Bayesian Modeling of Epidemiologic Data under Complex Sampling Schemes

by

Jaeil Ahn

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2011

Doctoral Committee:

Associate Professor	Bhramar Mukherjee, Co-chair
Research Associate Professor	Timothy D. Johnson, Co-chair
Professor	Stephen B. Gruber
Professor	Roderick J.A. Little

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my two advisors, Professor Bhramar Mukherjee and Professor Timothy D. Johnson, for their help and advice during my doctoral work. I want to thank my committee member Professor Stephen B. Gruber for his valuable insight into the application of the proposed methodology and generously sharing the data resource from the Molecular Epidemiology of Colorectal Cancer study. I extend my deep appreciation to Professor Roderick J.A. Little for serving on my committee and for his invaluable advices. I thank Professor Joseph N.S. Eisenberg and Professor Kathleen A. Cooney for sharing data from the Environmental change and Diarrheal disease in Ecuador study and the Flint Men's Health study respectively.

Next, I would like to thank my colleagues in the school of public health. I thank Jian Kang for sharing his expertise in C programming and in complex Bayesian Modeling. I thank Darlene Bhavnani for her efforts to keep updating the dataset for my last project and her helpful comments.

Finally, I would like to thank my family: my wife, Namhee Choi, my son, Benjamin Ahn, my parents, Hongmoon Ahn and Yekyung Park, and my parents-in-law, Seonok Choi and Misun Lee. Their help was and is invaluable to me. This thesis would not have been possible without their unconditional love and support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

CHAPTER I

Introduction

1.1 Background and Significance

Case-control studies are dominant analytic tools in epidemiologic research. In the frequentist domain, the classical theory for estimating relative risk parameters under the case-control sampling design has been well established (see Breslow, 1996 for a comprehensive review). However, with advancements in medical sciences, genomic technologies, global positioning system, and in our ability to collect and process complex data, new issues and challenges are emerging in such studies. The current thesis addresses three such problems in the general area of case-control studies using the Bayesian framework as our inferential strategy. The ultimate goal is to develop appropriate statistical methods to exploit and acknowledge the nuances of a complex data mechanism, leading to better understanding of risk factors. The set of risk factors, we consider, consists of both genetic and environmental predictors.

A case-control study is a retrospective study design in which one takes a random sample of subjects with the disease (cases) and a random sample of subjects without the disease (controls). Conditional on the disease status, exposure information on potential risk factors including demographic, behavioral, familial or possibly genetic information is then measured, depending upon the study objectives. The main goal of a case-control study is to identify potential risk factors for the disease phenotype defining the cases by comparing the distributions of exposures in cases and controls. Thus, a retrospective design is alternatively employed to gain insight on prospective quantities of interest, such as the disease odds ratio or the relative risk of a disease.

When the disease is rare, Cornfield (1951) showed that the exposure odds ratio approximates the relative risk of disease. In their seminal paper, Mantel and Haenszel (1959), further clarified the relationship between a retrospective case-control study and a prospective cohort study. Cornfield *et al.* (1961) demonstrated that the prospective probability of disease (D) given the exposure (X) or $\mathrm{P}(D|X)$ can be equivalently modeled via a logistic regression model assuming that the exposures in cases and controls, that is $\mathrm{P}(X|D)$, follow Normal distribution with different means and a common covariance matrix. Cox (1966) and Day and Kerridge (1967) demonstrated that the maximum likelihood estimator obtained from the prospective logistic likelihood under this normal exposure model is efficient, that is, the variance of estimator reach the variance lower bound. A more general class of results regarding equivalence of prospective and retrospective formulations for any exposure distribution, beyond the Normal, were derived in Anderson (1972) and Prentice and Pyke (1979). These results established that, for case-control data, though the intercept parameter is not identifiable in a prospective logistic likelihood, the corresponding estimating equations for the relative risk parameters are unbiased and the asymptotic standard errors are equivalent to those obtained from the retrospective likelihood.

Matching of case and control subjects in terms of certain demographic factors has been widely used as a design strategy to reduce potential bias caused by confounding. Under matched case control sampling, Breslow *et al.* (1978) developed a stratified logistic regression framework with a matched-set (or stratum) specific intercept parameter in a logistic regression model. Since the number of nuisance parameters grows with sample size, the naive unconditional maximum likelihood estimate (MLE) is biased and inconsistent under this model. Breslow *et al.* (1978) proposed the use of conditional logistic regression (CLR) by conditioning on the complete sufficient statistics for the nuisance parameters, namely the number of cases in each matched set, to eliminate the stratum-specific nuisance parameters.

Missingness in exposure variables entails loss of efficiency in case-control studies, especially in matched case-control studies where a single missing case-observation could lead to deletion of the entire stratum in naive complete-case CLR analysis. Several solutions have been proposed to address this issue, including a class of pseudo-likelihood based methods (Lipsitz, Parzen, and Ewell, 1998; Paik 2004), semiparametric maximum likelihood estimation (Rathouz, Satten, and Carroll, 2002), and full parametric likelihood based approaches involving modeling the distribution of the exposure variable in the control population with exponential family of distributions (Satten and Kupper, 1993; Paik and Sacco, 2000; Satten and Carroll, 2000; Sinha and Maiti 2008). The results regarding prospective and retrospective equivalence have also been extended to situations with missingness and measurement error in covariates (Roeder, Carroll, and Lindsay, 1996).

Despite the rich literature in the frequentist domain, Bayesian methods for case-control studies started to appear only in the late 1980's. Bayseian modeling is naturally appealing in epidemiologic studies as one can incorporate existing evidence in the form of elicited prior information on the model parameters. The hierarchical formulation allows uncertainty around prior guesses. Simple Bayesian analysis with a single binary exposure variable under a beta-binomial conjugate prior structure was introduced in several early papers (Zelen and Parker, 1986; Nurminen and Mutanen, 1987; Marshall, 1988; Ashby *et al.*, 1993). Zelen and Parker (1986) focused on inference regarding the log odds-ratio parameter while Nurminen and Mutanen (1987) considered posterior inference related to the risk ratio and the risk difference parameters. Marshall (1988) provided an approximate closed form expression for the posterior distribution of the odds ratio via power expansion of the hypergeometric function involved in the density. Ashby *et al.* (1993) stressed the practical utility and advantage of a Bayesian approach by eliciting informative prior information from historic data to analyze a follow-up case-control study.

With the advent of Markov chain Monte Carlo numerical integration techniques (Geman and Geman, 1994; Gelfand and Smith, 1994; Tierney, 1994) and efficient posterior sampling schemes, more complex models have been noted for case-control studies. Müller and Roeder (1997) presented a semiparametric Bayesian approach, based on a retrospective likelihood with measurement errors in the covariates. The exposure distribution had a flexible Dirichlet process mixture of normals prior. Seaman and Richardson (2001) considered a similar problem for categorical exposures and adopted a discrete Dirichlet distribution prior on the exposure distribution. Müller *et al.* (1999) proposed a hierarchical Bayesian approach to combine a case-control and cohort study to estimate the absolute risk of a disease. All of the above references consider Bayesian modeling of unmatched case-control data.

Bayesian analysis of matched case-control studies was first explored by Diggle *et al.* (2000) and Ghosh and Chen (2002). While the former used conditional likelihood to eliminate the stratum specific nuisance parameters and proceeded with Bayesian inference, the latter used the unconditional full likelihood as the basis for posterior inference. Rice (2003) considered the situation where a binary exposure is potentially misclassified in a matched case-control setting and proposed a full-likelihood based Bayesian approach. Rice (2004, 2008) characterized the class of priors leading to equivalent inference under a marginal likelihood and conditional likelihood. Prescott and Garthwaite (2005) presented Bayesian methods for analyzing matched case-control studies in which a binary exposure variable is sometimes measured with error, but whose correct values have been validated for a random sample of the matched case-control sets. Rice (2006) provided a nice comparison of the full likelihood based method of Rice (2003) to the two-stage approach presented by Prescott and

Garthwaite (2005). Liu *et al.* (2009) proposed a Bayesian adjustment for the misclassification of a binary exposure variable in a matched case-control study. The method admits a priori knowledge about both the misclassification parameters and the exposure-disease association. Sinha *et al.* (2004, 2005a,b, 2007) provided a unified Bayesian framework for analysis of matched case-control studies where the exposure distribution could depend on varying stratum effects. Mukherjee *et al.* (2007) considered matched case-control studies with multiple disease states using different links for ordered categorical outcomes.

Recently, case-control studies have been extensively used to study the association between disease, genes and their interplay with environmental risk factors. Zhang *et al.* (2006) presented a Bayesian approach to adjust for population stratification in a genetic association study. Zhang *et al.* (2008) considered flexible modeling of the random effects distribution in family based case-control studies. Mukherjee *et al.* (2007) presented the first Bayesian work on modeling case-control studies of gene-environment interaction assuming gene-environment independence. Mukherjee and Chatterjee (2008) proposed an empirical Bayes-type shrinkage estimator to analyze case control data to relax the gene-environment independence assumption in a data-adaptive manner. Mukherjee *et al.* (2008) provided an in-depth comparison of several concurrent approaches for testing gene-environment interaction. Mukherjee *et al.* (2009) investigated a Bayesian sample size determination problem for both estimation and the hypothesis testing regarding the gene-environment interaction parameter.

In this existing backdrop on Bayesian modeling for case-control data, this dissertation addresses largely three challenging problems with the connecting theme of modeling atypical data situations observed under variants of the case-control sampling scheme. Each problem is motivated by ongoing disease-exposure association studies led by researchers at the University of Michigan. More specific literature reviews and introductions to the specific application corresponding to each of the three problems appear in the respective chapters.

This dissertation is organized as follows. In Chapter II, we consider the problem of modeling disease subtypes in matched case-control studies. With advances in modern medicine and clinical diagnostic techniques, precise characterization of the histological and morphological features of a tumor are frequently available. Modeling risk profiles corresponding to tumor subtypes is an important area of research in cancer epidemiology. We use the *stereotype regression model* for categorical outcome data to achieve this modeling objective. Anderson (1984) introduced the stereotype regression model in his discussion paper on modeling ordered categorical responses. The stereotype regression model was somewhat unexplored in subsequent years. Classical frequentist inference for this class of models is complicated due to non-linearity in the parameter space as well as lack of identifiability under the global null hypotheses of null covariate effect on outcome. We introduce a Bayesian inferential paradigm that bypasses these problems under classical inference. Specifically, for modeling ordered subtypes within the cases in a matched case-control study, the stereotype model has the advantage of being amenable to the conditional likelihood principle and preserving consistent inference under prospective-retrospective conversion. We apply our method to ongoing case-control studies, prostate cancer study (The Flint Men's Health Study, FMHS). In subsequent Chapter III, we extend this basic Bayesian idea and consider matched case-control studies with partially missing exposure under the stereotype model, with the possibility of non-ignorable missingness. We propose two estimation strategies, namely, a full Bayes approach and an expectation conditional maximization (ECM) approach based on a joint retrospective likelihood. Extensive simulation studies and sensitivity analyses are carried out to assess the performances of our methods when compared with other alternatives. The methods are illustrated by applying them to colorectal cancer study (The Molecular Epidemiology of Colorectal Cancer, MECC Study).

In the second problem in Chapter IV, we develop methods for identifying gene-gene and gene-environment interactions under a two-phase design with non-monotone missing data patterns in the multiple genetic covariates. Two-phase sampling is often more efficient than standard case-control sampling where baseline information on certain demographic, environmental, behavioral, dietary exposures are available on the entire study population and a stratified sample based on disease status and some exposure is done to prioritize individuals for more expensive covariates such as genotyping in Phase II. We propose the first Bayesian approach in the context of two-phase studies. The proposed method is based on the retrospective likelihood which allows us to relax the gene-gene and gene-environment independence assumptions under multiple genes in a data-adaptive manner, thus striking a balance between bias and efficiency. Furthermore, the method offers an automated Bayesian variable selection feature to facilitate model selection in the presence of high dimensional genes, environmental covariates and their interactions in the disease risk model. We handles non-monotone missingness in the genotype data in Phase-II and posit a flexible model for the joint distribution of the Phase-I categorical variables using the non-parametric Bayes construction of Dunson and Xing (2009). Motivated by the MECC study, we investigate the interaction between the use of statins and 294 genetic markers in the lipid metabolism/cholesterol synthesis pathway. The sample on which these genetic markers were measured was enriched in terms of statin users in cases and controls. The natural missing data likelihood in two-phase studies, flexibility of leveraging the constraints around gene-environment and gene-gene independence, boosted with an adaptive variable selection algorithm, all make the Bayesian method more appealing.

In the third problem and last problem in Chapter V, we build a spatio-temporal stochastic point process model to understand the effect of spatially and temporally referenced social and natural environments; e.g. temperature and the remoteness on the diarrhea prevalence in a serial case-control study conducted in Esmeraldas province (the ECODESS study, 'http://www.sph.umich.edu/scr/ecodess/home.php'). The study largely investigates how sensitive incidence of diarrhea is to seasonal variables and social networks. Under a very complex spatial and longitudinal case pattern, we devise a new approach to analyze these case patterns as well as predict the disease prevalence over unsampled communities via building a multi-step procedures based on spatial point process models simultaneously. The study collected data in 21 communities in Esmeraldas province, while the region of interest has approximately 150 communities. During seven cycle, all diarrheal case coordinates during the study visit were recorded. To explain the clustering, we use the log Gaussian Cox process (LGCP) model as a primary tool to model spatial clustering and point patterns within sampled villages and extend it to accommodate a temporal drift. Prediction of case patterns at unsampled communities is performed by tools for the spatial misalignment problem (SMP) in the spatial regression literature that are used for interpolation with data collected on different scale and different levels of areal aggregation. We also attempted prediction of the future disease prevalence. A Bayesian paradigm is applied to model the effect of spatial covariates and residual spatial heterogeneity in a flexible and computationally tenable manner. The Metropolis-adjusted Langevin algorithm and hybrid Metropolis Hasting algorithm that facilitate posterior sampling are introduced in this context.

The current thesis explores three new problems under the case-control sampling design and thus leads to many important questions for future research. Chapter VI presents a discussion of the remaining work to be completed as a part of my dissertation and possible extensions of each chapter beyond this dissertation.

CHAPTER II

Bayesian Inference for the Stereotype Regression Model: Application to a Case-control Study of Prostate Cancer

2.1 Introduction

In his seminal discussion paper on regression and ordered categorical variables (Anderson, 1984), Anderson proposed a very general model for truly discrete outcomes called the *stereotype* regression model. The model can accommodate both ordered and unordered categorical outcomes and allows for inference regarding the order restrictions. Greenland (1994) pointed out that unlike models for ordinal data based on the cumulative logits, the stereotype model has the advantage of yielding valid inference under outcome dependent sampling, for example, in case-control studies. In addition to Greenland's observation, in the current Chapter we note that for matched case control data with fine degree of stratification and disease sub-classification, the stereotype model allows for the stratum-specific nuisance parameters to be eliminated by using the conditional likelihood principle (Breslow and Day, 1980). The proportional odds model does not yield any reduction due to sufficiency in such instances and the nuisance parameters remain even in the conditional likelihood (Mukherjee *et al.*, 2007). However, the stereotype model has been less used, primarily due to some of the computational complexities arising due to non-linearity in the parameters as well as non-standard testing theory for testing the global null hypothesis of no covariate effect (Anderson, 1984). Kuss (2006) presents a comprehensive overview of maximum likelihood estimation strategies for the stereotype regression model.

There is a growing literature on parametric and non-parametric Bayesian inference for ordinal data (Johnson and Albert, 1999; Congdon, 2005; Kottas *et al.*, 2005; Sinha *et al.*, 2004) but the stereotype model has not received attention in the Bayesian domain so far. A Bayesian approach for inference and model comparison appear to be a natural route for this class of models, and is a new contribution to the literature. Our goal in this Chapter is (i) To review the classical background and motivation for proposing this class of models for ordinal data and compare/contrast with other commonly used ordinal models; (ii) Point out limitations and issues with maximum likelihood based inference and illustrate the advantages that an alternative Bayesian route can offer; (iii) Describe Bayesian inference for this class of models, with special emphasis on case-control data with finer disease subclassification within the cases; This includes specification of priors, derivation of full conditional distribution of the model parameters and designing the computational algorithm for estimating posterior quantities of interest; (iv) Illustrate the methods using a case-control study of prostate cancer among African-American men, conducted in Flint, Michigan. We consider inference under both unmatched and matched case-control design. Sections 2.2-2.5 are organized sequentially according to the above four objectives. We also present a small-scale simulation study to compare the proposed Bayesian method with maximum likelihood based estimation in Section 2.6. Section 2.7 presents a brief concluding discussion.

2.2 Classical Background and Motivation for the Stereotype Regression Model

Two of the popular models for categorical outcomes are the polytomous logit model (or the baseline-category logit model) and the proportional odds model (Agresti, 2002; McCullagh and Nelder, 1983). Let $Y = 0, 1, \cdots K$ be an outcome with $K + 1$ categories and let the reference category be denoted by 0. Let $\boldsymbol{X}$ be a $p \times 1$ vector of p covariates. Then the proportional odds model is given by, for $k = 0, \cdots, K-1$:

$$\text{logit}[P(Y \le k|\boldsymbol{X})] = \beta_{0k} - \boldsymbol{\beta}^\top \boldsymbol{X}, \tag{2.1}$$

$\{\beta_{0k}\}$ are increasing in k and $P(Y \le K|\boldsymbol{X}) \equiv 1$. The number of parameters to be estimated in this model is $K + p$. The negative sign is used to make positive (negative) values of the elements of $\boldsymbol{\beta}$ correspond to positive (negative) association with the outcome.

On the other hand, the baseline-category logit or polytomous logistic regression model is given by, for $k = 0, 1, \cdots, K$,

$$P(Y = k|\boldsymbol{X}) = \frac{\exp(\beta_{0k} + \boldsymbol{\beta}_k^\top \boldsymbol{X})}{\sum_{k=0}^K \exp(\beta_{0k} + \boldsymbol{\beta}_k^\top \boldsymbol{X})} \tag{2.2}$$

with constraints $\beta_{00} \equiv 0$ and $\boldsymbol{\beta}_0 \equiv \mathbf{0}$. The number of parameters to be estimated in this model is $K + (p \times K)$.

The stereotype model, proposed by Anderson (1984) imposes a special structure on the parameters of the polytomous logistic model, namely,

$$\boldsymbol{\beta}_k = \phi_k \boldsymbol{\beta}. \tag{2.3}$$

Thus the stereotype model in its complete form is represented by,

$$P(Y = k|\boldsymbol{X}) = \frac{\exp(\beta_{0k} + \phi_k \boldsymbol{\beta}^\top \boldsymbol{X})}{\sum_{k=0}^K \exp(\beta_{0k} + \phi_k \boldsymbol{\beta}^\top \boldsymbol{X})} \tag{2.4}$$

for $k = 0, 1, \ldots, K$, with $\beta_{00} = \phi_0 \equiv 0$. An additional constraint has to be imposed on the other $\{\phi_k\}$ for identifiability, typically, $\phi_K \equiv 1$. The $\{\phi_k\}$ are regarded as scores for different response categories. Because of this structure, the number of parameters to be estimated in this model are: $K + (K-1) + p = 2K - 1 + p$, in between the dimensionality of models (2.2) and (2.1). Note that, akin to the proportional odds model, the stereotype model has the property of representing the effect of the predictors by a single parameter (for given scores) and is more parsimonious than the polytomous logistic model in (2.2).

