\boldsymbol{S})/P(D=0|G=g_m,E,\boldsymbol{S})}{P(D=1|E,\boldsymbol{S})/P(D=0|E,\boldsymbol{S})}.
\end{aligned}$$

Proof of Lemma 2

I use the following identity and apply Lemma 1.

$$\begin{aligned}
1 &= \sum_{m=0}^{M} P(G=g_m|D=1,E,\boldsymbol{S}) \\
&= \sum_{m=0}^{M} \frac{P(G=g_m|D=1,E,\boldsymbol{S})}{P(G=g_m|D=0,E,\boldsymbol{S})} P(G=g_m|D=0,E,\boldsymbol{S}),
\end{aligned}$$

Proof of Lemma 3

I begin with the identity

$$\frac{p(E|\boldsymbol{S},D=1)}{p(E|\boldsymbol{S},D=0)} = \frac{P(D=1|E,\boldsymbol{S})/P(D=0|E,\boldsymbol{S})}{P(D=1|\boldsymbol{S})/P(D=0|\boldsymbol{S})}.$$

Then I observe that

$$\begin{aligned}
P(D=1|\boldsymbol{S}) &= \int P(D=1|E,\boldsymbol{S})p(E|\boldsymbol{S})dE \\
&= \int \frac{P(D=1|E,\boldsymbol{S})}{P(D=0|E,\boldsymbol{S})} P(D=0|E,\boldsymbol{S})p(E|\boldsymbol{S})dE \\
&= \int \frac{P(D=1|E,\boldsymbol{S})}{P(D=0|E,\boldsymbol{S})} p(E|D=0,\boldsymbol{S})P(D=0|\boldsymbol{S})dE.
\end{aligned}$$

Proof of Result 1

By Lemma 2 and (4–1) one gets,

$$\begin{aligned}\frac{P(D=1|E,Z=i)}{P(D=0|E,Z=i)} &= \sum_{m=0}^{M}\exp(\beta_{0i}+\beta_{1m}+\beta_2 E+\beta_{3m}E)\times\frac{\exp(\gamma_{im})}{\sum_{k=0}^{M}\exp(\gamma_{ik})} \\ &= \frac{\exp(\beta_{0i}+\beta_2 E)\{1+\sum_{k=1}^{M}\exp(\beta_{1k}+\beta_{3k}E+\gamma_{ik})\}}{1+\sum_{k=1}^{M}\exp(\gamma_{ik})} \quad \text{(B–1)}\end{aligned}$$

Substituting model (4–1) and (B–1) into Lemma 1, I get

$$\begin{aligned}\frac{P(G=g_m|D=1,E,Z=i)}{P(G=g_m|D=0,E,Z=i)} &= \frac{\exp(\beta_{0i}+\beta_{1m}+\beta_2 E+\beta_{3m}E)}{\frac{\exp(\beta_{0i}+\beta_2 E)(1+\sum_{k=1}^{M}\exp(\beta_{1k}+\beta_{3k}E+\gamma_{ik}))}{\{1+\sum_{k=1}^{M}\exp(\gamma_{ik})\}}} \\ &= \frac{\exp(\beta_{1m}+\beta_{3m}E)\{1+\sum_{k=1}^{M}\exp(\gamma_{ik})\}}{1+\sum_{k=1}^{M}\exp(\beta_{1k}+\beta_{3k}E+\gamma_{ik})}.\end{aligned}$$

Now by (4–4), I get the result.

Proof of Result 2

Note that by (4–5), (4–6) and Lemma 2,

$$\begin{aligned}&\frac{P(D=1|E,Z=i)}{P(D=0|E,Z=i)}p(E|D=0,Z=i,\theta_i=\boldsymbol{\omega}_l) \\ =\ &\frac{\exp\{\beta_{0i}+\beta_2 E\}\{\sum_{k=0}^{m}\exp(\beta_{1m}+\beta_{3m}E+\gamma_{im})\}}{\sum_{k=0}^{M}\exp(\gamma_{ik})}\times\frac{1}{\sqrt{2\pi\sigma_l^2}}\exp\Big\{-\frac{(E-\mu_l)^2}{2\sigma_l^2}\Big\} \\ =\ &\frac{\exp(\beta_{0i})}{\sum_{k=0}^{M}\exp(\gamma_{ik})}\sum_{m=0}^{M}\Big[\exp\Big\{\beta_{1m}+\gamma_{im}+\frac{(\mu_l+\beta_2\sigma_l^2+\beta_{3m}\sigma_l^2)^2-\mu_l^2}{2\sigma_l^2}\Big\}\phi(E;\boldsymbol{\omega}_{lm}^*)\Big]\end{aligned}$$

where $\boldsymbol{\omega}_{lm}^* = (\mu_l+\beta_2\sigma_l^2+\beta_{3m}\sigma_l^2, \sigma_l^2)$. Now, by Lemma 3 and the above two results, I get

$$p(E|D=1,Z=i,\theta_i=\boldsymbol{\omega}_l)=\sum_{m=0}^{M}\underbrace{\frac{\exp\{\beta_{1m}+\gamma_{im}+\frac{(\mu_l+\beta_2\sigma_l^2+\beta_{3m}\sigma_l^2)^2}{2\sigma_l^2}\}}{\sum_{k=0}^{M}\exp\{\beta_{1k}+\gamma_{ik}+\frac{(\mu_l+\beta_2\sigma_l^2+\beta_{3k}\sigma_l^2)^2}{2\sigma_l^2}\}}}_{p_{ilm}}\times\phi(E;\boldsymbol{\omega}_{lm}^*).$$