In (2.4), variable X_r has coefficients $\phi_k\beta_r$ (r indexing the variable and k indexing the category) which represents the log odds-ratio corresponding to category k vs category 0 of Y. The constraint $\phi_K \equiv 1$ implies that β_r, corresponding to X_r represents the effect of unit change in X_r on the log OR of response in the highest category K vs category 0 of Y. In order to understand the interpretation of the parameters in the model and visualize the structure, let us consider the special case with $K = 2$, implying that the possible values of Y are $0, 1$, and 2, and a set of two covariates X_1 and X_2. Then the stereotype model and the polytomous logit model (in squared brackets) can be compared as,

$$\begin{aligned}\log[\frac{\pi_1(\boldsymbol{X})}{\pi_0(\boldsymbol{X})}] &= \beta_{01} + \phi_1(\beta_1 X_1 + \beta_2 X_2) = \Big[\beta_{01} + \beta_{11}X_1 + \beta_{12}X_2\Big] \\ \log[\frac{\pi_2(\boldsymbol{X})}{\pi_0(\boldsymbol{X})}] &= \beta_{02} + \phi_2(\beta_1 X_1 + \beta_2 X_2) = \Big[\beta_{02} + \beta_{21}X_1 + \beta_{22}X_2\Big]\end{aligned}$$

Just like the proportional odds assumption, the stereotype structure may not be realistic for a given dataset, and needs to be tested in terms of model diagnostics and fit statistics. However, it does allow some flexibility compared to the proportional odds model by introducing the category-specific score parameters which are assumed to be the same for each covariate. Allowing covariate-specific scores will lead to a model exactly equivalent to the multinomial logistic model (2.2).

The stereotype model can also be represented in terms of adjacent category logits

$$\log[\frac{\pi_k(\boldsymbol{X})}{\pi_{k+1}(\boldsymbol{X})}] \quad = \quad \beta_{0k} - \nu_k \boldsymbol{\beta}^\top \boldsymbol{X},$$

where $\nu_k = \phi_{k+1} - \phi_k$ in terms of the score parameters $\{\phi_k\}$s. With fixed choice of the scores, say equispaced, i.e.,$\phi_k = k/K$, the model reduces to simpler representations corresponding to the more standard adjacent category logit model as given in Agresti (2002, p. 287).

Anderson's original motivation was to propose a model for inherently discrete outcomes, different from models derived from grouped-continuous data, like the proportional odds model. The stereotype model captures a situation where the assignment into different categories are done by processing some indeterminate amount of information in assessor's mind who has "stereotypes" in which a new case is categorized. Anderson (1984) and Greenland (1994) both argue that collapsing invariance and invariance under reversal of outcome codes should not be treated as axioms of acceptability for ordered regression models. In some contexts the axiom is less attractive, particularly with *assessed* categorical outcomes.

Anderson (1984) suggested using the stereotype model for ordered data with monotone scores.

$$0 \equiv \phi_0 < \phi_1 < \ldots < \phi_K \equiv 1$$

This implies that for a unit increase in covariate x_r, the log(OR) $\phi_k \beta_r$ of category k vs baseline category 0 becomes larger when category k gets further than 0. For this ordered stereotype model the conditional distribution of $Y|\boldsymbol{X}$ is stochastically ordered in $\boldsymbol{\beta}^\top \boldsymbol{X}$ (Anderson, 1984). Note that these orderings are quite different than the ones given in models for grouped continuous data, for example, in the proportional odds model. There, the ordered categories are "given" and not necessarily ordered with respect to $\boldsymbol{X}$, and, the ordering exists even in the absence of any covariates $\boldsymbol{X}$. In the stereotype model, the ordering is intrinsically defined through the regression relationships between Y and $\boldsymbol{X}$. For example, if we had a single binary covariate X with values 0 and 1, and $p_{kl} = \mathrm{pr}(Y = k, X = l)$, $k = 0, \cdots, K$, $l = 0, 1$, denote the cell probabilities in a $2 \times (K + 1)$ contingency table, then the observations follow an *ordered stereotype* model if and only if the the probability ratios p_{k0}/p_{k1} are monotonically increasing or decreasing. In the spirit of Cornfield's (1951) construction of the logistic regression model in terms of the retrospective distribution of $\boldsymbol{X}|Y = k$, suppose that $\boldsymbol{X}|Y = k \sim N(\boldsymbol{\mu}_k, \Sigma)$, $k = 0, \cdots, K$; then the stereotype model implies that $(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)'\Sigma^{-1} = \phi_k \boldsymbol{\beta}$, or that the means follow a linear trend.

Before concluding this section we will like to point out two major advantages of the stereotype regression model when applied to case-control data with finer disease sub-classification within the cases, as in our real data example. First, unlike the cumulative logit models, the stereotype model is preserved under outcome dependent sampling. If $f_0, \cdots, f_K$ are the sampling fractions corresponding to each category then, under outcome-stratified sampling,

$$\frac{P(Y = k|\boldsymbol{X}, \text{sampled})}{P(Y = 0|\boldsymbol{X}, \text{sampled})} = \exp(\beta_{0k}^{*} + \phi_k \boldsymbol{\beta}^{\top} \boldsymbol{X})$$

where $\beta_{0k}^{*} = \beta_{0k} + \log(f_k/f_0)$. Thus under a correctly specified stereotype model the estimates from retrospectively sampled data will be consistent for true covariate effects (Greenland, 1994).

Second, in a matched case-control study with binary outcomes, the matched set specific nuisance parameters are eliminated by using conditional logistic regression (Breslow and Day, 1980). Mukherjee *et al.* (2007, 2008) illustrate that the conditioning principle does not apply to the cumulative logit model and the stratum specific parameters remain in the conditional likelihood. The number of such nuisance parameters grows with sample size and thus maximum likelihood estimation runs into inconsistency problems. Due to its multiplicative intercept structure, the stereotype model is amenable to the conditioning principle. With this backdrop in mind we proceed to discuss inferential issues associated with this model.

2.3 Inference in stereotype regression model: issues with maximum likelihood and advantages of a Bayesian paradigm

In spite of its parsimony when compared to (2.2), with unknown $\{\phi_k\}$, the non-linearity of the parameters in (2.4) makes estimation by a standard generalized linear model software infeasible. Greenland (1994) suggests a two-step estimation procedure which starts by fixing $\{\phi_k\}$, and then estimating $\boldsymbol{\beta}$ by fitting a baseline-category logit model with multiple predictors $\phi_k x_r$'s. At the second step, treat $\hat{\boldsymbol{\beta}}^{\top} \boldsymbol{X}$ as a single predictor and estimate $\{\phi_k\}$ by the same model fitting procedure. This iterative procedure does not guarantee convergence to the true MLE. The correct standard errors are recommended to be obtained by a subsequent bootstrap procedure (Greenland, 1994; Lall el al., 2002). Kuss (2006) discusses various computational methods to obtain the MLEs in the stereotype regression model. Among the algorithms proposed he discusses generalized least squares (GLS) and a quasi-Newton method for direct maximization of the likelihood. PROC NLMIXED, PROC MODEL in SAS and R function *gnm* can be modified to do the direct maximization. The model can be embedded in the class of reduced rank vector of generalized linear models (RR-VGLM) and can be fitted in R-package VGLM (Yee and Hastie, 2003). The ordered version of the stereotype model is often harder to fit due to multiple constraints and hitting boundaries of the parameter space and often not implementable in standard software.

A Bayesian procedure can potentially handle many of these difficulties associated with frequentist estimation. The ordering constraints are ensured by appropriate prior choices on the simplex $0 \equiv \phi_0 < \phi_1 < \ldots < \phi_K = 1$, so that the constraints are automatically satisfied while generating posterior samples. Bayesian computation is based on exact simulation of the posterior distribution of the parameters, thus calculating the posterior variance or highest posterior density (HPD) credible interval does not pose any additional challenges or require validity of large sample approximations. Moreover, if one is actually interested in estimating the log odds ratio parameters $\phi_k\beta_r$, corresponding to category k and covariate X_r, the Bayesian procedure can directly generate the posterior distribution of this quantity, whereas frequentist inference will require using the multivariate Delta theorem to derive a variance approximation for this product parameter, and rely on large samples in each response category for the validity of this approximation.

The multiplicative nature of the stereotype model poses some issues for testing the null hypothesis of independence $H_0 : \boldsymbol{\beta} = 0$ in the likelihood based framework. Under this global null hypothesis, the score parameters $\{\phi_k\}$ are not identifiable. McCullagh (1984) pointed out in the discussion of Anderson (1984) that the approximate null distribution of the likelihood-ratio statistic is that of the largest eigenvalue of a Wishart matrix (Habermean, 1981). The testing problem under such non-regular conditions remains to be explored in the frequentist domain for this class of models. Theories developed for modified partial likelihood ratio test (Hanfelt and Liang, 1995; Chen *et al.*, 2001) could be useful in deriving a suitable test under this class of models. We relegate the all discussion related to hypotheses testing to Appendix A.1.1 and focus on estimation and model comparison in the main text of the Chapter.

Model comparison between the different class of nested models as well as between ordered and unordered model can be carried out by using the Deviance Information Criterion (DIC)(Carlin and Louis, 2000; Gelman *et al.*, 2004). In a Bayesian numerical integration scheme with direct generation of posterior observations, the expected deviance function $E_{\boldsymbol{\theta}|y}(D(\boldsymbol{\theta}))$ can be estimated by the average deviances over the posterior realizations $\{\boldsymbol{\theta}_l : l = 1, \cdots, L\}$, expressed as,

$$\overline{D} = \frac{1}{L}\sum_{l=1}^{L} D(\boldsymbol{\theta}_l) = \frac{1}{L}\sum_{n=1}^{L}\{-2\log p(y|\boldsymbol{\theta}_l)\}.$$

The DIC is defined as the expected deviance, penalized by the effective sample size $DIC = \overline{D} + (\overline{D} - D(\boldsymbol{\theta}^*)) = 2\overline{D} - D(\boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ is the posterior mean. The advantage of DIC over the other model selection criteria is that it can be easily computed in a Markov chain Monte Carlo (MCMC) setting, in terms of the readily available posterior quantities.

We now proceed to describe our Bayesian approach with specification of prior, likelihood, posterior and outline the numerical scheme to generate observations from the posterior distribution of model parameters.

2.4 Bayesian Inference for the Stereotype Regression Model

We specifically formulate our notations to represent data from a case-control study though the Bayesian proposal is very general and can readily be applied to a prospective stereotype regression model. We first describe the methods for a matched case-control design and then proceed to discuss the modifications under an unmatched study design. Let Y_{ij} denote the disease state corresponding to the j-th subject in the i-th stratum (or matched set), with S_i denoting all variables which contributed explicitly or implicitly to the formation of the i-th stratum. The disease states are classified into one of the K distinct categories $1, 2, \cdots, K$, while the reference control group is denoted by $Y_{ij} = 0$. In each of the N strata we assume there is one case matched with M controls. The results could be directly generalized to the setting of more general $L_i : M_i$ matching ratio. We consider a vector of covariates $\boldsymbol{X}_{ij}$ corresponding to subject j in stratum i. The disease risk model is described as,

$$(2.5) \quad P(Y_{ij} = k|\boldsymbol{X}_{ij}, S_i) = \frac{\exp(\beta_{0k}(S_i) + \phi_k(\boldsymbol{\beta}^\top X_{ij}))}{\sum_{k=0}^{K} \exp(\beta_{0k}(S_i) + \phi_k(\boldsymbol{\beta}^\top X_{ij}))} \qquad \text{for } k = 0, \cdots, K$$

The $\beta_{0k}(S_i)$ are category specific intercepts which could vary with strata. For identifiability $\beta_{00}(S_i) \equiv 0$. Assuming $\beta_{0k}(S_i) \equiv \beta_{0k}$ renders an unmatched analysis where the category specific intercepts are assumed to be constant across matched sets. Under the ordered model we assume the constraint $0 \equiv \phi_0 < \phi_1 < \ldots < \phi_K \equiv 1$. Without loss of generality, let us assume that the first subject in each stratum is the case and remaining are controls. To eliminate the stratum specific nuisance parameters $\beta_{0k}(S_i)$ we use the conditional likelihood, by conditioning on the event $\sum_{j=1}^{M+1} Y_{ij} = k_i$, in the i-th stratum, where k_i is the observed disease state corresponding to the case subject in the i-th stratum, i.e. $Y_{i1} = k_i$. The corresponding conditional likelihood is given by

$$
\begin{aligned}
L_c &= \prod_{i=1}^{N} P\left(Y_{i1} = k_i, Y_{i2} = \cdots = Y_{iM+1} = 0 | \{\boldsymbol{X}_{ij}\}_{j=1}^{M+1}, S_i, \sum_{j=1}^{M+1} Y_{ij} = k_i\right) \\
&= \prod_{i=1}^{N} \left[\frac{\exp(\phi_{k_i}(\boldsymbol{\beta}^\top \boldsymbol{X}_{i1}))}{\sum_{j=1}^{M+1} \exp(\phi_{k_i}(\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}))}\right] \qquad (2.6)
\end{aligned}
$$

One could proceed with Bayesian inference by either assuming a hierarchical prior on the parameters of the prospective model (2.5) (including $\beta_{0k}(S_i)$), or via the conditional likelihood in (2.6) treating it as a genuine likelihood and impose prior structure on the parameters ϕ_k and $\boldsymbol{\beta}$. The justification for using L_c as a basis of Bayesian inference can be found in Rice (2003).

Priors: There are two sets of parameters under consideration here, the covariate effects $\boldsymbol{\beta}$ and the category-specific scores $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_{K-1})$ in the conditional likelihood (2.6). On the parameters of $\boldsymbol{\beta}$, we impose either independent normal or joint multivariate normal priors. In an *unordered* stereotype model we can similarly put independent or joint multivariate normal prior on $\phi_1, \cdots, \phi_{K-1}$, without an order constraint on the prior support.

For the *ordered* stereotype model, to handle the identifiability condition and the order restriction of the parameters $0 \equiv \phi_0 < \phi_1 < \ldots < \phi_{K-1} < \phi_K = 1$ in a natural way, we reparameterize the parameters in terms of the differences,

$$
\gamma_1 = \phi_1; \qquad \gamma_s = \phi_s - \phi_{s-1} \qquad \text{for } s = 2, \cdots, K-1
$$

Thus $\phi_k = \sum_{s=1}^{k} \gamma_s$ for $s = 2, \cdots, K$, and $\phi_0 = \gamma_0 \equiv 0$. The condition $\phi_K \equiv 1$ implies that $\gamma_K = 1 - \phi_{K-1} = 1 - \sum_{s=0}^{K-1} \gamma_s$. The stereotype model requires $\{\phi_k\}$ to be increasing in k and be bounded by 1, this implies that the sequence of parameters $\{\gamma_s\}$ are positive and lie in the $K-1$ dimensional simplex $0 < \sum_{s=1}^{K-1} \gamma_s < 1$. There are several strategies to ensure this, here, we focus on the prior structure involving a Dirichlet distribution as follows:

- If $0 \equiv \phi_0 < \phi_1 < \ldots < \phi_K = 1$ are $K-1$ order statistics from the uniform distribution

U(0,1), then the successive differences of the order statistics, namely, $\{\gamma_s\}$ as defined above, $s = 1, \cdots, K$, follow a Dirichlet $(1, \cdots, 1)$ distribution. This particular prior actually corresponds to centering your belief around equal spacing of the $\{\phi_k\}$. Generalizing this result, one can directly impose a Dirichlet$(\alpha_1, \cdots, \alpha_K)$ distribution jointly on the parameters $\{\gamma_s\}$, which is a natural prior with support as the above simplex.

Full Conditional Distributions: We first consider the *unordered* stereotype model with the following mutually independent normal priors on $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$

$$\begin{aligned} \boldsymbol{\beta} &\sim N(\boldsymbol{\mu}_\beta, \sigma^2_\beta I_p) \\ \boldsymbol{\phi} &\sim N(\boldsymbol{\mu}_\phi, \sigma^2_\phi I_{K-1}) \end{aligned}$$

Combining the conditional likelihood (2.6) with the assumed priors, we can derive the following joint posterior distribution,

$$p(\boldsymbol{\beta}, \boldsymbol{\phi} | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}) \propto \prod_{i=1}^{N} \left[\frac{\exp(\phi_{k_i}(\boldsymbol{\beta}^\top \boldsymbol{X}_{i1}))}{\sum_{j=1}^{M+1} \exp(\phi_{k_i}(\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}))} \right] \times \pi(\boldsymbol{\beta})\pi(\boldsymbol{\phi}) \tag{2.7}$$

The full conditional distribution of the respective parameters are given by:

$$\begin{aligned} \pi(\beta_r|\cdot) &\propto \frac{\exp\left(-\frac{1}{2\sigma^2_{\beta_r}}(\beta_r - \mu_{\beta_r} - \sigma^2_{\beta_r} \sum_{i=1}^{N} \phi_{k_i} X_{i1r})^2\right)}{\prod_{i=1}^{N} \sum_{j=1}^{M+1} \exp(\phi_{k_i}(\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}))} \\ \pi(\phi_k|\cdot) &\propto \frac{\exp\left(-\frac{1}{2\sigma^2_{\phi_k}}\left(\phi_k - \mu_{\phi_k} - \sigma^2_{\phi_k} \sum_{i=1}^{N} I(Y_{i1} = k)(\boldsymbol{\beta}^\top \boldsymbol{X}_{ij})\right)^2\right)}{\prod_{i=1}^{N} \sum_{j=1}^{M+1} \exp \phi_k(\boldsymbol{\beta}^\top \boldsymbol{X}_{ij})} \end{aligned}$$

Where X_{ijr} is the value of the r-th predictor corresponding to the covariate vector $\boldsymbol{X}_{ij}$, for the j-th subject, in the i-th stratum, and β_r is the parameter specific to covariate X_r; $i = 1, \cdots N, j = 1, \cdots, M+1, r = 1, \cdots, p$. For the ordered stereotype model, we consider the natural prior choice that $\{\gamma_s\}$, $s = 1, \cdots, K$ follows a Dirichlet $(\alpha_1, \cdots, \alpha_K)$ distribution. The full conditional for $\{\gamma_s\}$ is expressed as

$$\pi(\gamma_s|\cdot) \propto \prod_{i=1}^{N} \left[\frac{\exp(\gamma_s I(Y_{i1} \geq s)(\boldsymbol{\beta}^\top \boldsymbol{X}_{i1}))}{\sum_{j=1}^{M+1} \exp(\sum_{k=1}^{k_i} \gamma_k(\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}))} \right] \times {\gamma_s}^{\alpha_s - 1}.$$

For the special case of Dirichlet$(1,\cdots,1)$ prior, corresponding to a prior belief of equal spacing of the scores, the joint full conditional distribution of $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_{(K-1)})$ is expressed as following.

$$\pi(\boldsymbol{\phi}|\cdot) \propto \prod_{i=1}^{N} \left[\frac{\exp(\phi_{k_i}(\boldsymbol{\beta}^\top \boldsymbol{X}_{i1}))}{\sum_{j=1}^{M+1} \exp(\phi_{k_i}(\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}))} \right]$$

Remark: For **unmatched** case-control data with N subjects consisting of n_1 cases and n_0 controls, we have an intercept parameter that is constant across strata, namely, $\beta_{0k}(S_i) \equiv \beta_{0k}$ and the conditional likelihood need not be evoked, one is able to proceed with the unconditional prospective likelihood (2.5) with an additional set of priors on $\boldsymbol{\beta}_0 = (\beta_{01}, \cdots, \beta_{0K})$. The data vector now has only one index with $(Y_i, \boldsymbol{X}_i)$ corresponding to the observations on the i-th subject, $i = 1, \cdots, N$. For the *unordered* stereotype model, we assume that,

$$\begin{aligned} \boldsymbol{\beta}_0 &\sim N(\boldsymbol{\mu}_{\beta_0}, \sigma^2_{\beta_0} I_K) \\ \boldsymbol{\beta} &\sim N(\boldsymbol{\mu}_{\beta}, \sigma^2_{\beta} I_p) \\ \boldsymbol{\phi} &\sim N(\boldsymbol{\mu}_{\phi}, \sigma^2_{\phi} I_{K-1}) \end{aligned}$$

Combining the prospective likelihood with the assumed priors we have the posterior as,

$$p(\boldsymbol{\beta}, \boldsymbol{\phi}|\boldsymbol{Y}, \boldsymbol{X}) \propto \prod_{i=1}^{N} \left[\frac{\exp(I(Y_i = k)(\beta_{0k} + \phi_k \boldsymbol{\beta}^\top \boldsymbol{X}_i))}{\sum_{k=0}^{K} \exp(\beta_{0k} + \phi_k \boldsymbol{\beta}^\top \boldsymbol{X}_i)} \right] \times \pi(\boldsymbol{\beta})\pi(\boldsymbol{\phi}) \tag{2.8}$$

The full conditional distribution of the respective parameters are as follows:

$$\begin{aligned} \pi(\beta_{0k}|\cdot) &\propto \frac{\exp\left(-\frac{1}{2\sigma^2_{\beta_{0k}}}(\beta_{0k} - \mu_{\beta_{0k}} - \sigma^2_{\beta_{0k}} \sum_{i=1}^{N} I(Y_i = k))^2\right)}{\prod_{i=1}^{N} \sum_{k=0}^{K} \exp(\beta_{0k} + \phi_k \boldsymbol{\beta}^\top \boldsymbol{X}_i)} \\ \pi(\beta_r|\cdot) &\propto \frac{\exp\left(-\frac{1}{2\sigma^2_{\beta_r}}(\beta_r - \mu_{\beta_r} - \sigma^2_{\beta_r} \sum_{i=1}^{N} I(Y_i = k)\phi_k X_{ir})^2\right)}{\prod_{i=1}^{N} \sum_{k=0}^{K} \exp(\beta_{0k} + \phi_k \boldsymbol{\beta}^\top \boldsymbol{X}_i)} \\ \pi(\phi_k|\cdot) &\propto \frac{\exp\left(-\frac{1}{2\sigma^2_{\phi_k}}\left(\phi_k - \mu_{\phi_k} - \sigma^2_{\phi_k} \sum_{i=1}^{N} I(Y_i = k)(\boldsymbol{\beta}^\top \boldsymbol{X}_i)\right)^2\right)}{\prod_{i=1}^{N} \sum_{k=0}^{K} \exp(\beta_{0k} + \phi_k \boldsymbol{\beta}^\top \boldsymbol{X}_i)} \end{aligned}$$

Here X_{ir} is the r-th covariate corresponding to subject i, $i = 1, \cdots, N$, $r = 1, \cdots, p$. For the *ordered* stereotype model and unmatched data likelihood, the full conditional distribution of $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_K)$ under the Dirichlet(α_1, $\cdots$,α_k) prior is expressed as follows;

$$\pi(\gamma_s|\cdot) \propto \prod_{i=1}^{N} \left[\frac{\exp(\gamma_s I(Y_i \geq s)(\boldsymbol{\beta}^\top \boldsymbol{X}_i))}{\sum_{k=0}^{K} \exp(\beta_{0k} + \phi_k \boldsymbol{\beta}^\top \boldsymbol{X}_i)} \right] \times {\gamma_s}^{\alpha_s - 1}.$$

Bayesian Computation: The Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) is used to generate a sequence of random observations from the joint posterior distribution $[\boldsymbol{\beta}, \boldsymbol{\phi}|Y, \boldsymbol{X}, S]$ and Bayesian estimates are obtained from this generated sequence. In this algorithm, observations are sampled iteratively from the following conditional distributions: $p[\boldsymbol{\beta}|\boldsymbol{X}, Y, S, \boldsymbol{\phi}]$ and $p[\boldsymbol{\phi}|\boldsymbol{X}, Y, S, \boldsymbol{\beta}]$. Since none of the full conditionals have a standard distributional form, we used a hybrid of Gibbs sampler and Metropolis Hastings algorithm to generate random numbers from the full conditionals of the parameters as specified above. Noting that some of the above full-conditionals are log-concave, we use adaptive rejection sampling (ARS) (Gilks and Wild, 1992) for sampling $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ for the unordered model and only $\boldsymbol{\beta}$ for the ordered model. We use Metropolis-Hastings update of the $\boldsymbol{\phi}$ in the ordered model with proposal distribution on ϕ as order statistics from the uniform distribution on [0,1]. We typically ran the sampler 50,000 iterations and considered estimates based on the last 10,000 runs, allowing a burn-in of 40,000 iterations. To monitor the convergence of the chain, we use the 'potential scale reduction factor' diagnostic proposed by Gelman and Rubin (1992).

2.5 Example: The Flint Men's Health Study

Prostate cancer is the most common non-cutaneous cancer among American men and is the second leading cause of cancer deaths in the United States. African American men have a 1.5 fold higher incidence of prostate cancer compared to Caucasian men (Sarma and Schottenfield, 2002). The Flint Men's Health Study (FMHS) is a community-based case-control study of prostate cancer in African-American men between the ages of 40-79, from 1999-2002 (Cooney *et al.*, 2001). Case subjects were identified using the Genessee County Community-wide hospital oncology program (CHOP) registry. Along with in-home interview questionnaires on occupational and behavioral exposures, personal and family history of cancer, the study collected hospital record information on stage, Gleason's grade of differentiation, treatment and pre-diagnosis prostate-specific antigen (PSA) value. Several anthropometric measurements and blood sample were also collected. Disease-free controls were identified from a sample of African-American men in Flint, or in selected census tracts in neighboring communities. Controls were asked to undergo a prostate cancer screening protocol which included providing blood sample for PSA measurement and a comprehensive urological examination. Men with abnormal clinical examination or elevated PSA were excluded from control set and referred for prostate biopsy. There were 28 men who were subsequently diagnosed with prostate cancer from this referral, who were included as case-subjects. We had initial data on 144 cases and 434 controls which contained some missing information.

Our categorical outcome variable is the stage of prostate cancer. Two systems are in common use for the staging of prostate cancer. The Jewett system (1975, stages A through D) was described in 1975 and has since been modified. In 1997, the American Joint Committee on Cancer (AJCC) and the International Union Against Cancer adopted a revised tumor, nodes, metastasis (TNM) system that employs the same broad stage categories as the Jewett system but includes finer subcategories of stage. This revised TNM system is clinically useful and more precisely stratifies newly diagnosed patients. In 2002, the AJCC further revised the TNM classification system (Prostate IN., 2002). A thorough review of the controversies of staging in prostate cancer is contained in (Montie, 1995). We have used the AJCC recommended TNM based classifications to define our response variable. Notice that stage is not a one-dimensional continuum or discretized version of such a latent continuum. According to Anderson's original motivating framework, assignment of stage should be treated as summaries of multidimensional outcomes. The *stereotypes* assigned for different stages are based on many aspects of a clinical and histopathological examination. As two potential predictors, we used log(1+PSA) and age. We used pre-diagnosis total free PSA for the case subjects and the total free PSA measured at the time of entry in the study for control subjects as a measure of PSA. The age recorded at the time of the PSA measurement is used as a second covariate in the model.

The final dataset we use for illustration purposes, after removing subjects with missing information on stage, age or PSA consists of 433 controls and 132 cases. Among 132 cases, 64 subjects are in stage 1, 61 subjects are in stage 2 and 7 subjects are in stage 3 according to the AJCC/TNM classification as described above. Ages are noted to have a roughly uniform distribution over the entire range [40, 79] with mean 58 yrs and standard deviation 10.4 yrs. To avoid very small values for the estimated regression coefficient corresponding to age, we transform the ages into a [0, 1] interval by using the linear transformation, i.e. (Age-40)/(79-40). The marginal empirical distribution of log-transformed PSA (in combined case and control sample) is distributed roughly as a normal distribution with mean 1.12 and sd 0.84 truncated at 0. We first analyze the unmatched data using the unconditional likelihood (2.5) with $\beta_{0k}(S_i) = \beta_{0k}$, and explore the association of age and PSA with different stages of cancer. We fit both the ordered and unordered model in the proposed Bayesian framework. We used $N(0, 5^2)$ prior distributions for $\boldsymbol{\beta}, \boldsymbol{\phi}$ in unordered stereotype model and $\boldsymbol{\beta}$ in the ordered model. For $\boldsymbol{\phi}$ in the ordered model, we use the *Dirichlet*(1,1,1,1) as a prior density for the successive score differences $\{\gamma_s\}$ as described in the previous section.

In order to obtain the MLE under the unordered stereotype regression model, we follow the maximum likelihood algorithm implemented in SAS/ETS module PROC MODEL, by fitting the stereotype model as a general multivariate non-linear regression model (Kuss, 2006). Throughout the Chapter we refrain from maximizing the likelihood under the ordered stereotype model due to computational instability. We provide the result of baseline category or multinomial logistic regression model for reference. Though not quite comparable, we present the results of a proportional odds model fitted to the data as well.

Estimates of $\boldsymbol{\beta}, \boldsymbol{\phi}$ and their standard errors with 95% HPD for Bayesian methods or 95% confidence interval for frequentist methods under different models and approaches are presented in Table 2.1. We can note the well expected significant positive association between PSA and stage of cancer. The corresponding log-odds ratios given in Appendix Table A.1 also illustrate that the odds ratios corresponding to each stage as compared to the controls appear to be ordered in terms of PSA. On the other hand, the effect of age, after adjusting for PSA is not so pronounced and the direction is sensitive to the model choice, with the age effect not being significant under any method, except attaining borderline significance in the ML fitted stereotype model. The score parameters $\hat{\boldsymbol{\phi}}$ appear to be monotonically increasing across stage, though their values are quite close. The ordered and unordered Bayesian stereotype models have very similar values of estimated DIC criterion. The unordered stereotype regression model fitted by the approach followed in Kuss (2006) has AIC estimate (though directly not comparable to DIC) close to the Bayesian models. The multinomial logistic model has an AIC that is higher than the stereotype model. The variance of the log odds-ratio corresponding to category 3, is indeed worrisome in the multinomial logistic regression model (estimate 10.81, s.e.=7.47). This is due to the fact that there are only 7 subjects in category 3 who contribute to the estimation of this parameter, whereas the stereotype model uses common covariate-specific parameter for estimating this log odds-ratio and has much superior precision. The proportional odds model, on the other hand has a higher AIC value compared to both multinomial logistic and stereotype model, indicating that this model may not be the right choice in this context.

Table 2.1: Results of analysis of unmatched FMHS data with 132 cases and 433 controls. We used transformed age and $\log(1+PSA)$ as the two covariates X_1 and X_2. Under the Bayesian methods the "estimate" corresponds to the posterior mean whereas PSD and HPD correspond to posterior standard deviation and highest posterior density intervals respectively. For the MLE, CI corresponds to the Wald-type large sample confidence intervals.

Unordered Stereotype		β_1	β_2	ϕ_1	ϕ_2	DIC/AIC
Bayes	Estimate	-0.12	4.29	0.52	0.57	557.6
	PSD	1.11	0.46	0.06	0.06	
	95% HPD	(-2.31, 1.99)	(3.39, 5.15)	(0.40, 0.64)	(0.46, 0.70)	
MLE	Estimate	-1.92	8.72	0.35	0.39	563.5
	s.e.	0.88	0.82	0.04	0.04	
	95% CI	(-3.64, -0.19)	(7.11, 10.33)	(0.27, 0.43)	(0.31, 0.47)	
Ordered Stereotype		β_1	β_2	ϕ_1	ϕ_2	DIC
Bayes	Estimate	-0.01	4.44	0.50	0.56	555.9
	PSD	0.99	0.38	0.05	0.05	
	95% HPD	(-2.28, 2.24)	(3.57, 5.24)	(0.40, 0.61)	(0.45, 0.68)	
Alternative Models		β_1	β_2			AIC
Proportional	Estimate	0.01	2.13			570.0
Odds Logistic	s.e.	0.59	0.17			
Regression	95% CI	(-1.15, 1.17)	(1.79, 2.46)			
		β_{11}	β_{12}	β_{13}		AIC
Multinomial	Estimate	1.37	-0.35	-1.08		
Logistic	s.e.	0.79	0.83	6.15		
Regression	95% CI	(-0.18, 2.93)	(-1.98,1.27)	(-13.13, 10.96)		
		β_{21}	β_{22}	β_{23}		567.2
	Estimate	2.19	2.60	10.81		
	s.e.	0.24	0.26	7.47		
	95% CI	(1.71, 2.67)	(2.08, 3.11)	(-3.83, 25.45)		

†In Multinomial Logistic regression model, β_{rk} is the log odds-ratio of category k versus the controls corresponding to covariate X_r, $r=1,2$; $k=1,2,3$.

‡AIC is the Akaike Information Criterion and DIC is the Deviance Information Criterion.

For illustration purposes we created a matched case-control dataset from this database with 132 cases matched with three controls in terms of census tract and neighborhoods of residence to obtain a 1:3 matched dataset with 396 controls. With these 528 subjects and 132 matched sets, we fitted the stereotype and multinomial logistic model via the conditional likelihood. We omit presenting the results for the proportional odds model for the matched case, as it is not amenable to the conditioning principle and can only implement an unmatched analysis of a matched study.

For the stereotype models for matched data we consider both ordered and unordered models with the same prior distributions for unmatched data analysis under the Bayesian approach. Since the approach in Kuss (2006) does not discuss conditional likelihood, we obtain the conditional maximum likelihood estimates by direct maximization of (2.6) via the Nelder-Mead optimization routine. The optimization method though relatively stable for a few parameters, encounters many convergence issues with larger number of covariates or response categories.

The results of matched data analysis are presented in Table 2.2. Though there are certain numerical differences especially, in terms of the effects of age when compared with the unmatched analysis, the basic pattern of inference remains the same. The ordered stereotype model appears to provide a slightly better fit to the data in terms of DIC criterion. The posterior standard deviations are larger for the matched dataset, providing wider credible intervals. This may be due to the reason that we created artificial matching in the dataset for illustration purposes and the matching strategy may not have been efficient and created correlation among covariates within a matched set.

Table 2.2: Results of analysis of 1:3 matched FMHS data with transformed age and $\log(1 + PSA)$ as the two covariates X_1 and X_2. There are 132 matched sets. Under the Bayesian methods the "estimate" corresponds to the posterior mean whereas PSD and HPD correspond to posterior standard deviation and highest posterior density intervals respectively. For the MLE, CI corresponds to the Wald-type large sample confidence intervals. For Bayesian models we report DIC whereas for models fitted by maximum likelihood we report AIC.

Unordered Stereotype		β_1	β_2	ϕ_1	ϕ_2	DIC/AIC
Bayes	Estimate	1.57	4.56	0.54	0.63	145.6
	PSD	1.15	1.67	0.20	0.23	
	95% HPD	(-0.72 4.67)	(1.95, 8.59)	(0.20, 1.03)	(0.26, 1.13)	
MLE	Est.	1.54	4.37	0.51	0.58	148.6
	s.e.	1.33	1.78	0.20	0.23	
	95% CI	(-1.07, 4.15)	(0.88, 7.86)	(0.12, 0.90)	(0.13, 1.03)	
Ordered Stereotype		β_1	β_2	ϕ_1	ϕ_2	DIC
Bayes	Estimate	1.57	4.33	0.50	0.66	144.3
	PSD	1.18	1.08	0.13	0.16	
	95% HPD	(-0.73, 3.96)	(2.45, 6.50)	(0.29, 0.80)	(0.40, 0.96)	
Alternative Models		β_{11}	β_{12}	β_{13}		AIC
Multinomial	Estimate	1.30	0.29	-		
Conditional	s.e.	0.84	0.99	-		
Logistic	95% CI	(-0.33, 2.94)	(-1.66, 2.24)	-		
		β_{21}	β_{22}	β_{23}		-
	Estimate	2.09	2.61	-		
	s.e.	0.41	0.45	-		
	95% CI	(1.27,2.90)	(1.73, 3.49)	-		

†The Multinomial Logistic model for matched data was fitted by simultaneously maximizing three conditional likelihoods of categories 1 vs 0, 2 vs 0 and 3 vs 0. The conditional MLEs of β_{13}, β_{23} did not converge, thus the AIC measure is not available.
‡AIC is the Akaike Information Criterion and DIC is the Deviance Information Criterion.

Figures 2.1 and 2.2 show posterior densities of $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ for the ordered and the unordered model, overlaid on each other for unmatched and matched datasets respectively. The ordered and unordered model produce fairly similar results in the unmatched case (Figure 2.1), but certain differences are noted in the matched data context (Figure 2.2). For the ordered model, mixing of the Markov chain is slow with a single Dirichlet distribution as proposal density for the ϕ. We used the mixture of two Dirichlet distributions as proposal and increased the number of iterations to 100,000 for the ordered models, however certain roughness is still noted in the density plot due to ties in the posterior samples at boundaries.

An advantage of the Bayesian approach is to be able to carry out exact posterior inference on the log odds-ratio parameters of category k compared to category 0, namely $\phi_k\beta_r$, $k = 1, 2, 3$ and $r = 1, 2$, corresponding to covariate X_r. Appendix Tables A.1 and A.2 present the corresponding results for the unmatched and matched datasets. Appendix Figures A.1 and A.2 exhibit the posterior distribution of the log-odds ratio parameters $\phi_k\beta_r$, $r = 1, 2$, $k = 1, 2, 3$; which look fairly symmetric for the unmatched data but exhibit skewness in matched data. Based on our computational experience, it appears that both likelihood maximization and Bayesian estimation with the conditional likelihood for matched data is appreciably more challenging than the unconditional likelihood with stereotype link function.

2.6 Simulation Study

Since in a real dataset we can only illustrate the methods and the truth about the parameters is unknown, we conducted a small scale simulation study to evaluate our proposed methods, mimicking the real data analysis of FMHS in certain aspects. We consider two covariates, namely X_1 and X_2 resembling age and PSA in the FMHS study. We generate X_1 from a uniform distribution on [40,79] and $\log(X_2)$ from a N(1.12, 0.84^2) distribution. We then generate a matching variable S following a U[0,1] distribution which will only be used for generating the matched sample. Given the true parameter values and $\boldsymbol{X}$, we generate the outcome variable $Y|\boldsymbol{X}, S$ with five categories ranging from 0 to 4 from a multinomial distribution with the following response probability,

$$P(Y_i = k|\boldsymbol{X}_i, S_i) = \frac{\exp(\beta_{0k} + \beta_S S_i + \phi_k(\boldsymbol{\beta}^\top \boldsymbol{X}_i))}{\sum_{k=0}^{K} \exp(\beta_{0k} + \beta_s S_i + \phi_k(\boldsymbol{\beta}^\top \boldsymbol{X}_i))} \qquad \text{for} \qquad k = 0, \cdots, 4,$$

where β_{0k} are category-specific intercepts with β_{00}=0. We set the effect size corresponding to the matching variable, namely, β_S at 0.5. We consider the covariate-specific parameters (β_1, β_2) =(1.0, 2.0). We consider two scenarios with respect to the category specific scores: $\boldsymbol{\phi} = (\phi_0, \phi_1, \phi_2, \phi_3, \phi_4) = (0, 0.6, 0.9, 0.6, 1)$ reflecting that the ordering assumption is violated and $\boldsymbol{\phi} = (0, 0.25, 0.5, 0.75, 1)$ when the ordering assumption is preserved. We call these two situations unordered and ordered setting respectively. The category specific intercepts β_{0k} are set as (0,-5.5, -7, -6, -8) for unordered setting and (0,-4.5, -5.5, -6.5, -8) for ordered setting.

With the above distributional structure, we generate 500,000 independent realizations of the data vector $(Y, \boldsymbol{X}, S)$ resulting in a cohort with 500,000 subjects with roughly 92% controls and 8% cases. The percentage of cases in stages 1, 2, 3, and 4 were 45%, 30%, 15% and 10% respectively. We first randomly select 150 cases from the case population. In order to construct a 1:3 matched dataset, corresponding to each selected case we selected three controls having the value of the matching variable S within 0.03 of what was noted for the case subject. For generating the unmatched datasets we simply remove S from the above simulation scheme and generate $Y|\boldsymbol{X}$ only, and randomly select 150 cases and 450 controls from case and control population respectively. We generated 500 such matched and unmatched datasets by sampling from the cohort. Thus we have four simulation settings (a) unmatched, unordered (b) unmatched, ordered (c) matched, unordered and (d) matched, ordered as presented in Tables 2.3 and 2.4 respectively.