B.2 Likelihood for The EDPM Model

The components in the retrospective likelihood (4–3) are as the follows:

$$\begin{aligned} & P(G_j = g_m | Z_j = i, D_j = 0) P(E_j | Z_j = i, D_j = 0) \\ = & \frac{\exp\{\gamma_{im} - (E_j - \mu_l)^2/(2\sigma_l^2)\}}{\sqrt{2\pi\sigma_l^2}\sum_{k=0}^{M}\exp(\gamma_{ik})}; \\ & P(G_j = g_m | E_j, Z_j = i, D_j = 1) P(E_j | Z_j = i, D_j = 1) \\ = & \frac{\exp\{\beta_{1m} + \beta_2 E_j + \beta_{3m} E_j + \gamma_{im} - (E_j^2 - 2\mu_l E_j)/(2\sigma_l^2)\}}{\sum_{k=0}^{M}\exp\{\beta_{1k} + \gamma_{ik} + (\mu_l + \beta_2\sigma_l^2 + \beta_{3k}\sigma_l^2)/(2\sigma_l^2)\}}. \end{aligned}$$

B.3 Computational Details

For the Bayesian methods I need to simulate random numbers from the full conditional distributions of the parameters given the data. When the conditional distributions do not have a standard form I use independence sampler Metropolis-Hastings algorithm to generate random numbers from the respective conditionals. Under the EDPM model, a cycle of Gibbs sampler consists of the following steps.

Step 1. Drawing β_{1m}, β_{3m}, γ_{im}, $(m = 1, \cdots, M,\ i = 1, \cdots, I)$ and β_2 following the usual Metropolis-Hastings algorithm;

Step 2. Drawing observations from the posterior of DPM, following the no gaps algorithm proposed by MacEachern and Müller (1998) as the following:

Let $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_K)$ denote the set of distinct $\boldsymbol{\theta}_i$'s, where $K \leq I$ is the number of distinct elements in the vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_I)$. Let $\boldsymbol{s} = (s_1, \cdots, s_I)$ denotes the vector of configuration indicators defined by $s_i = k$ if and only if $\boldsymbol{\theta}_i = \boldsymbol{\omega}_k$, $i = 1, \cdots, I$. In this connection I use the term "cluster" where k^{th} cluster is defined as $I_k = \{i : s_i = k\}$ and define n_k as the size of the k^{th} cluster, and therefore, $\sum_{k=1}^{K} n_k = I$. Now it is obvious that $\boldsymbol{\omega}$ and $\boldsymbol{s}$ uniquely determine $\boldsymbol{\theta}$. However, to determine $\boldsymbol{\omega}$ and $\boldsymbol{s}$ uniquely from Θ I need to redefine $\boldsymbol{\omega}$ as follows.

Define $\boldsymbol{\omega}_1 = \boldsymbol{\theta}_1$ and for $j \geq 2$ $\boldsymbol{\omega}_j = \boldsymbol{\theta}_l$, where $l = \min\{r : \boldsymbol{\theta}_r \neq \boldsymbol{\omega}_1, \cdots,\ \boldsymbol{\theta}_r \neq \boldsymbol{\omega}_{j-1}\}$.

Now, instead of simulating $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_I)$ directly I simulate $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_K)$ and $\boldsymbol{s} = (s_1, \cdots, s_I)$ as they are in 1-1 relation.

I use "$-i$" to denote the situation when the observation i is removed. For example, $\boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \cdots, \boldsymbol{\theta}_I)$, K_{-i} denotes the number of clusters formed by $\boldsymbol{\theta}_{-i}$. Likewise I define $\boldsymbol{\omega}_{-i}$ and $n_{-i,k}$ as the distinct components in $\boldsymbol{\theta}_{-i}$ and cluster sizes after removing $\boldsymbol{\theta}_i$ Just to make a note that $\sum_{r=1}^{K} n_{-ir} = I - 1$.

1. Update s_i drawing from $[s_i|\boldsymbol{s}_{-i}, \boldsymbol{\omega}, \text{Data}]$ for every $i = 1, \cdots, I$

$$p(s_i = l|\boldsymbol{s}_{-i}, \boldsymbol{\omega}, \text{Data}) \propto p(s_i = l|\boldsymbol{s}_{-i}, \boldsymbol{\omega}) \prod_{j:Z_j=i} \Pr(E_j|Z_j, D_j, \boldsymbol{\theta}_i = \boldsymbol{\omega}_l),$$

   where $p(s_i = l|\boldsymbol{s}_{-i}, \boldsymbol{\omega})$ is defined below.

   • When $n_{s_i} > 1$, then

$$p(s_i = l|\boldsymbol{s}_{-i}, \boldsymbol{\omega}) = \begin{cases} cn_{-i,l} & \text{for} l = 1, 2, \cdots, K_{-i} \\ c\frac{\alpha}{K_{-i}+1} & \text{for} l = K_{-i} + 1 \end{cases} \tag{B–2}$$

   Note that if s_i happens to be $K_{-i} + 1$, then $\boldsymbol{\omega}_{K_{-i}+1}$ is simply a random draw from $\mathcal{P}_0(\cdot)$.

   • If $n_{s_i} = 1$ then $K_{-i} = K - 1$ and with probability $(K - 1)/K$ leave s_i unchanged, i.e., $\boldsymbol{\theta}_i = \boldsymbol{\omega}_{s_i}$; otherwise relabel the clusters such that $s_i = K$ and then resample s_i with probabilities in (B–2). Now if the new s_i happened to be $K_{-i} + 1 = K$, then the preceding relabeling keeps the previous values of $\boldsymbol{\theta}_i$ as $\boldsymbol{\omega}_k$ and nothing is changed except possible relabeling of $\boldsymbol{\omega}$ and hence of $\boldsymbol{s}$. If the new $s_i \leq K_{-i}$, the last element after relabeling in $\boldsymbol{\omega}$ is discarded.