We fitted both the ordered and unordered stereotype models using the Bayesian approach, with priors identical to our data analysis section. For the unordered model we consider maximum likelihood estimation as well. With unmatched, unordered data in Table 2.3 we notice that the Bayesian ordered stereotype model yields larger biases with respect to $\boldsymbol{\phi}$ though the estimation of β is quite satisfactory in terms of both bias and MSE. The unordered stereotype model as implemented by the methods illustrated in Kuss (2006) encounter convergence problems in roughly 30% of all generated datasets and even after removing the non-convergent datasets the MLE have larger bias and MSE than the corresponding Bayesian method for estimating $\boldsymbol{\beta}$. Under (a), the unordered Bayesian method performs best when estimation of $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ are considered simultaneously. In contrast, with (b) unmatched ordered data, the ordered stereotype model as implemented by the Bayesian method performs the best in terms of mean squared errors for both $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$. Note that ML approach tends to yield smaller biases in estimating $\boldsymbol{\phi}$ while producing larger biases in estimating $\boldsymbol{\beta}$, an observation that we could not explain theoretically or heuristically. Interestingly the average DIC for unordered (ordered) model is less for unordered (ordered)setting, thus the two simulation settings are reflected in the preferred model choice under the Bayesian approaches.

We repeat the simulation results with 500 *matched* datasets and results are given in Table 2.4. The basic pattern remains the same as in Table 2.3, with the Bayesian unordered model performing the best under (c) and the Bayesian ordered model performing the best under (d). Once again, the DICs are able to indicate the better model fit among the ordered and unordered model, consistent with the simulation setting of generating ordered or unordered data.

Appendix Tables A.3 and A.4 contain the corresponding results for the log odds ratio parameters and the same basic pattern is followed. Based on the simulation study and the issues encountered with convergence of MLE, it appears that fitting the ordered and unordered stereotype regression model by the Bayesian approach and assessing comparative

Table 2.3: The results of the simulation study under unmatched case-control sampling design. The results are based on 500 simulated datasets with 150 cases and 450 controls. For each parameter we report estimated bias and mean squared error based on the 500 replications. The outcome variable Y has five categories from 0 to 4. The true values for the parameters are: $\beta_1 = 1.0$, $\beta_2 = 2.0$; for unordered setting $\phi_1 = 0.6$, $\phi_2 = 0.9$, $\phi_3 = 0.6$ and for ordered setting $\phi_1 = 0.25$, $\phi_2 = 0.5$ and $\phi_3 = 0.75$.

Unordered Setting		β_1	β_2	ϕ_1	ϕ_2	ϕ_3
Unordered Stereotype	bias	-0.02	-0.06	0.04	0.03	0.06
Bayes	MSE	0.25	0.08	0.02	0.03	0.02
Unordered Stereotype	bias	0.20	0.14	0.04	-0.00	0.03
MLE	MSE	0.38	0.96	0.03	0.05	0.03
Ordered Stereotype	bias	0.05	0.08	-0.07	-0.18	0.18
Bayes	MSE	0.33	0.08	0.01	0.04	0.04
Ordered Setting		β_1	β_2	ϕ_1	ϕ_2	ϕ_3
Unordered Stereotype	bias	-0.12	0.14	0.01	0.02	0.03
Bayes	MSE	0.30	0.12	0.01	0.10	0.15
Unordered Stereotype	bias	0.12	0.13	-0.01	-0.00	-0.01
MLE	MSE	0.59	0.25	0.01	0.03	0.04
Ordered Stereotype	bias	-0.08	-0.08	0.00	0.00	0.00
Bayes	MSE	0.29	0.06	0.01	0.01	0.01

†The estimated DICs corresponding to Bayesian estimation of the unordered stereotype model and ordered stereotype model are 922.8 and 928.1 and for unordered setting, 973.5 and 968.1 for the ordered setting respectively.

model fit via Bayesian model fitting diagnostics is an attractive alternative for this class of models.

2.7 Discussion

The *stereotype* regression model is an interesting class of models for categorical outcomes which has remained somewhat unexplored in the literature due to problems with classical inference as discussed in Section 2.3. The Bayesian paradigm circumvents many of the issues with classical inference and provides an alternative approach to estimate the model parameters and carry out model comparisons and model selection for this class of models. In modern medicine, characterization of disease subclasses in terms of histological and morphological terms is often available. The *stereotype* model has a unique distinction among models for ordered data that is preserved under outcome-dependent sampling and can be applied to matched data with fine degree of stratification. The current work is the first attempt to-

Table 2.4: The results of the simulation study under matched case-control sampling design with 1:3 matching ratio. The results are based on 500 simulated datasets, each with 150 matched sets. For each parameter we report estimated bias and mean squared error based on the 500 replications. The outcome variable Y has five categories from 0 to 4. The true values for the parameters are: $\beta_1 = 1.0$, $\beta_2 = 2.0$; for unordered setting $\phi_1 = 0.6$, $\phi_2 = 0.9$, $\phi_3 = 0.6$ and for ordered setting $\phi_1 = 0.25$, $\phi_2 = 0.5$ and $\phi_3 = 0.75$.

Unordered Setting		β_1	β_2	ϕ_1	ϕ_2	ϕ_3
Unordered Stereotype	bias	0.00	0.01	0.09	0.14	0.15
Bayes	MSE	0.38	0.30	0.08	0.12	0.13
Unordered Stereotype	bias	0.15	0.30	0.01	0.01	0.05
MLE	MSE	0.77	1.02	0.06	0.10	0.11
Ordered Stereotype	bias	0.22	0.43	-0.15	-0.26	0.15
Bayes	MSE	0.46	0.34	0.03	0.07	0.03
Ordered Setting		β_1	β_2	ϕ_1	ϕ_2	ϕ_3
Unordered Stereotype	bias	0.01	-0.01	0.04	0.07	0.15
Bayes	MSE	0.39	0.27	0.02	0.05	0.15
Unordered Stereotype	bias	0.13	0.23	0.01	0.00	0.04
MLE	MSE	0.56	0.39	0.02	0.04	0.12
Ordered Stereotype	bias	0.12	0.17	-0.01	-0.02	-0.01
Bayes	MSE	0.43	0.17	0.01	0.01	0.01

†The estimated DICs for unordered stereotype Bayes model and ordered stereotype Bayes model are 281.8 and 285.3 and for unordered setting, and, 320.3 and 315.4 for the ordered setting respectively.

wards proposing Bayesian inference for this class of models with and without the ordering restrictions. We also aim to provide the reader with an classical overview of this class of models. We point out its advantages for unmatched and matched case-control designs and illustrate the methodology in a study of prostate cancer. We present a simulation study when the true outcome are generated from both ordered and unordered model and illustrate that the Bayesian model comparison approach can discern between the two situations and lead to an estimate with good mean-squared error properties.

There are many issues which needs further exploration related to the stereotype model from a frequentist analysis perspective as well, among which is the exploration of an appropriate testing strategy for the hypothesis of independence. Our study indicates that we need better computation strategies in the conditional likelihood setting. How to extend this model to accommodate missing data, correlated or clustered observations can be considered as topics of future research.

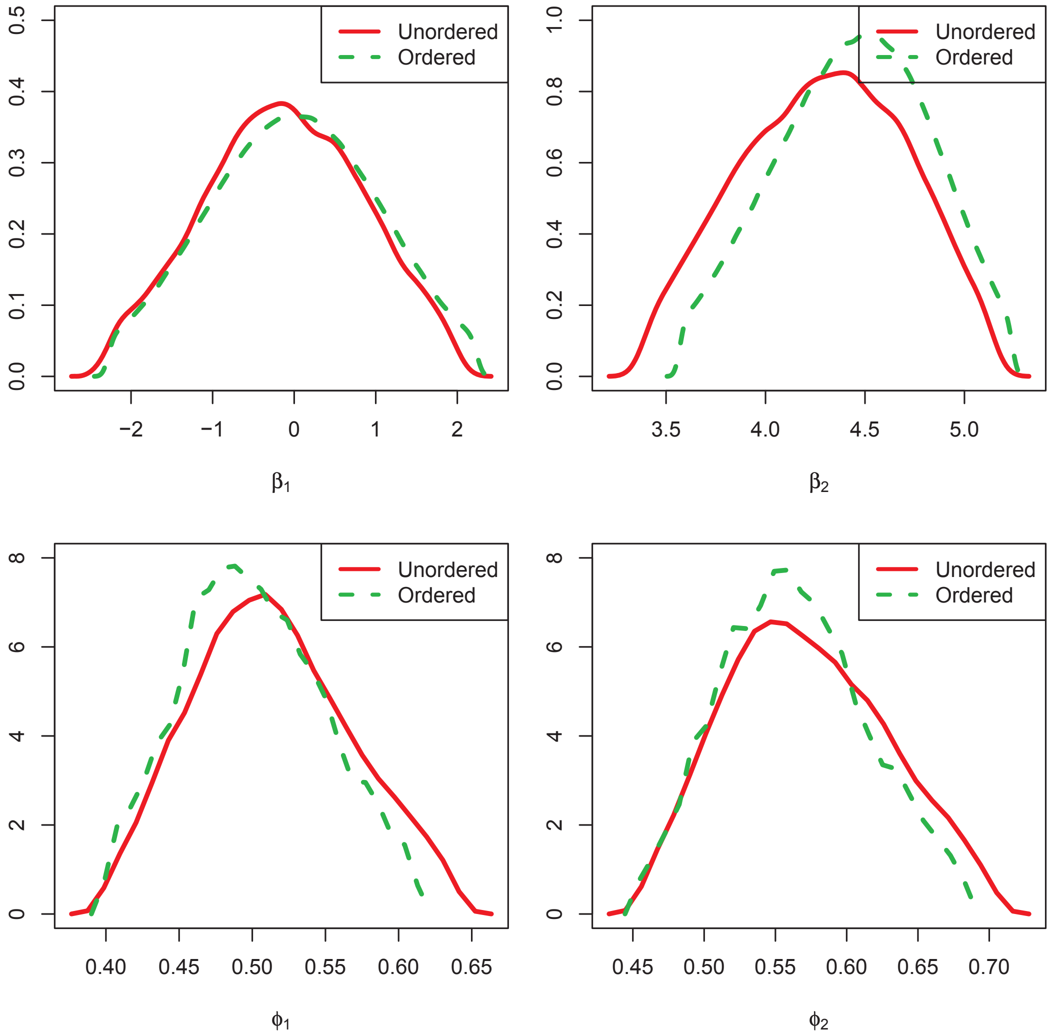


Figure 2.1: Posterior density estimates for covariate and category specific parameters of the stereotype model for unmatched FMHS data with numerical summaries as presented in Table 2.1. The results are based on 100,000 observations generated from the posterior distribution of each parameter. The solid line corresponds to the unordered model, whereas the dashed line corresponds to the ordered model.

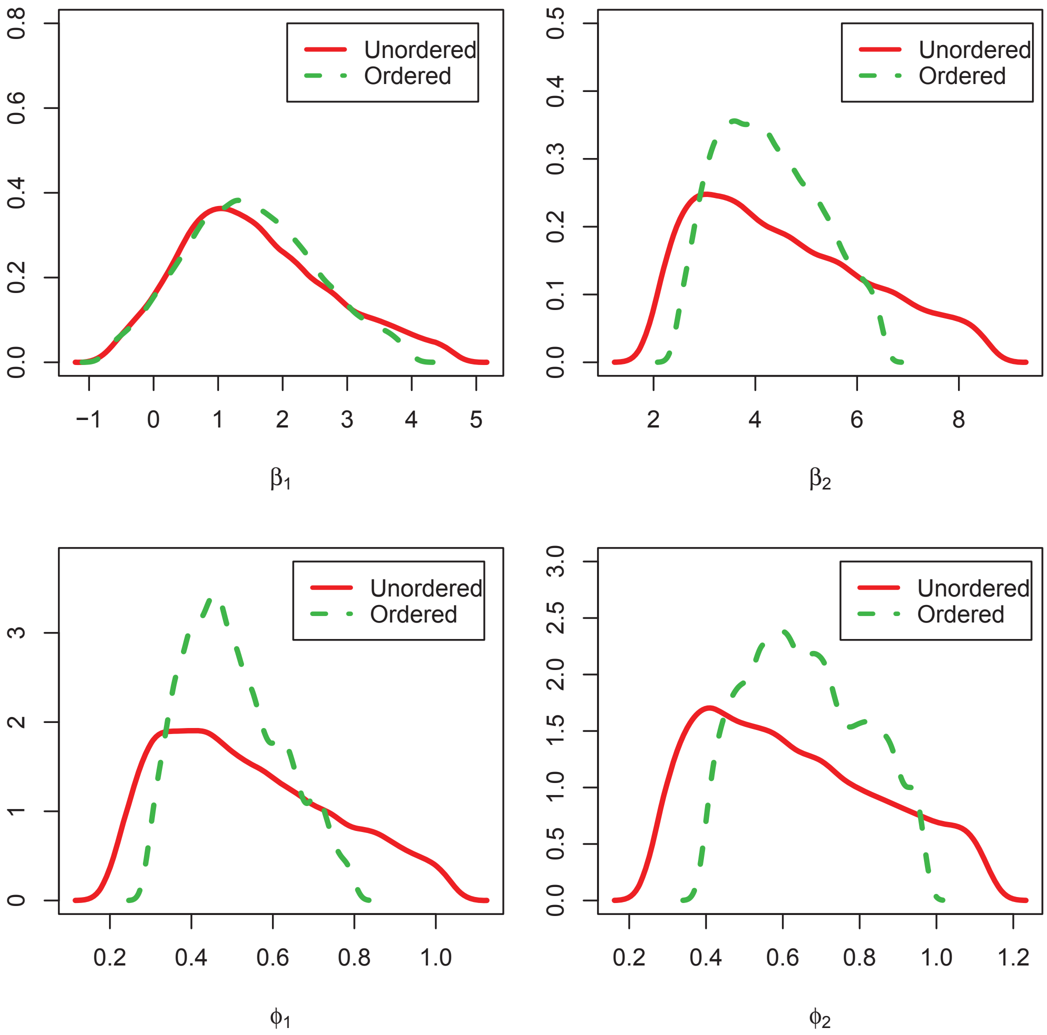


Figure 2.2: Posterior density estimates for covariate and category specific parameters of the stereotype model for 1:3 matched FMHS data with numerical summaries as presented in Table2.2. The results are based on 10,000 observations generated from the posterior distribution of each parameter. The solid line corresponds to the unordered model, whereas the dashed line corresponds to the ordered model.

CHAPTER III

Missing Exposure Data in Stereotype Regression Model: Application to Matched Case-Control Study with Diseases Subclassification

3.1 Introduction

In this Chapter we propose two methods for handling partially missing covariate data in a stereotype regression model while the data are collected through a matched case-control design. The stereotype regression model was proposed by Anderson (1984) for analyzing categorical outcome data by using category-specific scores and maintaining the homogeneous effect of covariates corresponding to each logit. The model stands intermediate between the baseline category logit model and the proportional odds model in terms of model flexibility and parsimony. The model can be adapted to ordered as well as unordered outcome setting whereas ordering assumption is required for the proportional odds model. The stereotype model, however, has been less attractive as an alternative to proportional odds model due to embedded computational burden caused by multiplicative structure of the model parameters. Since Anderson's initial paper, there has been only handful of follow-up papers on this class of models. Greenland (1994) proposed a two-step iterative algorithm followed by bootstrap for estimation of model parameters and their standard errors respectively. Holtbr$\ddot{u}$gge and

Schumacher (1991) used an iteratively reweighted least squares algorithm (Green, 1984) to obtain parameter estimates. Recently, Yee and Hastie (2003) considered the stereotype model as a special case of the reduced rank vector generalized linear model (RR-VGLM) and introduced a fitting approach in the R-package VGAM (Yee, 2010). Kuss (2006) presented an in-depth overview on the estimation of the parameters of a stereotype model by employing generalized least squares and discussed alternate implementation procedures in standard statistical software. Kuss (2004) contained an illustrative example using the random effects stereotype regression model. Lunt (2004) considered prediction of ordinal outcomes using this model. Ahn *et al.* (2009) presented Bayesian inference for ordered and unordered stereotype model.

Greenland (1994) pointed out an attractive feature of this model in terms of yielding valid inference under retrospective sampling, like in a case-control study. Alternative ordinal models such as the proportional odds or cumulative logit model do not preserve valid inference under outcome stratified sampling (Mukherjee and Liu, 2009; Mukherjee *et al.*, 2007). Moreover, for a matched case-control study, the conditional likelihood principle (Breslow and Day, 1980) may be invoked to eliminate stratum-specific nuisance parameters under this stereotype class of link functions, whereas the proportional odds model is not amenable to this principle (Mukherjee *et al.*, 2008). With advances in detection and diagnosis techniques for cancer, classification information into finer subtypes of cancers/tumors are often available in existing databases. The stereotype model presents an interesting alternative to model association of risk factors with such subtypes rather than just case-control status. The outcome categories or disease subtypes may or may not be ordered in terms of effect of covariates. The structure of the stereotype model allows a unique opportunity for testing such ordering restrictions. Thus the model appears to be an appealing tool for analyzing matched case-control data with finer disease subclassification.

Missingness in exposure values is frequently a concern in matched case-control studies. Naive use of the conditional logistic regression (CLR) renders deletion of the complete stratum containing any missing case observations in matched case-control studies. There exists a substantial amount of literature on handling missing data in matched case-control studies (Satten and Carroll, 2000; Paik and Sacco, 2000; Rathouz *et al.*, 2002; Rathouz, 2003; Sinha *et al.*, 2005). All of these papers consider missingness mechanisms that may or may not depend on observed data (missing completely at random (MCAR) and missing at random (MAR) as defined in Little and Rubin (1987)). Under MAR and MCAR, naive CLR is known to be inefficient.

If the missingness mechanism depends on unobserved exposure values, failure to incorporate the missingness information in the analysis can lead to biased and inconsistent results. Paik (2004) used a parametric approach to handle such informative missingness (IM) in matched case-control studies using a pseudo-likelihood. After the timely first investigation of Paik (2004) for handling IM in matched case-control studies, Sinha and Maiti (2008) carried out a comprehensive comparison of Paik's approach with an alternative full-likelihood based approach. Both of these papers use the expectation/maximization (EM) algorithm to estimate model parameters and to derive standard error estimates. None of the above papers, however, consider the problem of modeling disease subclassification, and do not involve the stereotype regression model. Sinha *et al.* (2004) did consider the problem of missing exposure data with multiple disease states using a polytomous regression model but not under IM. The parametric structure of the stereotype model leads to new computational issues and there is no literature on handling missingness under this class of models. In this article, we propose an expectation conditional maximization (ECM) approach and a full Bayesian (FB) approach to handle missing data under the stereotype model. The methods are applied to analyze the association between use of statins (a lipid lowering drug), physical activity and different stages of colorectal cancer (cancer staging based on Tumor, Nodes, and Metastasis criteria), in an ongoing population-based matched case-control study (Poynter *et al.*, 2005).

The rest of the article is organized as follows. In Section 3.2.1, we introduce the stereotype regression model. In Section 3.2.2, we describe the conditional likelihood under a matched case-control setting, without any missingness. In Section 3.2.3 we present the likelihood formulation with partially observed data, with a model for missing data and selection probability mechanism. In Section 3.3, we discuss the computational strategies to estimate the model parameters, namely the ECM and the full Bayes strategy. We illustrate our methods via analyzing data from the Molecular Epidemiology of Colorectal Cancer (MECC) Study in Section 3.4. Finally, we carry out a simulation study to compare properties of the different estimation strategies in terms of bias and mean squared error (MSE) under different missingness mechanisms in Section 3.5. Section 3.6 presents brief concluding remarks.

Before concluding this section, we highlight two novel features of this article. To the best of our knowledge, there is no literature on handling missing data under the stereotype link function. The current article is also the first one to present a full Bayesian framework to deal with non-ignorable missingness in matched case-control studies under any link function. We compare the performance of both the full Bayesian (FB) and the maximum likelihood based approach (ECM) in terms of simulation studies under an array of missingness mechanisms and model misspecification.

3.2 Models and Assumptions

In this section, we introduce the key ingredients of our likelihood specification, starting with the stereotype link function, the complete data likelihood, then followed by models for the selection probability and the distribution of the missing exposure.