2. Once the configuration indicators and the associated clusters are determined I move on to update $\boldsymbol{\omega}$'s. The full conditional distribution of $\boldsymbol{\omega}_l$ is

$$[\boldsymbol{\omega}_l|\boldsymbol{s}, \boldsymbol{\beta}_1, \beta_2, \boldsymbol{\beta}_3, \boldsymbol{\gamma}] \propto d\mathcal{P}_0(\boldsymbol{\omega}_l) \prod_{\{i:s_i=l\}} \prod_{j:Z_j=i} \Pr(E_j|Z_j, D_j, \boldsymbol{\theta}_i = \boldsymbol{\omega}_l), \tag{B–3}$$

   which is not in standard form, therefore I again use Metropolis-Hastings algorithm to update $\boldsymbol{\omega}_l$'s. Drawing a random number from the bivariate distribution

$\mathcal{P}_0(\cdot)$ is equivalent to first draw σ^2 from IG(s/2,S/2) and conditioned on σ^2, τ, and m_0, draw μ from $N(m_0, \tau\sigma^2)$.

Step 3. I update the hyperparameters as the following.

a) Conditional on $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_K)$ and τ, m_0 is conditionally independent of data and all other parameters, and follows a normal distribution with $E(m_0|\tau, \boldsymbol{\omega}) = (1-x)\mu_{m_0} + x\bar{V}\sum_{k=1}^{K}\mu_k/\sigma_k^2$ and $V(m_0|\tau, \boldsymbol{\omega}) = x\tau\bar{V}$, where $x = \sigma_{m_0}^2/(\sigma_{m_0}^2 + \tau\bar{V})$, and $\bar{V} = (\sum_{k=1}^{K}\sigma_k^{-2})^{-1}$.

b) Conditional on $\boldsymbol{\omega}$ and m_0, the distribution of τ is free of other parameters and the data and the full conditional distribution $\pi(\tau|\cdot)$ is $\text{IG}\Big(\frac{1}{2}(a_\tau + K), \frac{1}{2}(b_\tau + \sum_{k=1}^{K}\frac{(\mu_k - m_0)^2}{\sigma_k^2})\Big)$.

Step 4. We update the value of α in the following two steps.

i) Sample η from $p(\eta|\alpha, K) \propto \eta^\alpha(1-\eta)^{I-1}$;

ii) Sample α from $\pi_\eta\text{Gamma}(a_\alpha + K, b_\alpha - \log(\eta)) + (1-\pi_\eta)\text{Gamma}(a_\alpha + K - 1, b_\alpha - \log(\eta))$, where $\pi_\eta/(1-\pi_\eta) = (a_\alpha + K - 1)/\{I(b_\alpha - \log(\eta))\}$.

Convergence of the chain was assessed by computing the Gelman and Rubin (1992)(GR) diagnostic. The posterior means are taken the parameter estimates. To reduce auto-correlation among observations, I ran multiple chains and take every 5^{th} observation from the respective MCMC chain after the burn-in period of 30000 runs, and calculate the posterior mean, standard deviation and 95% HPD region based on observations from the last 10000 MCMC runs of each chain.

APPENDIX C
APPENDIX TO CHAPTER 5

Note: All the parameters are defined the same as in the text, except those defined separately here. Let $P(G=1)=q_G$ and $P(E=1)=q_E$.

C.1 The Constrained ML Equations under G-E Independence and Rare Disease Assumptions in Unmatched Case-Control studies

These constraint equations are obtained by differentiating the logarithm of the likelihood (5–3) with respect to the corresponding parameters:

$$\begin{aligned}
p^2(p_{04}r_{01}-p_{01}r_{04})+pp_{01}p_{04}(r_{11}-OR_{10}OR_{01}\psi r_{14}) &= p_{01}p_{04}(1-OR_{10}OR_{01}\psi)q \\
p^2(p_{04}r_{02}-p_{02}r_{04})+pp_{02}p_{04}OR_{10}(r_{12}-OR_{01}\psi r_{14}) &= p_{02}p_{04}(OR_{10}-OR_{10}OR_{01}\psi)q \\
p^2(p_{04}r_{03}-p_{03}r_{04})+pp_{03}p_{04}OR_{01}(r_{13}-OR_{10}\psi r_{14}) &= p_{03}p_{04}(OR_{01}-OR_{10}OR_{01}\psi)q \\
\frac{r_{12}+r_{14}}{OR_{10}}-\frac{(p_{02}+p_{04}OR_{01}\psi)n_1}{p} &= 0 \qquad \text{(A.1)} \\
\frac{r_{13}+r_{14}}{OR_{01}}-\frac{(p_{03}+p_{04}OR_{10}\psi)n_1}{p} &= 0 \qquad \text{(A.2)} \\
\frac{r_{14}}{\psi}-\frac{p_{04}OR_{10}OR_{01}n_1}{p} &= 0, \qquad \text{(A.3)}
\end{aligned}$$

where $q = p_{01}r_{11}+p_{02}OR_{10}r_{12}+p_{03}OR_{01}r_{13}+p_{04}OR_{10}OR_{01}\psi r_{14}$. Recall $p_{04} = 1-p_{01}-p_{02}-p_{03}$ and $p = p_{01}+p_{02}OR_{10}+p_{03}OR_{01}+p_{04}OR_{10}OR_{01}\psi$. The solutions to the above equations subject to the restriction of $p_{01}p_{04}=p_{02}p_{03}$. The following I show how to obtain those restricted MLEs:

1) Plugging (A.1), (A.2) and (A.3), I have

$$\begin{aligned}
\hat{p} &= \hat{p}_{01}+\widehat{OR}_{10}(\hat{p}_{02}+\hat{p}_{04}\widehat{OR}_{01}\hat{\psi})+\widehat{OR}_{01}(\hat{p}_{03}+\hat{p}_{04}\widehat{OR}_{10}\hat{\psi})-\hat{p}_{04}\widehat{OR}_{10}\widehat{OR}_{01}\hat{\psi} \\
&= \hat{p}_{01}+\frac{r_{12}+r_{14}}{n_1}\hat{p}+\frac{r_{13}+r_{14}}{n_1}\hat{p}-\frac{r_{14}}{n_1}\hat{p} \\
&= \hat{p}_{01}+\frac{n_1-r_{11}}{n_1}\hat{p},
\end{aligned}$$

thus

$$\frac{r_{11}}{n_1}\hat{p} = \hat{p}_{01} \quad \text{and} \quad \hat{p}_{11} = \frac{\hat{p}_{01}}{\hat{p}} = \frac{r_{11}}{n_1}.$$

Also by (A.1), (A.2) and (A.3), I can obtain

$$\hat{p}_{1j} = \frac{r_{1j}}{n_1}, \quad j = 2, 3, 4.$$

2) Thus I can write the profile likelihood as the following

$$L_p(\boldsymbol{p}_0, \hat{\boldsymbol{p}}_1) = \prod_{j=1}^{4} {p_{0j}}^{r_{0j}} \prod_{j=1}^{4} \hat{p}_{1j}^{r_{1j}} \propto \prod_{j=1}^{4} {p_{0j}}^{r_{0j}}.$$

By the *G*-*E* independence and rare disease assumptions, I have

$$p_{01}p_{04} = p_{01}(1 - p_{01} - p_{02} - p_{03}) = p_{02}p_{03},$$

i.e.,

$$p_{01} = p_{01}^2 + p_{01}p_{02} + p_{01}p_{03} + p_{02}p_{03} = (p_{01} + p_{02})(p_{01} + p_{03}),$$

similarly, I have

$$\begin{aligned} p_{01} &= (p_{01} + p_{02})(p_{01} + p_{03}) \\ p_{02} &= (p_{01} + p_{02})(p_{02} + p_{04}) \\ p_{03} &= (p_{01} + p_{03})(p_{03} + p_{04}). \end{aligned} \tag{A.4}$$

So writing

$$L_p(\boldsymbol{p}_0, \hat{\boldsymbol{p}}_1) \propto (p_{01} + p_{02})^{r_{01}+r_{02}}(p_{01} + p_{03})^{r_{01}+r_{03}}(p_{02} + p_{04})^{r_{02}+r_{04}}(p_{03} + p_{04})^{r_{03}+r_{04}},$$

I have

$$\begin{aligned}
\widehat{p_{01}^{IR}+p_{02}^{IR}} &= \frac{r_{01}+r_{02}}{n_0} \\
\widehat{p_{01}^{IR}+p_{03}^{IR}} &= \frac{r_{01}+r_{03}}{n_0} \\
\widehat{p_{02}^{IR}+p_{04}^{IR}} &= \frac{r_{02}+r_{04}}{n_0} \\
\widehat{p_{03}^{IR}+p_{04}^{IR}} &= \frac{r_{03}+r_{04}}{n_0}.
\end{aligned}$$

Plugging into (A.4), I have (5–7).

The estimated asymptotic variance-covariance matrix can be obtained by the inverse of the observed information matrix. The observed information matrix is constructed by taking the second derivative of the log-likelihood with respect the parameters, and evaluating them at the MLEs of the parameters, which are the solutions to the above equations.

Here I state how I use delta method along with the properties of a multinomial distribution and a binomial distribution to obtain the estimated asymptotic variance of the odds ratios.

First I consider OR_{10}, whose MLE is

$$\widehat{OR_{10}^{IR}} = \frac{r_{12}}{r_{11}} \cdot \frac{r_{01}+r_{03}}{r_{02}+r_{04}} = \frac{\hat{p}_{12}}{\hat{p}_{11}} \cdot \frac{\hat{p}_{01}+\hat{p}_{03}}{\hat{p}_{02}+\hat{p}_{04}},$$

where $\hat{p}_{dj} = r_{dj}/n_d$ ($d = 0, 1$ and $j = 1, 2, 3, 4$). Let $\phi = \hat{p}_{01} + \hat{p}_{03}$, I artificially build a distribution P_A which includes independent distributions P_m and P_b to satisfy this particular odds ratio estimate,

$$P_A = P_m P_b \propto p_{11}^{r_{11}} p_{12}^{r_{12}} (1-p_{11}-p_{12})^{r_{13}+r_{14}} \phi^{r_{01}+r_{03}} (1-\phi)^{r_{02}+r_{04}}.$$

Note that for the multinomial distribution $P_m \propto p_{11}^{r_{11}} p_{12}^{r_{12}} (1-p_{11}-p_{12})^{r_{13}+r_{14}}$,

$$\begin{pmatrix} \hat{p}_{12} \\ \hat{p}_{11} \end{pmatrix} \sim AN\Big(\begin{pmatrix} p_{12} \\ p_{11} \end{pmatrix}, \Sigma \Big), \quad \text{with} \quad \Sigma = \frac{1}{n_1} \begin{pmatrix} p_{12}(1-p_{12}) & -p_{11}p_{12} \\ -p_{11}p_{12} & p_{11}(1-p_{11}) \end{pmatrix}.$$