3.2.1 The Stereotype Regression Model

The stereotype model is nested within the family of polytomous logistic regression models. The polytomous logistic regression model for a categorical response variable Y with $K+1$ categories and a p-dimensional vector of explanatory variables $\boldsymbol{X}$ is denoted by

$$p(Y=k|\boldsymbol{X}) = \frac{\exp(\beta_{0k} + \boldsymbol{\beta}_k^\top \boldsymbol{X})}{\sum_{k=0}^{K} \exp(\beta_{0k} + \boldsymbol{\beta}_k^\top \boldsymbol{X})}, \tag{3.1}$$

for $k = 0, 1, \ldots, K$ with constraints $\beta_{00} \equiv \boldsymbol{\beta}_0 \equiv 0$. The $p \times 1$ parameter vector $\boldsymbol{\beta}_k$ denotes the log odds ratio of category $Y = k$ relative to baseline category $Y = 0$. Anderson (1984) proposed the stereotype model by imposing a structure on $\boldsymbol{\beta}_k$ such that $\boldsymbol{\beta}_k = \phi_k \boldsymbol{\beta}$. The *stereotype* regression model can thus be represented as,

$$p(Y=k|X) = \frac{\exp(\beta_{0k} + \phi_k \boldsymbol{\beta}^\top \boldsymbol{X})}{\sum_{k=0}^{K} \exp(\beta_{0k} + \phi_k \boldsymbol{\beta}^\top \boldsymbol{X})}, \tag{3.2}$$

for $k = 0, 1, \ldots, K$. For identifiability of the parameters, we assume $\beta_{00} = \phi_0 \equiv 0$ and, $\phi_K \equiv 1$. The model can be extended to accommodate ordered outcomes with a monotonicity constraint on the category-specific scores, namely, $0 \equiv \phi_0 \leq \phi_1 \leq \ldots \leq \phi_K \equiv 1$. The ordering constraint can be tested in light of the data by comparing the ordered and unordered model by a likelihood ratio test. The number of parameters to be estimated in (3.2) is $(2K-1)+p$, compared to $K + (p \times K)$ parameters in the polytomous logit model (3.1). The stereotype model allows a bit more flexibility than the proportional odds model which assumes an identical effect of the covariates for each cumulative probability, further reducing the number of parameters to be estimated to $K + p$. One can actually test the indistinguishability of covariate effects on outcome categories k and l by testing $H_0 : \phi_k = \phi_l$ in (3.2) and potentially collapse categories with similar category-specific scores. However, the limitations of the model are non-linearity in the parameters due to product terms in $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ and the lack of identifiability of the parameters under the global null hypotheses of $H_0 : \boldsymbol{\beta} = 0$, leading to non-standard asymptotic theory for likelihood based inference.

3.2.2 Stereotype Regression in Matched Case-Control Studies

As Greenland (1994) pointed out, the stereotype model leads to consistent and asymptotically efficient estimates of the parameters of interest, namely, $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$, under outcome-stratified sampling. Anderson (1984) specifically recommended this model for categorical outcomes that are not generated by segmenting a latent continuous scale, but are summaries of truly discrete multidimensional outcomes. A natural example for such an outcome is stages of cancer which are typically assessed based on multiple criteria in a case-control study. For matched case-control studies with finer disease subclassification, the stereotype model provides additional flexibility in terms of eliminating the matched set specific parameters via the conditional likelihood.

We now describe the stereotype regression model for the specific setting of a matched case-control design. Let Y_{ij} denote the disease state corresponding to the j^{th} subject in the i^{th} stratum (or matched set), with S_i denoting variables which contributed explicitly or implicitly to the formation of the i-th stratum. The disease states are classified into one of the K distinct categories $1, 2, \cdots, K$, while the reference control group is denoted by $Y_{ij} = 0$. In each of the N strata we assume there is one case matched with M controls. The results could be directly generalized to the setting of more general $L_i : M_i$ matching ratio. For ease of notation, we restrict our attention to a single covariate X_{ij} with potential missingness, the results could again be extended to a set of covariates containing missingness in a straightforward way (Sinha *et al.*, 2008). Let $\boldsymbol{Z}_{ij}$ denote the vector of p completely observed covariates $\boldsymbol{Z}_{ij} = [Z_{ij1} \ldots Z_{ijp}]^T$ corresponding to the j-th subject in the i-th stratum. The stratified disease risk model for $Y = 0, \ldots, K$ is described as,

$$(3.3) \qquad p(Y_{ij} = k | X_{ij}, \boldsymbol{Z}_{ij}, S_i) = \frac{\exp\{\beta_{0k}(S_i) + \phi_k(\beta_1 X_{ij} + \boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij})\}}{\sum_{k=0}^{K} \exp\{\beta_{0k}(S_i) + \phi_k(\beta_1 X_{ij} + \boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij})\}}.$$

The $\beta_{0k}(S_i)$ are category specific intercepts which could vary with strata. For identifiability,

$\beta_{00}(S_i) = \phi_0 \equiv 0$ and $\phi_K \equiv 1$. The change in the log odds of an individual being in the k^{th} disease category versus being a control, for each unit increase in X is given by $\phi_k\beta_1$. Without loss of generality, let us assume that the first subject in each stratum is the case and remaining are controls. To eliminate the stratum specific nuisance parameters $\beta_{0k}(S_i)$, we use the conditional likelihood, by conditioning on the event $\sum_{j=1}^{M+1} Y_{ij} = k_i$, in the i-th stratum, where k_i is the observed disease state corresponding to the case subject in the i-th stratum, $k_i = 1, \cdots, K$.

Thus the conditional likelihood when we have complete data is given by,

$$\begin{aligned} L_c &= \prod_{i=1}^{N} P\left(Y_{i1} = k_i, Y_{i2} = \cdots = Y_{iM+1} = 0 | \{X_{ij}, \boldsymbol{Z}_{ij}\}_{j=1}^{M+1}, S_i, \sum_{j=1}^{M+1} Y_{ij} = k_i\right) \\ &= \prod_{i=1}^{N} \frac{\exp\{\phi_{k_i}(\beta_1 X_{i1} + \boldsymbol{\beta}_2^\top \boldsymbol{Z}_{i1})\}}{\sum_{j=1}^{M+1} \exp\{\phi_{k_i}(\beta_1 X_{ij} + \boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij})\}}. \qquad (3.4) \end{aligned}$$

For completely observed data one could proceed with Bayesian inference via the above conditional likelihood treating it as a genuine likelihood and impose prior structure on the parameters ϕ_k, β_1 and $\boldsymbol{\beta}_2$ (Ahn *et al.* 2009). The justification for using L_c as a basis of Bayesian inference can be found in Rice (2004).

3.2.3 Likelihood formulation under missingness in exposure values

Matched case-control analysis is often challenged with issues involving missingness in exposure values. The conditional likelihood approach as described in (3.4) is inefficient in such situation. Moreover, if the missing data mechanism is non-ignorable, then a naive complete-case analysis may produce biased and inconsistent estimates. Paik (2004) and Sinha and Maiti (2008) present elegant EM-based maximum likelihood estimation strategies to handle non-ignorable missingness in matched case-control study. However, none of these papers consider a full Bayesian approach, or the issue of finer disease subclassification. The stereotype link function has also not been studied therein. We now present a general framework for handling missingness in exposure values by modeling both the exposure distribution and the missingness process under the model (3.3).

Let R_{ij} denote the indicator variable assuming the value 1 if X_{ij} is observed and is 0 otherwise. The joint conditional likelihood we consider as a basis of our inference is given by

$$L_{cm} = \prod_{i=1}^{N} L_{cm}^{i} = \prod_{i=1}^{N} p\Big(\{R_{ij}, X_{ij}, Y_{ij}\}_{j=1}^{M+1} | \, \{\boldsymbol{Z}_{ij}\}_{j=1}^{M+1}, S_i, \sum_{j=1}^{M+1} Y_{ij} = k_i\Big).$$

The likelihood contribution of the i-th stratum is factorized as below,

$$\begin{aligned} L_{cm}^{i} &= p(\{R_{ij}, X_{ij}, Y_{ij}\}_{j=1}^{M+1} | \Big\{\boldsymbol{Z}_{ij}\Big\}_{j=1}^{M+1}, S_i, \sum_{j=1}^{M+1} Y_{ij} = k_i) \\ &= \prod_{j=1}^{M+1} \Big\{p(R_{ij}|X_{ij}, Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \times p(X_{ij}|Y_{ij}, \boldsymbol{Z}_{ij}, S_i)\Big\} \\ &\times \; p(Y_{i1} = k_i, Y_{i2} = \cdots = Y_{iM+1} = 0|\boldsymbol{Z}_{ij}, S_i, \sum_{j=1}^{M+1} Y_{ij} = k_i). \end{aligned}$$

In order to evaluate this likelihood, we first assume a selection probability model, namely,

$$p(R_{ij} = 1|X_{ij}, Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \;=\; H(\delta_0 + \delta_1 X_{ij} + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^{\top} \boldsymbol{Z}_{ij}), \tag{3.5}$$

where $H(u)$ is the logistic link function defined as $H(u) = \{1 + \exp(-u)\}^{-1}$.

We now need to specify a model for $p(X_{ij}|Y_{ij}, \boldsymbol{Z}_{ij}, S_i)$. However, using the results of Satten and Kupper (1993), Satten and Carroll (2000), and Sinha and Maiti (2008), by specifying a model for $p(X_{ij}|Y_{ij} = 0, \boldsymbol{Z}_{ij}, S_i)$, and the prospective disease risk model (3.3), one can obtain the distribution of X_{ij} in all disease subclasses, namely, $p(X_{ij}|Y_{ij} = k, \boldsymbol{Z}_{ij}, S_i)$, $k = 1, \cdots K$. This well-known result is presented in Lemma 1 of Appendix A.2.2. The last term in L_{cm}^{i}, which remains to be expressed as a function of the ingredients of the assumed model components, can be simplified as,

$$\begin{aligned} & p(Y_{i1} = k_i, Y_{i2} = \cdots = Y_{iM+1} = 0 | \boldsymbol{Z}_{ij}, S_i, \sum_{j=1}^{M+1} Y_{ij} = k_i) \\ & \qquad = \{p(Y_{i1} = k_i | \boldsymbol{Z}_{i1}, S_i)/p(Y_{i1} = 0 | \boldsymbol{Z}_{i1}, S_i)\}/ \\ & \qquad \sum_{j=1}^{M+1} p(Y_{ij} = k_i | \boldsymbol{Z}_{i1}, S_i)/p(Y_{ij} = 0 | \boldsymbol{Z}_{ij}, S_i). \end{aligned}$$

The marginal odds of the disease $p(Y = k|\boldsymbol{Z}, S)/p(Y = 0, \boldsymbol{Z}, S)$ can again be represented in terms of the control distribution for X and the parameters of the disease risk model. The exact representation is stated in Lemma 2 of Appendix A.2.2. The marginal likelihood of observed data after integrating with respect to the distribution of the missing exposure is given by,

$$\begin{aligned} L_{cm}^{obs} &= \prod_{i=1}^{N} \prod_{j=1}^{M+1} \Big\{ p^{R_{ij}}(R_{ij} = 1 | X_{ij}^o, Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \times p^{R_{ij}}(X_{ij}^o | Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \Big\} \\ &\times \int p^{(1-R_{ij})}(R_{ij} = 0 | X_{ij}^m, Y_{ij}, \boldsymbol{Z}_{ij}, S_i) dF(X_{ij}^m | Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \\ &\times \prod_{i=1}^{N} \Big[\{p(Y_{i1} = k_i | \boldsymbol{Z}_{i1}, S_i)/p(Y_{i1} = 0 | \boldsymbol{Z}_{i1}, S_i)\}/ \\ & \sum_{j=1}^{M+1} p(Y_{ij} = k_i | \boldsymbol{Z}_{ij}, S_i)/p(Y_{ij} = 0 | \boldsymbol{Z}_{ij}, S_i) \Big], \end{aligned} \tag{3.6}$$

where X^o denotes the observed X-values and X^m are the unobserved missing data on X.

Instead of Monte Carlo evaluation of the above integrated likelihood followed by maximization procedures, both of our estimation strategies FB and ECM will be based on the following complete data likelihood,

$$\begin{aligned} L_{cm}^{comp} &= \prod_{i=1}^{N} \prod_{j=1}^{M+1} \Big\{ p^{R_{ij}}(R_{ij} = 1 | X_{ij}^o, Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \times p^{R_{ij}}(X_{ij}^o | Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \\ &\times p^{(1-R_{ij})}(R_{ij} = 0 | X_{ij}^m, Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \times p^{(1-R_{ij})}(X_{ij}^m | Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \Big\} \\ &\times \prod_{i=1}^{N} \Big[\{p(Y_{i1} = k_i | \boldsymbol{Z}_{i1}, S_i)/p(Y_{i1} = 0 | \boldsymbol{Z}_{i1}, S_i)\}/ \\ & \sum_{j=1}^{M+1} p(Y_{ij} = k_i | \boldsymbol{Z}_{ij}, S_i)/p(Y_{i1} = 0 | \boldsymbol{Z}_{ij}, S_i) \Big]. \end{aligned} \tag{3.7}$$

Remark 1: Note that in our formulation so far, any parametric and non-parametric model can be used for the distribution of X. One can restrict attention to the class of exponential family models like in Paik (2004) or model the distribution of X non-parametrically by using a Dirichlet process mixture of normals model as in Mukherjee *et al.* (2007). The methodology is not restricted to a class of parametric models for X. In Appendix A.2.3, we provide the version of Lemma 1 and 2, specifically for the general class of exponential family of distributions. We illustrate specific examples with the Normal and Binomial distribution, just to provide the reader a sense of how the expressions can be simplified in such commonly occurring instances.

Remark 2: When the missingness mechanism is MAR, $p(R|X, Y, \boldsymbol{Z}, S) = p(R|Y, \boldsymbol{Z}, S)$ and the above likelihood L_{cm}^{comp} reduces to the likelihood used in Satten and Carroll (2000) and Sinha *et al.* (2005), by simply removing two terms in (3.7) involving the selection probability model. In that case, there is an implicit assumption that $p(R|Y, \boldsymbol{Z}, S)$ does not involve any parameters of interest, so the contribution of that term to the likelihood can be ignored in both ML and FB inference.

3.3 Parameter estimation and inference

3.3.1 The ECM approach

Based on the complete data likelihood L_{cm}^{comp}, we devise an ECM approach to estimate the model parameters. Let $\boldsymbol{\theta}$ denote the parameters governing the assumed control distribution $p(X|Z, S, D = 0)$. For example, if we assume that the exposure distribution in controls belongs to an exponential family, i.e.,

$$f(X_{ij}|Y_{ij} = 0, \mathbf{Z}_{ij}, S_i) = \exp[\xi_{ij}\{\theta_{ij}X_{ij} - b(\theta_{ij})\} + c(\xi_{ij}, X_{ij})],$$

where the canonical parameters θ_{ij} are modeled as a regression function of the completely observed covariates, namely, $\theta_{ij} = \kappa_0 + \boldsymbol{\kappa}_1^\top \mathbf{Z}_{ij} + \kappa_2 S_i$, capturing the dependence of the distribution X on $\boldsymbol{Z}$ and S and ξ_{ij} are the scale parameters. In that case, $\boldsymbol{\theta} = (\kappa_0, \boldsymbol{\kappa}_1, \kappa_2, \xi)$. If we denote the entire parameter vector, as $\Theta = (\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\delta})$, by using Lemma 1 and 2 of Appendix A.2.3, the complete-data log-likelihood, say $l_{cm}^{comp}(\Theta)$ can be obtained via taking log of (3.7) as,

$$\underbrace{\sum_{(i,j):R_{ij}=1} \left[\xi_{ij}\{\theta^*_{ij}X^o_{ij} - b(\theta^*_{ij})\} + c(\xi_{ij}, X^o_{ij}) + \log H(\delta_0 + \delta_1 X^o_{ij} + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij})\right]}_{L_1(\Theta)}$$

$$+ \underbrace{\sum_{(i,j):R_{ij}=0} \left[\xi_{ij}\{\theta^*_{ij}X^m_{ij} - b(\theta^*_{ij})\} + c(\xi_{ij}, X^m_{ij}) + \log\{1 - H(\delta_0 + \delta_1 X^m_{ij} + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij})\}\right]}_{L_2(\Theta)}$$

$$+ \underbrace{\sum_{i=1}^{N} \left\{\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{i1} + \xi_{i1}\{b(\theta^*_{i1}) - b(\theta_{i1})\} + \log\left(\sum_{j=1}^{M+1} \exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \xi_{ij}\{b(\theta^*_{ij}) - b(\theta_{ij})\}\right]\right)\right\}}_{L_3(\Theta)},$$

(3.8)

where $\theta^*_{ij} = \theta_{ij} + I(Y_{ij} = k_i)\xi_{ij}^{-1}\phi_{k_i}\beta_1$. Let $L_1(\Theta)$, $L_2(\Theta)$, and $L_3(\Theta)$ denote the first, the second, and the third term of the log likelihood in (3.8), we can characterize the E-step at the $(t+1)^{th}$ iteration of a standard EM algorithm by computing the expectation of $l_{cm}^{comp}(\Theta)$ as,

$$E\{l_{cm}^{comp}(\Theta^{(t+1)})\} = L_1(\Theta^{(t+1)}) + E\{L_2(\Theta^{(t+1)})\} + L_3(\Theta^{(t+1)}), \tag{3.9}$$

where the expectation E is taken with respect to $p(X^m_{ij}|Y_{ij}, \boldsymbol{Z}_{ij}, S_i, R_{ij} = 0, \Theta^{(t)})$ which in turn can be expressed as

$$\begin{aligned} &\frac{p(R_{ij} = 0|X^m_{ij}, Y_{ij}, \boldsymbol{Z}_{ij}, S_i)p(X^m_{ij}|Y_{ij}, \boldsymbol{Z}_{ij}, S_i)}{p(R_{ij} = 0|Y_{ij}, \boldsymbol{Z}_{ij}, S_i)} \\ = &\frac{p(R_{ij} = 0|X^m_{ij}, Y_{ij}, \boldsymbol{Z}_{ij}, S_i)p(X^m_{ij}|Y_{ij}, \boldsymbol{Z}_{ij}, S_i)}{\int p(R_{ij} = 0|X^m_{ij}, Y_{ij}, \boldsymbol{Z}_{ij}, S_i)dF(X^m_{ij}|Y_{ij}, \boldsymbol{Z}_{ij}, S_i)}. \end{aligned} \tag{3.10}$$

The integral in (3.10) is replaced by sum for a discrete exposure X. If we have a standard distributional form for (3.10), e.g., when X^m is binary, we can obtain an analytical expression for $E\{L_2(\Theta^{(t+1)})\}$. However, Monte Carlo generation may be necessary at the E-step, depending on the form of the distribution of $p(X_{ij}^m|Y_{ij}, \boldsymbol{Z}_{ij}, S_i)$. In the M-step, we maximize (3.9) at the $(t+1)^{th}$ iteration with respect to $\Theta^{(t+1)}$ conditioning on the previously obtained values of $\Theta^{(t)}$.

The above M-step may lead to computational complexity with high dimensional parameter spaces. To handle this difficulty, a modification was proposed by Meng and Rubin (1993) to accelerate the EM algorithm by replacing the M-step with a rather simpler conditional maximization (CM) step. With the non-linearity in ϕ and $\boldsymbol{\beta}$, adopting the ECM is extremely helpful for the stereotype link function where the EM often fails to converge. The ECM is in the spirit of Greenland's two-step procedure for stereotype models (Greenland, 1994), where the maximization problem is simplified by iteratively maximizing in terms of $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$. In the $(t+1)^{th}$ step of ECM, we maximize the likelihood in terms of $\boldsymbol{\beta}^{(t+1)}$, conditioning on previously obtained $(\boldsymbol{\phi}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\delta}^{(t)})$ rather than maximizing the joint likelihood in terms of all parameters $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\delta})$. Then we maximize the likelihood with respect to $\boldsymbol{\phi}^{(t+1)}$ conditioning on $(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\delta}^{(t)})$ and continue iteratively. Similar to the EM, we repeat E-step and CM-step until the convergence condition is met. In this article, the conditional maximization is performed via the Nelder-Mead optimization routine.