Let $g(x,y) = \log(x) - \log(y)$, then

$$\frac{\partial g(x,y)}{\partial x} = \frac{1}{x} \text{ and } \frac{\partial g(x,y)}{\partial y} = -\frac{1}{y},$$

thus by delta method,

$$\log\Big(\frac{\hat{p}_{12}}{\hat{p}_{11}}\Big) \sim AN\Big(\log\Big(\frac{p_{12}}{p_{11}}\Big), \begin{pmatrix} \frac{1}{p_{12}}, & -\frac{1}{p_{11}} \end{pmatrix} \Sigma \begin{pmatrix} \frac{1}{p_{12}}, & -\frac{1}{p_{11}} \end{pmatrix}^T\Big),$$

so the estimated asymptotic variance of $\log\Big(\frac{\hat{p}_{12}}{\hat{p}_{11}}\Big)$,

$$\widehat{AVAR}\Big(\log\Big(\frac{\hat{p}_{12}}{\hat{p}_{11}}\Big)\Big) = \frac{1}{n_1}\Big(\frac{1}{\hat{p}_{11}} + \frac{1}{\hat{p}_{12}}\Big) = \frac{1}{r_{11}} + \frac{1}{r_{12}}.$$

Similarly, for the binomial distribution $P_b \propto \phi^{r_{01}+r_{03}}(1-\phi)^{r_{02}+r_{04}}$,

$$\frac{\hat{\phi}}{1-\hat{\phi}} \sim AN\Big(\frac{\phi}{1-\phi}, \frac{1}{n_0}\Big(\frac{1}{\phi^2(1-\phi)^2}\Big)\Big).$$

Let $g(x) = \log(x) - \log(1-x)$, then

$$\frac{dg(x)}{dx} = \frac{1}{x} + \frac{1}{1-x} = \frac{1}{x(1-x)},$$

thus by delta method,

$$\log\Big(\frac{\hat{\phi}}{1-\hat{\phi}}\Big) \sim AN\Big(\log\Big(\frac{\phi}{1-\phi}\Big), \frac{1}{n_0}\Big(\frac{1}{\phi} + \frac{1}{1-\phi}\Big)\Big),$$

so the estimated asymptotic variance of $\log\Big(\frac{\hat{\phi}}{1-\hat{\phi}}\Big)$,

$$\widehat{AVAR}\Big(\log\Big(\frac{\hat{\phi}}{1-\hat{\phi}}\Big)\Big) = \frac{1}{n_0}\Big(\frac{1}{\hat{\phi}} + \frac{1}{1-\hat{\phi}}\Big) = \frac{1}{r_{01}+r_{03}} + \frac{1}{r_{02}+r_{04}}.$$

Since

$$\log(\widehat{OR}_{10}^{IR}) = \log(\hat{p}_{12}) - \log(\hat{p}_{11}) + \log(\hat{\phi}) - \log(1-\hat{\phi}),$$

then

$$\begin{aligned}\widehat{AVAR}\Big(\log(\widehat{OR}_{10}^{IR})\Big) &= \widehat{AVAR}\Big(\log\Big(\frac{\hat{p}_{12}}{\hat{p}_{11}}\Big)\Big) + \widehat{AVAR}\Big(\log\Big(\frac{\hat{\phi}}{1-\hat{\phi}}\Big)\Big) \\ &= \frac{1}{r_{11}} + \frac{1}{r_{12}} + \frac{1}{r_{01}+r_{03}} + \frac{1}{r_{02}+r_{04}}.\end{aligned}$$

Calculating the estimated asymptotic variances of $\widehat{OR}_{01}^{IR}$ and $\hat{\psi}^{IR}$ follows the same ideas.

C.2 Obtain Restriction (5–6)

Following (5–5), I have

$$(1-q_G)(1-q_E) = (1-\pi)p_{01} + \pi p_{01}/p \tag{A.5}$$

$$(1-q_G)q_E = (1-\pi)p_{02} + OR_{10}\pi p_{02}/p \tag{A.6}$$

$$q_G(1-q_E) = (1-\pi)p_{03} + OR_{01}\pi p_{03}/p \tag{A.7}$$

$$q_G q_E = (1-\pi)p_{04} + \pi OR_{10}OR_{01}\psi p_{04}/p. \tag{A.8}$$

Note that by (A.6) and (A.8)

$$q_E = (1-\pi)p_{02} + OR_{10}\pi p_{02}/p + (1-\pi)p_{04} + \pi OR_{10}OR_{01}\psi p_{04}/p \tag{A.9}$$

and by (A.7) and (A.8)

$$q_G = (1-\pi)p_{03} + OR_{01}\pi p_{03}/p + (1-\pi)p_{04} + \pi OR_{10}OR_{01}\psi p_{04}/p, \tag{A.10}$$

thus, I have

$$\begin{aligned}&(1-\pi)p_{04} + \pi OR_{10}OR_{01}\psi p_{04}/p \\ =\ &\{(1-\pi)p_{02} + OR_{10}\pi p_{02}/p + (1-\pi)p_{04} + \pi OR_{10}OR_{01}\psi p_{04}/p\} \\ \times\ &\{(1-\pi)p_{03} + OR_{01}\pi p_{03}/p + (1-\pi)p_{04} + \pi OR_{10}OR_{01}\psi p_{04}/p\},\end{aligned}$$

which is equivalent to (5–6).

C.3 Proof of Remark 3

In fact, by the results in Table 5–2, I can easily see Remark 3 holds. However, here I use a slight different way to see the role of the independence and rare disease assumptions in the proof. Ie only consider in the absence of misclassification, for the case of misclassification, the proof is straightforward.