Remark 3: The standard errors corresponding to the estimated parameters can be obtained by inverting the observed Fisher information as prescribed in Louis (1982):

$$I(\Theta) = -E\left[\frac{\partial^2}{\partial\Theta\partial\Theta^\top}\left\{\log L_{cm}^{comp}(\Theta)\right\}\right]_{\Theta=\hat{\Theta}} . \tag{3.11}$$

We compute the above expectation with respect to the conditional distribution $p(X^m|Y, \boldsymbol{Z}, S, R = 0)$ by Monte carlo average of the second derivative of the log likelihood. Here, note that we evaluated each corresponding full dimensional Hessian via a numerical approximation employing Richardson extrapolations provided in R package 'hessian:numDeriv' not by conditioning on remaining estimated parameters as in two step type process, which is known to suffer from invalid standard error (Lall *et al.*, 2002).

3.3.2 Bayesian approach

Prior Specification: The likelihood used for Bayesian inference is again the complete data likelihood in (3.7). There are four subsets of parameters $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\delta})$ under consideration. Our main interest lies in $\boldsymbol{\beta}^{(p+1)\times 1} = (\beta_1, \boldsymbol{\beta}_2^{p\times 1})$ and $\boldsymbol{\phi}^{(K-1)\times 1} = (\phi_1, \cdots, \phi_{K-1})$ in the disease risk model (3.4). The two ancillary sets of parameters involve the $\boldsymbol{\delta}^{(p+4)\times 1} = (\delta_0, \delta_1, \delta_2, \delta_3, \boldsymbol{\delta}_4^{p\times 1})$ parameters in the selection probability model and the parameters $\boldsymbol{\theta} = (\kappa^{(p+2)\times 1}, \xi)$, where $\kappa^{(p+2)\times 1} = (\kappa_0, \boldsymbol{\kappa}_1^{p\times 1}, \kappa_2)$, used in modeling the exposure distribution in the control population. To formulate the full conditionals, we assume series of prior distributions on these four sets of parameters.

In this article, we generally consider the following set of mutually independent priors on Θ :

$$\begin{aligned} \pi(\boldsymbol{\beta}) &\overset{iid}{\sim} N_{(p+1)}(\boldsymbol{\mu}_\beta, \sigma_\beta^2 \boldsymbol{I}), \;\; \pi(\boldsymbol{\delta}) \overset{iid}{\sim} N_{(p+4)}(\boldsymbol{\mu}_\delta, \sigma_\delta^2 \boldsymbol{I}), \\ \pi(\boldsymbol{\kappa}) &\overset{iid}{\sim} N_{(p+2)}(\boldsymbol{\mu}_\kappa, \sigma_\kappa^2 \boldsymbol{I}), \;\; \pi(\boldsymbol{\phi}) \overset{iid}{\sim} N_{(K-1)}(\boldsymbol{\mu}_\phi, \sigma_\phi^2 \boldsymbol{I}). \end{aligned} \tag{3.12}$$

On ξ, the scale parameter of the exponential family, we adopt a suitable prior given the specific distribution, for example, we can assume a uniform prior on the logarithmic standard deviation for X^m following a Normal distribution. Based on the complete data likelihood in (3.7) and the priors described above, we can elicit full conditionals, that are described in detail for specific examples in the following section and in Appendix A.2.4.

Bayesian Computation: Following the data augmentation idea of Tanner and Wong (1987) we iterate the following two steps for iteratively generating observations from the joint full conditional of $(\boldsymbol{X}^m, \Theta | Y, Z, S, \boldsymbol{X}^o)$. At iteration $t+1$,

- (a) : Sample $\boldsymbol{X}^m_{(t+1)}$ from density $P(\boldsymbol{X}^m | \Theta_{(t)}, Y, \boldsymbol{X}^o, \boldsymbol{Z}, S, R)$,

- (b) : Sample $\Theta_{(t+1)}$ from density $P(\Theta | Y, \boldsymbol{X}^o, \boldsymbol{X}^m_{(t+1)}, \boldsymbol{Z}, S, R)$, where $\Theta^{(t)} = (\boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\delta}^{(t)})$ are obtained at the previous iteration t. As Tanner and Wong (1987) pointed out, the first step (a), where we sample $\boldsymbol{X}^m$ from the full conditional distribution, is analogous to 'multiple imputation' of filling in the missing data values. Also note that in step (a), we in fact, sample X^m from the same full conditional distribution that we use at the E-step in ECM as given in (3.9). In step (b), or the 'posterior' step, we generate posterior sample of Θ conditional on augmented data. However, instead of working with a finite number of imputed datasets as in multiple imputation, we iterate this process in our Monte Carlo sampling scheme and continue until stochastic convergence.

Given the full conditionals and employing the above data augmentation step, we use a Gibbs sampler (Geman and Geman, 1984) to generate samples from the full conditional distribution of $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\kappa}, \boldsymbol{\delta})$ given the augmented data. Note that though the full conditionals do not often have a standard form, they are log-concave when the distribution of X^m is assumed to belong to a general exponential family. In this case, we use the adaptive rejection sampling or ARS (Gilks and Wild, 1992). For situations when the full conditionals are not log-concave, we can adopt the adaptive rejection Metropolis sampling (ARMS) (Gilks *et al.*, 1995). For each parameter, we generate 50,000 posterior samples and discard the first 10,000 iterations as 'burn-in'. In order to reduce the inner-cycle correlation, a thinning of 5 observations was applied. We monitor convergence of the chains using the diagnostic 'potential scale reduction factor' (Gelman and Rubin, 1992) provided in the R package CODA (Plummer *et al.*, 2009). Finally, the remaining posterior sequences are analyzed for evaluating the Bayesian estimates and credible intervals.

3.3.3 Special Case: Binary Exposure

We turn our attention to the situation when exposure distribution in the controls arise from a Bernoulli distribution with $p_{ij} = H(\theta_{ij})$ where $\theta_{ij} = \kappa_0 + \boldsymbol{\kappa}_1^\top \boldsymbol{Z}_{ij} + \kappa_2 S_i$, namely,

$$f(X_{ij}|Y_{ij} = 0, \mathbf{Z}_{ij}, S_i) = \exp\left[\theta_{ij}X_{ij} - \log\left\{1 + \exp(\theta_{ij})\right\}\right].$$

We can then express the exposure distribution within sub-types of cases as $f(X_{ij}|Y_{ij} = k_i, \boldsymbol{Z}_{ij}, S_i) = \exp\left[\theta^*_{ij}X_{ij} - \log\left\{1 + \exp(\theta^*_{ij})\right\}\right]$ where $\theta^*_{ij} = \theta_{ij} + \phi_{k_i}\beta_1$ based on Lemma 1 of Appendix A.2.3. Using these expressions in (3.7), we have, for the Bernoulli case,

$$\begin{aligned}
L_{cm}^{comp} &= \prod_{i=1}^{N}\prod_{j=1}^{M+1}\Big[H(\delta_0 + \delta_1 X^o_{ij} + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij})^{R_{ij}} \\
&\times \left\{1 - H(\delta_0 + \delta_1 X^m_{ij} + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij})\right\}^{1-R_{ij}}\Big] \\
&\times \prod_{i:Y_{i1}=k_i}^{N} \exp\Big(\left[\{R_{ij}X^o_{ij} + (1 - R_{ij})X^m_{ij}\}(\theta_{ij} + \phi_{k_i}\beta_1) - \log\left\{1 + \exp(\theta_{ij} + \phi_{k_i}\beta_1)\right\}\right]\Big) \\
&\times \prod_{i=1}^{N}\prod_{j=2}^{M+1} \exp\Big(\{R_{ij}X^o_{ij} + (1 - R_{ij})X^m_{ij}\}\theta_{ij} - \log\left\{1 + \exp(\theta_{ij})\right\}\Big) \\
&\times \prod_{i=1}^{N} \frac{\exp\left[\phi_{k_i}\beta_2^\top \boldsymbol{Z}_{i1} + \log\left\{\frac{1+\exp(\theta_{i1}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{i1})}\right\}\right]}{\sum_{j=1}^{M+1}\exp\left[\phi_{k_i}\beta_2^\top \boldsymbol{Z}_{ij} + \log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]}. \qquad (3.13)
\end{aligned}$$

Both the ECM and FB inference are developed based on the above complete data likelihood. Similar expressions for X from a Normal distribution is presented in Appendix A.2.5.

The ECM: We can easily calculate the expected complete data log-likelihood in (3.13) based on the fact that $p(X^m_{ij} = 1|Y_{ij}, \boldsymbol{Z}_{ij}, S_i, R_{ij} = 0)$ has a Bernoulli distribution with known structure. Let $H\{\psi_{ij}(\theta)\} = p(X^m_{ij} = 1|Y_{ij}, \boldsymbol{Z}_{ij}, S_i, R_{ij} = 0)$ where $\psi_{ij}(\theta) = \theta_{ij} + I(Y_{ij} = k_i)\phi_{k_i}\beta_1 + \log\{\bar{\pi}_{ij}(1)/\bar{\pi}_{ij}(0)\}$. Here $\pi_{ij}(s)$ denotes $H(\delta_0 + \delta_1 s + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij})$ and $\bar{\pi}_{ij}(s) = 1 - \pi_{ij}(s)$. We can now express the three terms in (3.13) as:

$$\begin{aligned}
L_1(\Theta^{(t+1)}) &= \sum_{(i,j):R_{ij}=1} \big[X_{ij}\{\theta_{ij} + I(Y_{ij} = k_i)\phi_{k_i}\beta_1\} \\
&- \log\left[1 + \exp\{\theta_{ij} + I(Y_{ij} = k_i)\phi_{k_i}\beta_1\}\right] + \log\{\pi_{ij}(X_{ij})\}\big], \\
E\{L_2(\Theta^{(t+1)})\} &= \sum_{(i,j):R_{ij}=0} (H\{\psi_{ij}(\theta)\}[\theta_{ij} + I(Y_{ij} = k_i)\phi_{k_i}\beta_1 + \log\{\bar{\pi}_{ij}(1)\}]) \\
&+ \sum_{(i,j):R_{ij}=0} \Big([1 - H\{\psi_{ij}(\theta)\}]\log\{\bar{\pi}_{ij}(0)\} - \log\left[1 + \exp\{\theta_{ij} + I(Y_{ij} = k_i)\phi_{k_i}\beta_1\}\right]\Big), \\
L_3(\Theta^{(t+1)}) &= \sum_{i=1}^{N}\left\{\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{i1} + \log\left\{\frac{1 + \exp(\theta_{i1} + \phi_{k_i}\beta_1)}{1 + \exp(\theta_{i1})}\right\}\right. \\
&- \left.\log\Big(\sum_{j=1}^{M+1} \exp\Big[\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \log\left\{\frac{1 + \exp(\theta_{ij} + \phi_{k_i}\beta_1)}{1 + \exp(\theta_{ij})}\right\}\Big]\Big)\right\}.
\end{aligned}$$

We then follow the ECM steps outlined in Section 3.3.1.

The Bayesian Route:

We can obtain the following full conditional distributions of the model parameters, given the augmented data, by using the likelihood in (3.13) and the prior structure as in (3.12).

$$\begin{aligned}
\pi(\beta_1|\cdot) &\propto \frac{\exp\left(-\frac{1}{2\sigma^2_{\beta_1}}\left[\beta_1 - \mu_{\beta_1} - \sigma^2_{\beta_1}\sum_{i=1}^N \phi_{k_i}\left\{R_{i1}X^o_{i1} + (1 - R_{i1})X^m_{i1}\right\}\right]^2\right)}{\prod_{i=1}^N \sum_{j=1}^{M+1} \exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]}, \\
\pi(\beta_{2r}|\cdot) &\propto \frac{\exp\left\{-\frac{1}{2\sigma^2_{\beta_2}}(\beta_{2r} - \mu_{\beta_{2r}} - \sigma^2_{\beta_2}\sum_{i=1}^N \phi_{k_i}Z_{i1r})^2\right\}}{\prod_{i=1}^N \sum_{j=1}^{M+1} \exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]}, \\
\pi(\phi_k|\cdot) &\propto \frac{\exp\left\{-\frac{\left(\phi_k - \mu_{\phi_k} - 2\sigma^2_{\phi_k}\sum_{i=1}^N I(Y_{i1}=k)\left[\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{i1} + \beta_1\{R_{i1}X^o_{i1} + (1-R_{i1})X^m_{i1}\}\right]\right)^2}{2\sigma^2_{\phi_k}}\right\}}{\prod_{i=1}^N \sum_{j=1}^{M+1} \exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]}, \\
\pi(\delta_q|\cdot) &\propto \frac{\exp\left\{-\frac{1}{2\sigma^2_\delta}(\delta_q - \mu_{\delta_q} - \sigma^2_\delta\sum_{i=1}^N\sum_{j=1}^{M+1} R_{ij}V_{ijq})^2\right\}}{\prod_{i=1}^N\prod_{j=1}^{M+1}\left[1 + \exp\left\{\delta_0 + \delta_1(R_{ij}X^o_{ij} + (1 - R_{ij})X^m_{ij}) + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij}\right\}\right]}, \\
\text{where} \quad & V_{ij0} = 1,\ V_{ij1} = Y_{ij},\ V_{ij2} = X_{ij},\ V_{ij3} = S_i. \\
\pi(\delta_{4r}|\cdot) &\propto \frac{\exp\left\{-\frac{1}{2\sigma^2_\delta}(\delta_{4r} - \mu_{\delta_{4r}} - \sigma^2_\delta\sum_{i=1}^N\sum_{j=1}^{M+1} R_{ij}Z_{ijr})^2\right\}}{\prod_{i=1}^N\prod_{j=1}^{M+1}\left[1 + \exp\left\{\delta_0 + \delta_1(R_{ij}X^o_{ij} + (1 - R_{ij})X^m_{ij}) + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij}\right\}\right]},
\end{aligned}$$

$$
\begin{aligned}
\pi(\kappa_0|\cdot) \;&\propto\; \frac{\exp\left(-\frac{1}{2\sigma_\kappa^2}\left[\kappa_0-\mu_{\kappa_0}-\sigma_\kappa^2\sum_{i=1}^N\sum_{j=1}^{M+1}\left\{R_{ij}X_{ij}^o+(1-R_{ij})X_{ij}^m\right\}\right]^2\right)}{\prod_{i=1}^N\sum_{j=1}^{M+1}\exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top\boldsymbol{Z}_{ij}+\log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]} \\
&\times\; \prod_{i=1}^N\prod_{j=1}^{M+1}\frac{1}{1+\exp(\theta_{ij})}, \\
\pi(\kappa_{1r}|\cdot) \;&\propto\; \frac{\exp\left(-\frac{1}{2\sigma_\kappa^2}\left[\kappa_{1p}-\mu_{\kappa_{1p}}-\sigma_\kappa^2\sum_{i=1}^N\sum_{j=1}^{M+1}\left\{R_{ij}X_{ij}^o+(1-R_{ij})X_{ij}^m\right\}Z_{ijr}\right]^2\right)}{\prod_{i=1}^N\sum_{j=1}^{M+1}\exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top\boldsymbol{Z}_{ij}+\log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]} \\
&\times\; \prod_{i=1}^N\prod_{j=1}^{M+1}\frac{1}{1+\exp(\theta_{ij})},\; r=1,\ldots,p, \\
\pi(\kappa_2|\cdot) \;&\propto\; \frac{\exp\left(-\frac{1}{2\sigma_\kappa^2}\left[\kappa_2-\mu_{\kappa_2}-\sigma_\kappa^2\sum_{i=1}^N\sum_{j=1}^{M+1}\left\{R_{ij}X_{ij}^o+(1-R_{ij})X_{ij}^m\right\}S_i\right]^2\right)}{\prod_{i=1}^N\sum_{j=1}^{M+1}\exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top\boldsymbol{Z}_{ij}+\log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]} \\
&\times\; \prod_{i=1}^N\prod_{j=1}^{M+1}\frac{1}{1+\exp(\theta_{ij})}
\end{aligned}
$$

where $r=1,\ldots,p,\; k=1,\ldots,K-1$ and $q=0,\ldots,3$. Conditional on the current value of the sampled Θ, we sample X_{ij}^m from the conditional distribution $p(X_{ij}^m=1|Y_{ij},\boldsymbol{Z}_{ij},S_i,R_{ij}=0)=H\{\psi_{ij}(\theta)\}$, where $\psi_{ij}(\theta)$, is exactly as defined in the ECM approach. The Bayesian iterative computation scheme as described in Section 3.3.2 is then followed.

3.4 Example: The Molecular Epidemiology of Colorectal Cancer Study

Colorectal cancer (CRC) is the third most common cancer in the western world (WHO, 2006). The Molecular Epidemiology of Colorectal Cancer (MECC) study is a population-based case-control study of patients diagnosed with colorectal cancer in northern Israel between March 31, 1998 and March 31, 2004. Controls were 1:1 matched according to age, sex, clinic, and ethnic group (Jewish vs. non-Jewish). Subjects were interviewed on an array of dietary and behavioral risk factors including levels of physical activity and use of medications. Physical activity is known to reduce the risk of CRC by 30 to 40 percent according to the informational website of the National Cancer Institute (NCI, 2009). In the MECC dataset 20% of subjects had missing information on the variable measuring participation in sports or other physical activities. In a high profile article from the MECC study, (Poynter *et al.*, 2005) were the first to point out that the use of statins, a drug used for hypercholesterol, can reduce the risk of colorectal cancer (reported OR 0.53, 95% CI: (0.38,0.74)) after adjusting for other known risk factors, like physical activity. However, no analysis stratified in terms of subtypes of CRC were done in the original study. In the current Chapter, we consider CRC Stage, assigned according to the TNM (Tumor, Node, Metastasis) criteria recommended by American Joint Committee on Cancer (AJCC, 2002) as our categorical outcome ranging from (0 to IV) that represents different degree of disease progression. We investigate the effect of physical activity and statin use across CRC stages via fitting the stereotype model.

We analyzed data on 1,841 matched pairs with completely observed data on CRC stage (Y) and statin use (Z) and partially missing data on physical activity (X). In our analysis, we treat age as a single matching variable S that can affect our selection probability model and the model for control distribution of X. To avoid sparse frequencies, the cancer stage variable Y, was re-grouped into four categories 0 (consisting of 1841 controls), 1 (Stages I), 2 (Stages II), and 3 (Stages III and IV). The distribution of subjects in the three case categories were 306 (16.6%), 844 (45.9%), and 691 (37.5%) respectively. The completely observed variable Z or statin use contained 90% “No” and 10% “Yes”. While physical activity or X contained 29% “No”, 51% “Yes”, and 20% missing values. Age (observed range 19-97) was linearly transformed into a [0, 1] interval. The empirical distribution of transformed age was well-approximated by a Normal distribution with mean 0.64 and sd 0.14.

We analyzed the MECC data by (a) the direct maximizing the conditional likelihood (3.4) with completely observed data (CMLE), (b) the ECM approach, and (c) the full Bayesian method (FB). In order to obtain the CMLE estimates based on complete data, we used direct maximization of (3.4) via the Nelder-Mead optimization. Note that using CMLE restricted to completely observed data result in 33% loss of information due to deletion of the entire stratum with any missing covariate. We allowed the missingness mechanism to potentially depend on (Y, X, Z, S) under ECM and FB. For FB, we choose a relatively non-informative $N(0, 10^4)$ prior on each component of Θ as described in (3.12).

We present the results of this analysis in Table 3.1. For computing standard errors corresponding to CMLE and ECM, we inverted the observed Fisher information matrix based on complete data and the Monte Carlo evaluated conditional expectation of the Fisher information matrix as specified in (3.11) respectively. The posterior standard deviations (PSD) for the FB approach were obtained from the standard deviation of the generated posterior sequence. All three methods produced similar estimates of β_1 and β_2. The estimated covariate-specific coefficients imply the protective effect of physical activity and use of statins across CRC stages and the effects are highly significant under all methods. Both FB and ECM have smaller standard errors than the CMLE, due to gain in information by properly using partially observed covariate information. FB and ECM are fairly comparable in terms of the standard errors of the parameter estimates.