To obtain MLEs of the parameters, I differentiate the logarithm of the likelihood (5–3) with respect to the corresponding parameters and I have the equations as in C.1. Thus I have

$$\begin{aligned} OR_{10} &= \frac{r_{12}p}{n_1 p_{02}} \\ OR_{01} &= \frac{r_{13}p}{n_1 p_{03}} \\ \psi &= \frac{r_{14}n_1 p_{02}p_{03}}{r_{12}r_{13}p_{04}p}, \end{aligned} \tag{A.11}$$

Also note that under G-E independence and rare disease assumptions, I have an additional restriction as in (5–4). So together with the definition of p_{1j} $(j = 1, 2, 3, 4)$ in (5–2), and $p_{11}n_1 = r_{11}$, I have

$$\hat{\psi}^{IR} = \frac{r_{14}n_1\hat{p}_{02}\hat{p}_{03}}{r_{12}r_{13}\hat{p}_{04}\hat{p}} \overset{(5\text{–}4)}{=} \frac{r_{14}n_1\hat{p}_{01}}{r_{12}r_{13}\hat{p}} \overset{(5\text{–}2)}{=} \frac{r_{14}n_1\hat{p}_{11}}{r_{12}r_{13}} \overset{\hat{p}_{11}n_1 = r_{11}}{=} \frac{r_{14}r_{11}}{r_{12}r_{13}} = \hat{\psi}_{CO}.$$

REFERENCES

Agresti, A. (2001). *Categorical Data Analysis*, Second Edition. New York: Wiley.

Albert, P. S., Ratnastingle, D., Tangrea, J., and Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interaction. *American Journal of Epidemiology*, **154**: 687–693.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, **59**: 19–35.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to non-parametric problems. *The Annals of Statistics*, **2**: 1152–1174.

Antoniou, A. C., Pharoah, P. P, Smith, P., and Easton, D. F. (2004). The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *British Journal of Cancer*, **91**: 1580 -1590.

Ashby, D., Hutton, J. L., and McGee, M. A. (1993). Simple Baysian analyses for case-controlled studies in cancer epidemiology. *The Statistician*, **42**: 385–389.

Baker, S. G. (1994). The multinomial Poisson transformation. *The Statistician*, **43**: 495–504.

Bashir, S. A., and Duffy, S.W. (1997) The correction of risk estimates for measurement error. *Annals of Epidemiology*, **7**: 156 164.

Botto, L.D., and Khoury M.J. (2001). Commentary: Facing the challenge of gene-environment interaction: the two by four table and beyond. *American Journal of Epidemiology*, **153**: 1016–1020.

Brennan, P. (2002). Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it? *Carcinogenesis*, **23**: 381–87.

Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, **91**: 14–28.

Breslow, N. E., and Powers, W. (1978). Are there two logistic regressions for retrospective studies? *Biometrics*, **34**: 100–105.

Breslow, N. E., Robind, D., Tangrea, J. M., and Wellner, J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli*, **6**: 447–455.

Carroll, R. J., Wang, S., and Wang, C. Y. (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association*, **90**: 157–169

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

Chatterjee, N. Kalaylioglu, Z., and Carroll, R. (2005). Exploiting gene-environment independence in family-based case-control studies: Increase power detecting associations, interactions and joint effects. *Genetic Epidemiology*, **28**: 138–156.

Clayton, D. and McKeigue, P. M. (2001). Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet*, **358**: 1356–1360.

Committee on DNA Forensic Science: An Update (1996) The evaluation of forensic DNA evidence. National Academy Press, Washington DC.

Cornfield, J. (1951). A method of estimating comparative rates form clinical data: applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, **11**: 1269–1275.

Cornfield, J. (1956). A statistical problem arising form retrospective studies *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, Berkeley, CA: University of California Press, pp 135–146.

Cornfield, J., Gordon, T., and Smith, W. W. (1961). Quantal Response Curves for exponentially uncontrolled variables. *Bulletin of the International Statistical Institute*, **38**: 97–115.

Couch, F. J., DeShano, M. L., Blackwood, M. A., Calzone, K., Stopfer, J., Campeau, L., Ganguly, A., Rebbeck, T., Weber, B. L., Jablon, L., Cobleigh, M. A., Hoskins, K., and Garber, J.E. (1997). BRCA1 mutations in women attending clinics that evaluate the risk of breast cancer. *The New England Journal of Medicine*, **336**: 1409–1415.

Cox, D. R. (1984). Interaction (with discussion) *International Statistical Review*, **52**: 1–32.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal*

of the Royal Statistical Society, Series B, **41**: 1–31.

Day, N. E., and Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics*, **23**: 313–323.

Diggle, P., Heagert, P., Liang, K., and Zeger, S. L. (1994). *Analysis of Longitudinal data*. New York: Oxford University Press.

Diggle, P. J., Morris, S. E., and Wakefield, J. C. (2000). Point-source modeling using matched case-control data. *Biostatistics*, **1**: 89–105.

Escobar, M.D, and West, M.(1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, **90**: 577–588.

Falush, D., Stephens, M., and Pritchard, J. K. (2003) Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics*, **164**: 1567–1587.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, **52**: 399–433.

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N., Pato, M. T., Petryshen, T. L., Kolonel, L. N., Lander, E. S., Sklar, P., Henderson, B., Hirschhorn, J. N., and Altshuler, D. (2004). Assessing the impact of population stratification on genetic association studies. *Nature Genetics*, **36**: 388–393.

Garcia-Closas, M., Rothman, N., and Lubin, J. (1999). Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. *Cancer Epidemiology, Biomarkers & Prevention*, **8**: 1043–1050.

Garcia-Closas, M., Thompson, W. D., and Robins, J. M. (1998). Differential misclassification and the assessment of gene-environment interactions in case-control studies. *American Journal of Epidemiology*, **147**, 426-433.

Gatto, N. M., Campbell, U. B., Rundle, A. G., and Ahsan, H. (2004). Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. *International Journal Epidemiology*, **33**(5): 1014-1024.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**: 457–472.

Ghosh, M., and Chen, M-H. (2002). Bayesian inference for matched case-control studies. *Sankhya*, Series B, Pt. 2, **64**: 107–127.

Ghosh, M., Ghosh, A., Chen, M-H, and Agresti, A. (2000). Noninformative priors for one-parameter item response models. *Journal of Statistical Planning and Inference*, **88**: 99–115.

Greenland S. (1989) On correcting for misclassification in twin studies and other matched pair studies. *Statistics in Medicine*, **8**: 825-829.

Greenland, S., and Kleinbaum, D. G. (1983) Correcting for misclassification in two-way tables and matched-pair studies. *International Journal of Epidemiology*, **12**: 93-97.

Gustafson, P., Le, N. D., and Vallee, M. (2002). A Bayesian approach to case-control studies with errors in covariables. *Biostatitics*, **3**: 229–243.

Hayakawa, T., Nagai, Y., Kahara, T., Yamashita, H., Takamura, T., Abe, T., Nomura, G., and Kobayashi, K. (2000) Gln27Glu and Arg16Gly polymorphisms of the beta2-adrenergic receptor gene are not associated with obesity in Japanese men. *Metabolism*, **49**: 1215–8.

Hoggart, C. J., Parra, E. J., Shriver, M. D., Bonilla, C., Kittles, R. A., Clayton, D. G., and McKeigue, P.M. (2003) Control of confounding of genetic associations in stratified populations. *American Journal of Human Genetics*, **72**: 1492-1504.

Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G., and McKeigue, P. M. (2004) Design and analysis of admixture mapping studies. *American Journal of Human Genetics*, **74**: 965-978.

Johnson, J. A., and Terra, S.G. (2002) b-Adrenergic receptor polymorphisms: cardiovascular disease associations and pharmacogenetics. *Pharmaceutical Research*, **19**: 1779-1787.

Khoury, M. J., and Flanders, W. D. (1996). Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls. *American Journal of Epidemiology*, **144**: 207–213.

Knowler, W.C., Williams, R.C., Pettitt, D.J., and Steinberg, A. G. (1988). Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *American Journal of Human Genetics*, **43**: 520-526.

Little, R. J. A., and Rubin, D. B. (2002). Statistical Analysis with Missing Data, Second Edition. New York: Wiley.

Lander, E. S., and Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, **265**: 2037–2048.

Lin, M., Aquilante, C., Johnson, J. A., and Wu, R. (2005) Sequencing drug response with HapMap. *The Pharmacogenomics Journal*, **5**: 149-156.

Lindley, D. V. (1964). The Bayesian analysis of contingency tables. *The Annals of Mathematical Statistics* **35**: 1622–1643.

Lubin, J. H., and Gail, M. (1990). On power and sample size for studying features of the relative odds of disease. *American Journal of Epidemiology*, **131**, 552-566.

Lynch, M., and Walsh, B. (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.

MacEachern, S. N., and Muller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**: 223–238.

Madigan, D., and Raftery, A. E. (1994). Model selection and model uncertainty in graphical models using Occam's Window. *Journal of the American Statistiscal Association*, **89**: 1535–1546

Mantel, N. (1973). Synthetic retrospective studies and realted topics. *Biometrics*, **29**: 479–486.

Mantel, N., and Haenszel, W.(1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**: 719–748.

Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P., (2004). The effects of human population structure on large genetic association studies. *Nature Genetics*, **36**: 512–517.

Marshall, R. J. (1988). Baysian analysis of case-control studies. *Statistics in Medicine*, **7**: 1223–1230.

McKeigue, P. M. (2005). Prospects for admixture mapping of complex traits. *American Journal of Human Genetics*, **76**: 1–7.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**: 153–157.

Modan, M. D., and Hartge, P., Hirsh-Yechezkel, G., Chetrit, A., Lubin, F., Beller, U., Ben-Baruch, G., Fishman, A., Menczer, J., Struewing, J. P., Tucker, M. A., Ebbers, S. M., Friedman, E., Piura, B., Wacholder, S.(2001). Parity, oral contraceptives and the risk of ovarian cancer among carriers and noncarriers of a

BRCA1 or BRCA2 mutation. *New England Journal of Medicine*, **345**: 235–240.

Morton, N. E., and Collins, A. (1998). Tests and estimates of allelic association in complex inheritance. *Proceedings of the National Academy of Sciences*, **95**: 11389–393.

Müller, P., Parmigiani, G., Schildkraut, J., and Tardella, L. (1999). A Baysian hierarchical approach for combining case-control and prospective studies. *Biometrics*, **55**: 858–866.

Müller, P., and Roeder, K. (1997). A Baysian semiparametric model for case-control studies with errors in variables. *Biometrika*, **84**: 523–537.

Nurminen, M., and Mutanen, P. (1987). Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics*, **14**: 67–77.

Parmigiani, G., Berry, D. A., and Aguilar, O. (1998). Determining Carrier Probabilities for Breast CancerSusceptibility Genes BRCA1 and BRCA2. *American Journal Human Genetics*, **62**: 145-158.

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., Hauser, S. L., Smith, M. W., OBrien, S. J., Altshuler, D., Daly, M. J., and Reich, D. (2004). Methods for High-Density Admixture Mapping of Disease Genes. *American Journal of Human Genetics*, **74**: 979–1000.

Pfeiffer, R., Gail, M, H., and Pee, D. (2001). Inference for covariates that accounts for ascertainment and random genetic effects in family studies. *Biometrika*, **88**: 933–948.