Note that the estimated stage-specific parameters $\boldsymbol{\phi}$ are also fairly consistent across FB and ECM, while CMLE shows certain numerical differences. It is fairly clear from the analysis that the protective effect of statins and participation in sports are not homogeneous across different stages of cancer, as the values of ϕ_1 and ϕ_2 differ significantly. A large estimate of ϕ_2, approximately 1.80 from both ECM and FB, indicates that the protective effects were more pronounced in Stage 2. The estimates of ϕ_1 and ϕ_2 also imply that there is departure from monotone ordering of the categories in terms of covariate effects, thus the ordered Stereotype model (Anderson, 1984) does not appear to be appropriate for the

Table 3.1: Analysis results of the matched MECC study data with participation in sports X and statin use Z covariates with corresponding coefficients β_1 and β_2. 1841 cases are 1:1 matched to controls. For the CMLE, the conditional likelihood (3.4) is directly maximized with completely observed data. Under the FB methods the 'Est.' corresponds to the posterior mean whereas PSD corresponds to posterior standard deviation. For the disease risk parameters, we present 95% Wald confidence intervals (CMLE and ECM) whereas for FB we present 95% Highest Posterior Density (HPD) intervals.

	Method								
	CMLE			ECM			FB		
Parameter	Est.	SD	(95% CI)	Est.	SD	(95% CI)	Est.	PSD	(95% HPD)
β_1	-0.34	0.12	(-0.57,-0.11)	-0.34	0.09	(-0.55, -0.20)	-0.31	0.09	(-0.47, -0.14)
β_2	-0.71	0.19	(-1.08,-0.33)	-0.75	0.16	(-0.96, -0.34)	-0.72	0.15	(-1.03, -0.42)
ϕ_1	0.56	0.41	(-0.25, 1.36)	0.62	0.37	(0.01, 1.47)	0.62	0.32	(0.00, 1.25)
ϕ_2	1.59	0.51	(0.60, 2.58)	1.56	0.46	(0.87, 2.68)	1.67	0.43	(0.98,2.52)
κ_0				0.10	0.18	(-0.25,0.45)	0.05	0.09	(-0.13, 0.22)
κ_1				0.13	0.12	(-0.11, 0.36)	0.12	0.12	(-0.11, 0.35)
κ_2				-0.97	0.26	(-1.48,-0.46)	-0.80	0.18	(-1.13,-0.45)
δ_0				0.58	0.22	(0.15, 1.01)	0.68	0.14	(0.40, 0.93)
δ_1				0.06	0.10	(-0.14, 0.26)	0.03	0.15	(-0.25, 0.33)
δ_2				-0.00	0.04	(-0.08, 0.08)	-0.01	0.04	(-0.09,0.07)
δ_3				0.24	0.16	(-0.07, 0.55)	0.29	0.15	(-0.43, 0.35)
δ_4				-0.14	0.31	(-0.75, 0.47)	-0.04	0.22	(-0.47, 0.39)

current analysis. In fact, the posterior probability of the ordering of the categories, i.e., $p(\phi_0 \equiv 0 < \phi_1 < \phi_2 < \phi_3 \equiv 1|\text{Data})$ was computed from the posterior samples as 0.0012, indicating no evidence in favor of the ordered stereotype model.

We will like to point out that in the above stereotype model, the log odds-ratio parameters corresponding to each category k as compared to the controls, is obtained by the parameters $\phi_k\beta_1$ (for X) and $\phi_k\beta_2$ (for Z), $k = 1, 2, 3$. Bayesian inference has the added advantage of directly generating the posterior of these log odds-ratio parameters directly, instead of resorting to delta theorems and variance approximations needed in frequentist inference. Based on the FB analysis, the posterior estimate (95% HPD) of the odds-ratios (relative to controls) for physical activity corresponding to categories 1, 2 and 3 are 0.83 (0.70,0.99), 0.62 (0.51,0.73), 0.74 (0.62,0.88) respectively. For use of statins, the corresponding odds ratios are given by 0.65 (0.42,0.94), 0.31 (0.21,0.44) and 0.48 (0.36,0.65) respectively. Figure 3.1 presents estimated posterior densities of the log odds ratios of each CRC stage versus controls with respect to participation in sports and use of statins respectively. As pointed out earlier, non-monotone trend in the log odds ratios demonstrates that the ordering assumption

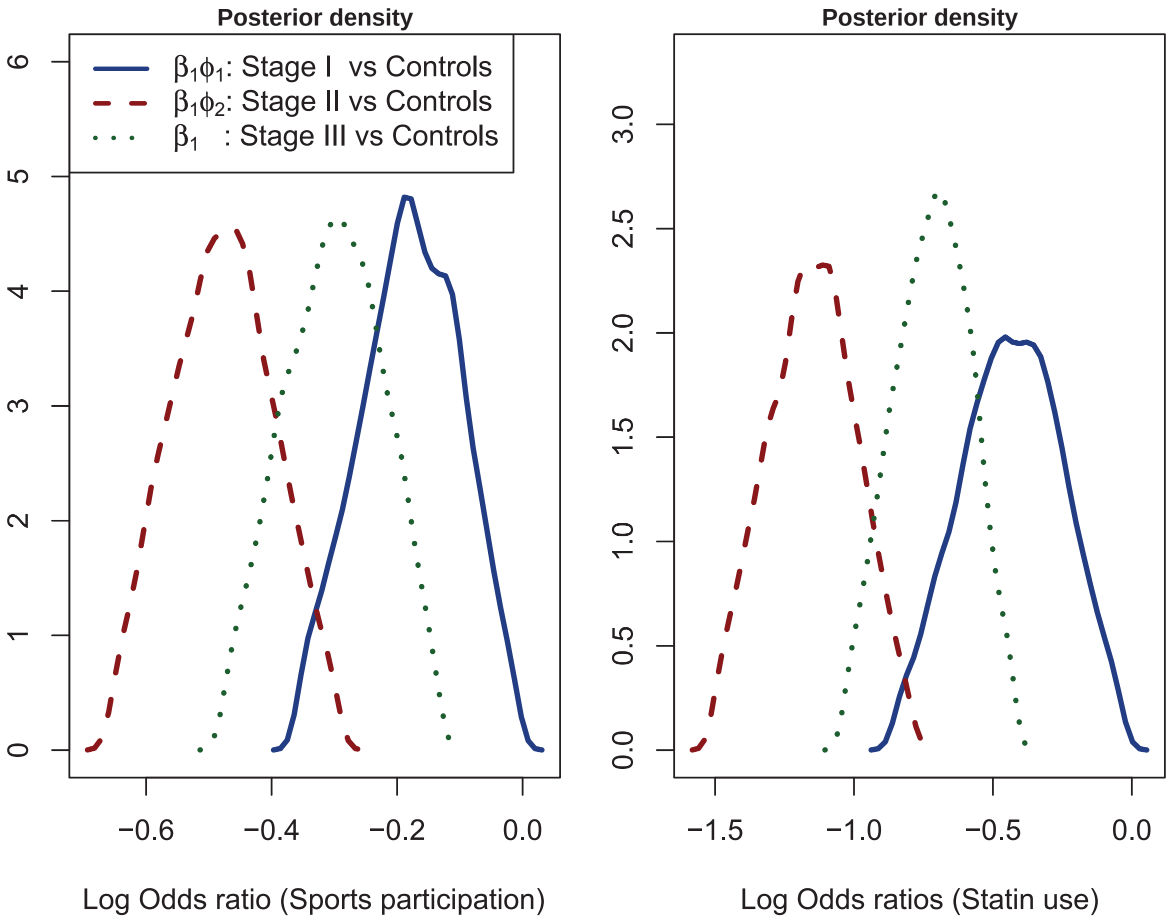


Figure 3.1: Posterior density plot corresponding to the log odds ratio parameters in 1:1 matched MECC study data with numerical summaries and estimates as presented in Table 3.1. The left plot corresponds to participation in sports (X) and the right plot corresponds to statin use (Z). The results are based on 10,000 samples generated from the posterior distribution of each parameter.

regarding the category specific parameters is not tenable for this study. We also tried fitting a proportional odds model to the completely observed data, ignoring the stratification due to matching and the proportional odds assumption was clearly violated with each collapsing of the stage category leading to significantly different estimates for the cumulative relative risk parameter corresponding to each covariate.

Regarding the missing data model, the coefficient δ_1 is not statistically significant under both ECM and FB with p-value 0.55 corresponding to the Wald test for ECM, suggesting that the missingness does not depend on X (physical activity). We can also note the agreement of the point estimates corresponding to the CMLE and after modeling missingness mechanism (ECM and FB).

3.5 Simulation Study

We evaluate and compare the performances of the three methods by conducting a small-scale simulation study. The purpose of the simulation study was to assess the methods under various models for the selection probability and the exposure distribution in terms of efficiency and robustness under model misspecification. Mimicking the real data analysis results, we fixed our true parameter values $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\delta})$ in the range of the point estimates obtained by the three methods. We first generate a large cohort of 500,000 subjects, containing information on (Y, X, Z, S). Akin to the statin use variable, we generated Z, from a Bernoulli distribution of success $p = 0.1$. We then generated a potential matching variable S from a Normal(0.6, 0.1^2) distribution, mirroring the age variable in the MECC study. Conditional on Z and S, we generated a binary X from several probability mechanisms as described below in detail. Conditional on X, Z, we generated Y from an unmatched stereotype model. We set the covariate specific parameters $\beta_1, \beta_2 =$(-0.3, -0.7) and the category specific scores as $\boldsymbol{\phi} = (\phi_0, \phi_1, \phi_2) = (0, 0.8, 1.7, 1)$. We selected the three case-category specific intercepts as (-1.5, -0.5, -0.9) to make the relative frequency distribution of Y similar to the real data analysis. With this large population base of 500,000 records on $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{S}$, we created a matched case-control dataset in the following way. First, we randomly sampled 1,000 cases $(Y \neq 0)$ from this large population. Corresponding to each selected case, we chose a matched control randomly from the set of all controls having the value of the matching variable S within 0.05 of the S-value for the selected case. We replicated the aforementioned process 200 times to create 200 matched case-control datasets from this large population under each simulation setting.

Under each simulation configuration, we considered five different schemes of selection probability models. The first four models fall under the class of missingness models we consider in (3.5), whereas MM5 involves non-linear terms in X and Y and violate the modeling assumption of (3.5).

MM1. Missing Completely at Random (MCAR) : $logit\{p(R_{ij}=1|Y_{ij},X_{ij},Z_{ij},S_i)\}$=0.8,

MM2. Missing at Random (MAR) : $logit\{p(R_{ij}=1|Y_{ij},X_{ij},Z_{ij},S_i)\}$=$Y_{ij}+0.5$,

MM3. Informative Missingness (IM): $logit\{p(R_{ij}=1|Y_{ij},X_{ij},Z_{ij},S_i)\}$=$X_{ij}$,

MM4. IM : $logit\{p(R_{ij}=1|Y_{ij},X_{ij},Z_{ij},S_i)\}$=$0.5X_{ij}+0.5Y_{ij}+0.5$,

MM5. IM : $logit\{p(R_{ij}=1|Y_{ij},X_{ij},Z_{ij},S_i)\}$=$X_{ij}Z_{ij}+Y_{ij}X_{ij}+1$.

The parameters for the above models are chosen in a way to yield the marginal probability of missingness to approximately 20% in each case.

To assess the robustness of our proposed methods under different departures from the assumed model for missing exposure, we consider three scenarios : (a) The exposure model is correctly specified (Table 3.2); (b) The exposure model is mis-sepecified in terms of a covariate (Table 3.3); (c) The exposure model is misspecified in terms of a link function (Table 3.4).

Under each simulation setting, we evaluated the performance of three methods: CMLE, ECM, and FB. The corresponding results are presented in terms of the average bias and mean squared errors across the 200 datasets (Tables 3.2-3.4). In approximately 3% cases, we failed to obtain estimates from the CMLE approach due to lack of convergence and those simulation iterations are deleted for a fair comparison across the three methods .

Table 3.2 presents simulation results when the exposure model is correctly specified. We generated exposure $X|Z,S$ from $H(0.3+0.3Z-1.5S)$. In the presence of non-informative missingness (MM1, MM2), the CMLE yields less efficient estimates than the ECM and the FB methods while all three methods are approximately unbiased. With informative missingness and a correctly specified selection model (MM3, MM4), the ECM and the FB produce less biased estimates than the CMLE in terms of β_1, the coefficient corresponding to X, which is noted to be affected most in presence of missingness. When model violation exists in terms of the selection probability model having non-linear product terms XZ and YX (MM5), all three methods produce large biases. Overall, the FB appears to have slightly better mean squared error properties than the ECM.

To assess the effect of model misspecification in the exposure distribution, for example, due to missing a correct covariate term, we introduce a quadratic term S^2, and generate $X|Z, S$ from $H(0.3+0.3Z-1.5S^2)$ everything else being identical to Table 3.2 settings. Contrary to our expectation that the full-likelihood based estimates from both FB and ECM will yield enhanced biases compared to the CMLE, which does not make any parametric assumption regarding the exposure distribution, we notice that the results are fairly similar across Table 3.3 and Table 3.2 for MM1-MM4 though there is marginally larger bias compared to Table 3.2. This can be possibly explained by the fact that S^2 and S are not abundantly apart to affect the estimation. Model misspecification in both selection probability and the exposure distribution (MM5), however, results in substantial increase in bias and MSE in the ECM and the FB as shown under MM5.

Lastly, we investigate the situation where the link function corresponding to generating $X|Z, S$ departs from the logistic link function. Here we generated $X|Z, S$ from a mixture of the Burr family of distributions (Burr, 1942),

$$X|\boldsymbol{Z}, S \sim \begin{cases} \text{Bernoulli with } p(X=1|\boldsymbol{Z}, S) = 1-\{1+\exp(0.3+0.3Z)\}^{-0.7}, & S < 0.5 \\ \text{Bernoulli with } p(X=1|\boldsymbol{Z}, S) = 1-\{1+\exp(0.3+0.3Z)\}^{-1.3}, & S \geq 0.5. \end{cases}$$

The biases corresponding to the FB and the ECM in Table 3.4 increase when compared to Table 3.2 and Table 3.3 with some loss in efficiency. This indicates that this type of

Table 3.2: Simulation results under correct specification of the exposure model. Here, binary exposure $X|Z,S$ is generated from $P(X=1|Z,S) = H(0.3+0.3Z-1.5S)$. The CMLE, the ECM and the FB methods are considered. The results are based on 200 simulated datasets, each with 1,000 cases and 1,000 controls. For each parameter of interest in the disease risk model, we report estimated bias and mean squared error based on the 200 replications. The true values for the parameters of interest are: $\beta_1 = -0.3$, $\beta_2 = -0.7$, $\phi_1 = 0.8$, and $\phi_2 = 1.7$.

	Method					
	CMLE		ECM		FB	
Parameter	Bias	MSE	Bias	MSE	Bias	MSE
			Complete Data			
β_1	0.007	0.009	0.007	0.008	0.030	0.008
β_2	-0.007	0.029	0.002	0.021	0.039	0.021
ϕ_1	-0.053	0.145	-0.072	0.102	-0.036	0.112
ϕ_2	-0.003	0.285	0.003	0.200	0.108	0.223
	MM1. logit$\{p(R_{ij}=1\|Y_{ij},X_{ij},Z_{ij},S_i)\}$=0.8					
β_1	-0.023	0.021	-0.015	0.010	0.058	0.011
β_2	-0.046	0.068	-0.027	0.033	0.006	0.031
ϕ_1	0.046	0.301	-0.009	0.132	0.069	0.151
ϕ_2	0.071	0.371	-0.002	0.208	0.112	0.236
	MM2. logit$\{p(R_{ij}=1\|Y_{ij},X_{ij},Z_{ij},S_i)\}$=$Y_{ij}+0.5$					
β_1	-0.035	0.034	0.033	0.011	-0.010	0.025
β_2	-0.067	0.100	-0.016	0.046	-0.011	0.041
ϕ_1	-0.082	0.307	0.010	0.279	0.026	0.190
ϕ_2	0.070	0.387	0.090	0.293	0.050	0.277
	MM3. logit$\{p(R_{ij}=1\|Y_{ij},X_{ij},Z_{ij},S_i)\}$=$X_{ij}+1$					
β_1	-0.016	0.026	-0.011	0.013	0.059	0.015
β_2	-0.050	0.075	-0.020	0.040	-0.001	0.039
ϕ_1	0.170	0.485	0.118	0.158	0.165	0.175
ϕ_2	0.202	0.647	0.092	0.226	0.171	0.283
	MM4. logit$\{p(R_{ij}=1\|Y_{ij},X_{ij},Z_{ij},S_i)\}$=$0.5X_{ij}+0.5Y_{ij}+0.5$					
β_1	-0.115	0.043	-0.036	0.012	-0.051	0.017
β_2	-0.057	0.064	-0.033	0.024	-0.048	0.028
ϕ_1	0.051	0.367	0.050	0.126	0.052	0.099
ϕ_2	0.053	0.686	-0.006	0.150	-0.076	0.245
	MM5. logit$\{p(R_{ij}=1\|Y_{ij},X_{ij},Z_{ij},S_i)\}$=$X_{ij}Z_{ij}+Y_{ij}X_{ij}+1$					
β_1	0.186	0.036	0.170	0.035	0.202	0.046
β_2	-0.126	0.102	0.065	0.043	0.081	0.038
ϕ_1	0.104	0.377	0.046	0.358	0.087	0.259
ϕ_2	0.207	0.969	0.329	0.672	0.342	0.531

link misspecification is possibly more severely affecting the parametric methods of the ECM and the FB than covariate misspecification. Thus the performance of our methods can be dependent upon the nature of the departure from the correct exposure model, producing slightly larger biases than CMLE under MAR data (MM1-MM2). However, with IM, both the ECM and the FB lead to improved Bias and MSE properties than the CMLE as the exposure misspecification bias appears to be less, compared to the bias generated by failure to account for non-ignorable missingness.

Summarizing our findings, our proposed methods present more efficient estimates than

Table 3.3: Simulation results under exposure model misspecification in terms of non-linear predictor in the exposure model. Here, a binary exposure $X|Z, S$ is generated under $P(X = 1|Z, S) = H(0.3 + 0.3Z - 1.5S^2)$. The CMLE, the ECM and the FB methods are considered. The results are based on 200 simulated datasets, each with 1,000 cases and 1,000 controls. For each parameter we report estimated bias and mean squared error based on the 200 replications. The true values for the parameters are: $\beta_1 = -0.3$, $\beta_2 = -0.7$, $\phi_1 = 0.8$, and $\phi_2 = 1.7$.