Phillips, A., and Holland, P. W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics*, **43**: 425–431.

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

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

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000a). Inference of population structure using multiocus genotype data. *Genetics*, **155**: 945–959.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000b). Association mapping in structured populations. *American Journal of Human Genetics*, **67**: 170–181.

Risch, H. A., McLaughlin J. R., Cole, D. E. C., Rosen, B., Bradley, L., Kwan, E., Jack, E., Vesprini, D. J., Kuperstein, G., Abrahamson, J. L. A., Fan, I., Wong, B., and Narod, S. A. (2001). Prevalence and Penetrance of Germline BRCA1 and BRCA2 Mutations in a Population Series of 649 Women with Ovarian Cancer. *American Journal of Human Genetics*, **68**: 700–710.

Rice, K. (2003). Full-likelihood approaches to misclassification of a binary exposure in matched case-control studies. *Statistics in Medicine*, **22**: 3177–3194.

Rice, K.M. (2004). Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications. *Journal of the American Statistical Association*, **99**: 510–522.

Rice, K., and Holmans, P. (2003). Allowing for genotyping error in analysis of unmatched case-control studies. *Annals of Human Genetics*, **67**: 165–174.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex diseases. *Science*, **273**: 1516–517.

Robins, J., Breslow, N., and Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, **42**: 311–323.

Roeder, K., Carroll, R. J., and Lindsay, B. G. (1996). Semiparametic mixture approach to case-control studies with errors in covariates. *Journal of the American Statistical Association*, **91**: 722–732.

Sala, A., Penacino, G., Carnese, R., and Corach, D. (1999). Reference database of hypervariable genetic markers of Argentina: application for molecular anthropology and forensic casework. *Electrophoresis*, **20**: 1733–739.

Sala, A., Penacino, G., and Corach, D. (1998). Comparison of allele frequencies of eight Loci from Argentinean Amerindian and European populations. *Human Biology*, **70**: 937–947.

Satten, G. A., Flanders, W. D., and Yang, Q. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *American Journal of Human Genetics*, **68**: 466–477.

Satten, G., and Kupper, L. (1993). Inference about exposure-disease associations using probability-of-exposure information. *Journal of the American Statistical Association*, **88**: 200–208.

Schmidt, S., and Schaid, D. J. (1999). Potential misinterpretation of the case-only study to assess gene-environment interaction. *American Journal of Epidemiology*,

150: 878–885.

Seaman, S. R., and Richardson, S.(2001). Bayesian analysis of case-control studies with categorical covariates. *Biometrika*, **88**: 1073–1088.

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

Seigel, D. G., and Greenhouse, S. W. (1973). Multiple relative risk functions in case-control studies. *American Journal of Epidemiology*, **97**: 324–331.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**: 639–650.

Sinha, S. Mukherjee, B., and Ghosh, M. (2004). Bayesian analysis of matched case-control studies with multiple disease states. *Biometrics*, **60**: 41–49.

Sinha, S. Mukherjee, B., Ghosh, M., Mallick, B.,K, and Carroll, R. J (2005a). Bayesian semiparametric modeling for matched case-control studies with multiple disease states. *Journal of the American Statistical Association*, **100**: 591–601.

Sinha, S. Mukherjee, B., and Ghosh, M.(2005b). Modeling association among multivariate exposures in matched case-control studies. *Preprint*.

Spielman, R.S., and Ewens, W.J. (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *American Journal of Human Genetics*, **62**: 450–458.

Spielman, R.S., McGinnis, R.E., and Ewens, W.J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, **52**: 506–516.

Srivastava, M. (1984). Estimation of interclass correlations in familial data. *Biometrika*, **71**: 177–185.

Srivastava, M., Keen, K. J., and Katapa, R. S. (1988). Estimation of interclass and intraclass correlations in multivariate familial data. *Biometrics*, **44**: 141–150.

Sullivan, P. F., Eaves, L. J., Kendler, K. S., and Neale, M. C. (2001). Genetic case-control association studies in neuropsychiatry. *Archives of General Psychiatry*, **58**: 1015–024.

Takami, S., Wong, Z. Y. H., Stebbing, M., and Harrap, S. B. (1999) Linkage

analysis of glucocorticoid and b2-adrenergic receptor genes with blood pressure and body mass index *American Journal of Physiology, Heart and Circulatory Physiology*, **276**: 1379–1384.

Umbach, D. M., and Weinberg, C. R. (1997). Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine*, **16**: 1731–43.

West, M., Müller, P., and Escobar, M. D. (1994), Hierarchical priors and mixture models, with application in regression and density estimation, in *Aspects of Uncertainty. A Tribute to D. V. Lindley*, A.F.M. Smith and P. Freeman, (eds.), pp 363–386, Wiley: New York.

Witte, J. S., Gauderman, J., and Thomas, D. C. (1999). Asymptotic bias and efficiency in case-control studies of candidates genes and gene-environment interactions: basic family design. *American Journal of Epidemiology*, **149**: 693–705.

Zelen, M., and Parker, R. A. (1986). Case-control studies and Bayesian inference. *Statistics in Medicine*, **5**: 261–269.

Zhao, H. (2000). Family-based association studies. *Statistical Methods in Medical Research*, **9**: 563–587.

BIOGRAPHICAL SKETCH

Li Zhang was born on June 19, 1976, in Jiangsu, China. She received her B.E. in civil engineering from Southeast University, Nanjing, China, in 1999. She joined the graduate program in statistics at the University of Florida in January 2001. Her research interests include case-control studies, genetic epidemiology and Bayesian method. She started to work on her dissertation in the spring of 2004.