	Method					
	CMLE		ECM		FB	
Parameter	Bias	MSE	Bias	MSE	Bias	MSE
	Complete Data					
β_1	-0.011	0.011	-0.012	0.010	0.013	0.009
β_2	-0.052	0.032	-0.049	0.031	-0.011	0.029
ϕ_1	0.038	0.126	0.036	0.119	0.080	0.143
ϕ_2	0.010	0.182	0.017	0.172	0.113	0.206
	MM1. logit$\{p(R_{ij} = 1\|Y_{ij}, X_{ij}, Z_{ij}, S_i)\}$*=0.8*					
β_1	-0.045	0.023	-0.028	0.017	0.038	0.012
β_2	-0.029	0.068	0.006	0.039	0.023	0.035
ϕ_1	0.047	0.464	0.028	0.261	0.077	0.232
ϕ_2	0.096	0.501	0.090	0.291	0.087	0.290
	MM2. logit$\{p(R_{ij} = 1\|Y_{ij}, X_{ij}, Z_{ij}, S_i)\}=Y_{ij} + 0.5$					
β_1	-0.025	0.021	0.046	0.012	-0.021	0.035
β_2	-0.035	0.054	0.044	0.034	0.040	0.043
ϕ_1	0.044	0.499	-0.039	0.201	0.032	0.119
ϕ_2	0.131	0.503	0.123	0.419	0.063	0.343
	MM3. logit$\{p(R_{ij} = 1\|Y_{ij}, X_{ij}, Z_{ij}, S_i)\}=X_{ij} + 1$					
β_1	-0.005	0.015	-0.016	0.011	0.071	0.013
β_2	-0.008	0.062	-0.018	0.027	0.024	0.026
ϕ_1	0.085	0.262	0.055	0.118	0.120	0.135
ϕ_2	0.194	0.565	0.056	0.172	0.151	0.209
	MM4. logit$\{p(R_{ij} = 1\|Y_{ij}, X_{ij}, Z_{ij}, S_i)\}=0.5X_{ij} + 0.5Y_{ij} + 0.5$					
β_1	-0.132	0.046	-0.044	0.016	-0.055	0.020
β_2	-0.029	0.060	-0.013	0.037	-0.044	0.028
ϕ_1	-0.052	0.489	0.006	0.124	0.038	0.101
ϕ_2	0.037	0.725	0.070	0.301	-0.052	0.235
	MM5. logit$\{p(R_{ij} = 1\|Y_{ij}, X_{ij}, Z_{ij}, S_i)\}=X_{ij}Z_{ij} + Y_{ij}X_{ij} + 1$					
β_1	0.176	0.048	0.187	0.041	0.222	0.036
β_2	-0.086	0.109	0.047	0.048	0.033	0.041
ϕ_1	0.003	0.570	0.026	0.340	0.102	0.294
ϕ_2	0.243	1.096	0.390	0.867	0.194	0.670

Table 3.4: Simulation results under misspecification in terms of link function corresponding to the exposure distribution. Here, a binary $X|Z, S$ is generated from a mixture of Burr family of link functions, $P(X = 1|Z, S) = 1 - \{1 + \exp(0.3 + 0.3Z)\}^{-0.7}$ when $S < 0.5$ and $P(X = 1|Z, S) = 1 - \{1 + \exp(0.3 + 0.3Z)\}^{-1.3}$ otherwise. The CMLE, the ECM and the FB methods are considered. The results are based on 200 simulated datasets, each with 1,000 cases and 1,000 controls. For each parameter we report estimated bias and mean squared error based on the 200 replications. The true values for the parameters are: $\beta_1 = -0.3$, $\beta_2 = -0.7$, $\phi_1 = 0.8$, and $\phi_2 = 1.7$.

	Method					
	CMLE		ECM		FB	
Parameter	Bias	MSE	Bias	MSE	Bias	MSE
Complete Data						
β_1	-0.008	0.011	-0.021	0.015	0.017	0.011
β_2	-0.036	0.030	0.081	0.019	0.007	0.036
ϕ_1	0.041	0.149	0.098	0.114	0.050	0.142
ϕ_2	0.014	0.273	0.185	0.297	0.131	0.283
MM1. logit$\{p(R_{ij} = 1\|Y_{ij}, X_{ij}, Z_{ij}, S_i)\}$=*0.8*						
β_1	-0.033	0.021	-0.064	0.016	0.045	0.015
β_2	-0.030	0.069	0.058	0.035	0.013	0.034
ϕ_1	0.008	0.328	0.124	0.241	-0.016	0.194
ϕ_2	0.086	0.539	0.112	0.330	0.101	0.299
MM2. logit$\{p(R_{ij} = 1\|Y_{ij}, X_{ij}, Z_{ij}, S_i)\}$=$Y_{ij} + 0.5$						
β_1	-0.041	0.023	0.082	0.013	0.089	0.014
β_2	-0.015	0.052	0.076	0.050	0.023	0.039
ϕ_1	-0.025	0.342	0.161	0.301	0.021	0.214
ϕ_2	0.109	0.601	0.148	0.463	0.110	0.387
MM3. logit$\{p(R_{ij} = 1\|Y_{ij}, X_{ij}, Z_{ij}, S_i)\}$=$X_{ij} + 1$						
β_1	0.173	0.051	0.154	0.036	0.119	0.043
β_2	-0.151	0.077	0.095	0.052	0.041	0.029
ϕ_1	0.043	0.389	-0.094	0.203	0.077	0.261
ϕ_2	-0.008	0.644	-0.016	0.240	0.246	0.454
MM4. logit$\{p(R_{ij} = 1\|Y_{ij}, X_{ij}, Z_{ij}, S_i)\}$=$0.5X_{ij} + 0.5Y_{ij} + 0.5$						
β_1	-0.017	0.045	0.065	0.022	0.051	0.021
β_2	0.019	0.057	0.001	0.044	0.034	0.026
ϕ_1	0.196	0.412	0.094	0.191	0.099	0.189
ϕ_2	0.238	0.847	0.125	0.318	0.102	0.332
MM5. logit$\{p(R_{ij} = 1\|Y_{ij}, X_{ij}, Z_{ij}, S_i)\}$=$X_{ij}Z_{ij} + Y_{ij}X_{ij} + 1$						
β_1	0.201	0.063	0.179	0.072	0.202	0.077
β_2	0.087	0.123	0.074	0.053	0.081	0.061
ϕ_1	-0.036	0.481	-0.071	0.362	0.130	0.351
ϕ_2	-0.341	1.112	0.229	0.803	0.366	0.671

the naive CMLE using completely observed data in presence of missingness in covariates. In addition, the proposed methods appear to be fairly robust under modest misspecification in the missing exposure distribution. Our approaches do suffer under the incorrect model for informative missingness mechanism. In other more extensive simulation studies under more dramatic departures from the exposure model, we noticed that the ECM approach is less robust than the FB (results are not included). Among the three methods, the FB method has the smallest MSE by virtue of introducing shrinkage effect through prior information. Regarding the secondary model parameters corresponding to the selection probability and the exposure distribution, namely, $\boldsymbol{\delta}$ and $\boldsymbol{\kappa}$, ECM and the FB provide roughly unbiased estimates except for severe model misspecification (MM5 or situation (c)). Finally, we will also like to point out that the computational costs for the ECM are substantially less than the FB in terms of computing time.

3.6 Discussion

This article presents a comprehensive approach to handle non-ignorable missingness in covariates under the stereotype regression model. Though we focus on matched case-control studies with finer disease sub-classification as our primary example, the methods can be adapted to prospective analysis of categorical response data with ordered or unordered response categories using the stereotype class of link functions. We develop an expectation/conditional maximization algorithm as well as a full Bayes procedure with data augmentation and compare these approaches with naive use of conditional maximum likelihood based on complete data. Our real data analysis as well as simulation study establish the methods lead to substantial gain in efficiency compared to the CMLE and are fairly robust under modest departures from the model for missing exposure. However the methods could perform poorly if the selection probability model is grossly misspecified.

Inference under the stereotype model is burdened with computational and analytical challenges due to embedded non-linearity and lack of identifiability in the parametric structure. Missingness further compounds the complexity. The Bayesian paradigm offers flexible alternative modeling approaches and inferential solutions for this class of models. For matched case-control data, the model has an added distinction of accommodating highly stratified data via conditioning and preserving prospective-retrospective conversion of the parameters of interest. The current Chapter involves handling a general form of missingness in this class of models. Future research involves considering a more flexible semi-parametric model for the exposure distribution, for the missingness mechanism and considering missingness with correlated or clustered observations as in a longitudinal cohort study under the stereotype link function.

CHAPTER IV

Bayesian Modeling of Studies of Gene-Environment Interaction under Two-phase Sampling

4.1 Introduction

Case-control studies are popular analytical tools, particularly in cancer epidemiology, for assessing gene-disease association where the allele/genotype frequencies at a bi-allelic single nucleotide polymorphism (SNP) locus are compared between cases and controls. Recent genomewide case-control association studies (GWAS) have been remarkably successful in identifying susceptibility loci for many cancers (Yeager et al., 2007; Hunter et al., 2007; Amundadottir, 2009). A large fraction of variability in the different cancer traits still remain unexplained, with the identified SNPs contributing modestly to prediction of disease risk (Wacholder et al., 2010; Park et al., 2010). It is thus natural to study the genetic architecture of a cancer phenotype in conjunction with the known environmental risk factors (environmental toxins, dietary exposures, physical activity levels, medication use, and other behavioral risk factors). In the post-GWAS era, more efficient statistical approaches to characterize such complex gene-environment (G x E) interactions, in terms of both design and analytic tools, have become a pressing need in cancer epidemiology research.

Variants of the case-control sampling design have been often employed in epidemiologic studies. Two-phase stratified sampling (Neyman, 1938) is an efficient alternative to the traditional cohort and case-control designs (Cochran, 1963) from cost and resource-saving perspectives. A typical application of two-phase sampling is for collecting expensive covariate information, for example, novel biomarkers or genotype data on a prioritized sub-sample of the initial study base. In particular, we will consider the following set-up: the disease outcome D, some relatively inexpensive covariates ($\boldsymbol{S}$) and environmental data (E) are collected at phase I (P_1). At phase II (P_2), genotype data (G) is collected on a subset selected from phase I sample. To select this phase II sub-sample, stratified sampling with strata defined by phase I data (D, E and possibly $\boldsymbol{S}$) is implemented.

There is a large amount of literature on two-phase designs, using different likelihood based approaches (Horvitz and Thompson, 1952; Flanders and Greenland, 1991; Breslow and Cain, 1988) or estimating score approaches (Reilly and Pepe, 1995; Chatterjee et al., 2003; Robins et al., 1994). Maximum likelihood inference for such problems was considered in the pioneering work of Scott and Wild (1997) and Breslow and Holubkov (1997a, b). Lawless et al. (1999) and Breslow and Chatterjee (1999) compare and contrast several different approaches for analyzing two-phase data. It has been noted that adding more phases can lead to further efficiency gains, consequently, the two-phase design has been generalized to multi-phase designs (Whittemore and Halpern, 1997; Lee et al., 2010). Haneuse and Chen (2011) propose an intermediate phase between phase I and phase II to reduce participation bias caused by differential participation.

The potential for such sampling designs for G x E studies has been indicated in Thomas (2010). Many GWAS adopt this sampling at the design phase, but little attention is paid at the analysis stage to address the sampling design, thus potentially leading to biased estimates. To the best of our knowledge, literature on two-phase studies of G x E interaction is very limited. Chatterjee and Chen (2007) proposed maximum likelihood inference using a novel regression model for G x E interaction studies where second stage sampling was carried out based on disease outcome and family history. Asymptotic theories were established under the assumption of independence of the genetic and environmental factors in the population.

Multiple papers (Piegorsch et al., 1994; Umbach and Weinberg, 1997; Chatterjee and Carroll, 2005) attest the phenomenon of gaining efficiency in studies of G x E by exploiting independence between the genetic and environmental factors under case-control sampling. Under such constraints, it is beneficial to use the retrospective likelihood for estimating interaction parameters instead of standard prospective logistic regression. However, with departures from these constraints, biases in estimating the interaction parameter can occur under retrospective methods. Several researchers have addressed this issue and proposed more robust strategies for testing one SNP at a time G x E interaction (Mukherjee et al., 2008, 2010; Mukherjee and Chatterjee, 2008; Vansteelandt et al., 2008; Li and Conti, 2009; Murcray et al., 2009). There is no literature on handling multiple genetic markers for G x G and G x E studies even under case-control sampling, that uses gene-gene and gene-environment independence on multiple SNP x E interaction parameters.

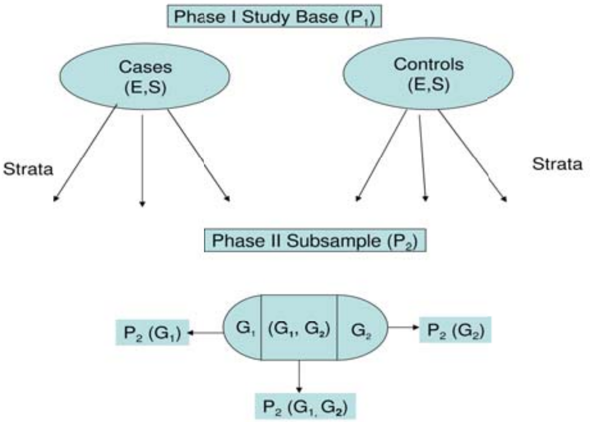
Bayesian literature on two-phase studies, even beyond the context of G x E studies is also very limited. Haneuse and Wakefield (2007) presented the first hierarchical Bayesian work that closely relate to such data structure. The Bayesian framework appears to be a natural route to explore in the current problem for multiple reasons. First, Bayesian estimation can lead to efficient computational algorithms as the two-phase likelihood is naturally a missing data likelihood. Second, for G x E studies, Bayesian methods provide data-adaptive shrinkage to leverage the constraints of gene-environment independence by imposing informative priors around this assumption. Third, we incorporate Bayesian variable selection features which help us to handle a potentially high dimensional disease risk model with main effects and interactions of multiple genes and environmental factors simultaneously. Fourth, we use the clever non-parametric Bayesian construction of Dunson and Xing (2009) as a substitute for profile likelihood in the frequentist setting to construct the retrospective likelihood under two-phase sampling. The current Chapter thus contributes to analysis of G x E studies with multiple markers/environmental exposures under an outcome-exposure stratified two-phase sampling design by offering a new Bayesian treatment of the problem.

This Chapter is largely motivated by an example that originates from a population based case-control study of colorectal cancer (CRC) in Israel, namely, the Molecular Epidemiology of Colorectal Cancer (MECC) study. Statins (our environmental factor E) are a class of lipid-lowering drugs used by more than 25 million individuals worldwide for reducing cardiovascular disease risk. The MECC study was the first to establish a chemoprotective association of statins with risk of CRC (Poynter et al., 2005). Follow-up individual studies and a meta analysis of 18 studies have confirmed this association (Hachem et al., 2009). The benefit of statins for reducing CRC risk has been shown to vary with genetic variations in HMGCR (3-Hydroxy-3-methylglutaryl coenzyme A reductase) gene, a gene involved in cholesterol synthesis (Lipkin et al., 2010). To understand the mechanism of effect modification further, investigators measured 294 SNPs in 40 genes, including HMGCR (our set of genetic factors $\boldsymbol{G}$), selected in the cholesterol synthesis/lipid metabolism pathway. The sub-sample selected for genotyping from the study population of all cases and controls was chosen by stratified sampling conditional on statin use (E) and case-control status (D) where statin users were purposefully oversampled. This sampling strategy was adopted due to limited budgetary resources and DNA samples. Complete statin use (E) data and other basic demographic covariates ($\boldsymbol{S}$) were available on the entire study base (phase I or P_1), and genetic data on these 294 SNPs were only available for the phase II subsample (P_2).

In the MECC study, due to experimental and laboratory logistics, genotype data were missing on a subset of individuals selected in P_2 on a group of genes ($\boldsymbol{G}_1$, say) and on a different subset of individuals on another group of genes ($\boldsymbol{G}_2$, say). This led to a non-monotone missing data structure with some individuals in P_2 having observations on both

$(\boldsymbol{G}_1, \boldsymbol{G}_2)$ (subset denoted by $P_2(\boldsymbol{G}_1, \boldsymbol{G}_2)$) and some only on $\boldsymbol{G}_1$ (subset denoted by $P_2(\boldsymbol{G}_1)$) and some only on $\boldsymbol{G}_2$ (subset denoted by $P_2(\boldsymbol{G}_2)$). Figure 4.1 is a visual representation of the sampling scheme and missingness pattern in the data. We would like to point out that beyond this particular example, when two cohorts are combined for genetic analysis, which is routinely the case for G x E studies, very similar missing data patterns may occur.

Figure 4.1: Data Structure under two-phase sampling and missingness pattern at phase II genetic covariate in the Molecular Epidemiology of Colorectal Cancer study



The rest of the Chapter is organized as follows. In Section 4.2, we present the model ingredients: the likelihood, priors and posteriors. In Section 4.3, we discuss the analysis of statin x gene interaction in the MECC study. In Section 4.4, we conduct a simulation study to compare the various maximum likelihood and score based approaches with the Bayesian approach. Section 4.5 concludes with a discussion.

4.2 Proposed Methods

4.2.1 The likelihood

We refer to Figure 4.1 for understanding the data structure and construction of our likelihood. Let u and D denote the subject ID and disease indicator respectively, with $W = (E, \boldsymbol{S})$. Here, E is environmental exposure and $\boldsymbol{S}$ are basic demographic covariates as described before. There are N individuals in phase I and M individuals in phase II. To simplify notations we write the retrospective likelihood corresponding to a two-gene model (G_1, G_2), with the understanding that the methods/notations can be directly extended to gene-sets $(\boldsymbol{G}_1, \boldsymbol{G}_2)$ where each contain multiple SNPs. The two-phase likelihood has the following form to capture the sampling phases and the missingness patterns in $\boldsymbol{G}$ (Figure 4.1),

$$\begin{aligned} \mathrm{L}^{\mathrm{TP}} &= \prod_{u \in P_1 \setminus P_2} \mathrm{P}(W_u|D_u) \times \prod_{u \in P_2(G_1)} \mathrm{P}(G_{1u}, W_u|D_u) \\ &\times \prod_{u \in P_2(G_2)} \mathrm{P}(G_{2u}, W_u|D_u) \times \prod_{u \in P_2(G_1, G_2)} \mathrm{P}(G_{1u}, G_{2u}, W_u|D_u). \end{aligned}$$

Each term in L^{TP} can be factorized by using $\mathrm{P}(G_1, G_2, W|D) = \{\mathrm{P}(D|G_1, G_2, W)\mathrm{P}(G_1, G_2|W)$ $\mathrm{P}(W)\}/\mathrm{P}(D)$. This retrospective likelihood is then marginalized over the missing data in each component. We assume missing completely at random (Little and Rubin, 2002) herein. The likelihood is then expressed as,

$$\begin{aligned} \mathrm{L}^{\mathrm{TP}} &= \prod_{u \in P_1 \setminus P_2} \sum_{g_1, g_2} \mathrm{P}(D_u|g_1, g_2, W_u)\mathrm{P}(g_1, g_2|W_u)\mathrm{P}(W_u)/\mathrm{P}(D_u) \\ &\times \prod_{u \in P_2(G_1)} \sum_{g_2} \mathrm{P}(D_u|G_{1u}, g_2, W_u)\mathrm{P}(G_{1u}, g_2|W_u)\mathrm{P}(W_u)/\mathrm{P}(D_u) \\ &\times \prod_{u \in P_2(G_2)} \sum_{g_1} \mathrm{P}(D_u|g_1, G_{2u}, W_u)\mathrm{P}(g_1, G_{2u}|W_u)\mathrm{P}(W_u)/\mathrm{P}(D_u) \\ &\times \prod_{u \in P_2(G_1, G_2)} \mathrm{P}(D_u|G_{1u}, G_{2u}, W_u)\mathrm{P}(G_{1u}, G_{2u}|W_u)\mathrm{P}(W_u)/\mathrm{P}(D_u), \end{aligned} \tag{4.1}$$

where $\mathrm{P}(D_u) = \sum_{g_1, g_2} \int_w \mathrm{P}(D_u|g_1, g_2, w)\mathrm{P}(g_1, g_2|w)\mathrm{P}(dw)$ with the integral replaced by the sum when components of W are discrete. Corresponding to this likelihood, there are three model ingredients:

1. A DISEASE RISK MODEL. We assume, $\mathrm{P}(D = 1|G_1 = g_1, G_2 = g_2, W = w; \boldsymbol{\beta}) = H[\{\beta_0 + m(g_1, g_2, w; \boldsymbol{\beta})\}]$, where H is the logistic function $H(u) = \{1 + \exp(-u)\}^{-1}$. Typical choice of m involves, say for two genes G_1 and G_2, $m(g_1, g_2, w; \boldsymbol{\beta}) = \beta_{G_1} g_1 + \beta_{G_2} g_2 + \beta_E e + \boldsymbol{\beta}_S^\top \boldsymbol{s} + \beta_{G_1 G_2} g_1 g_2 + \beta_{G_1 E} g_1 e + \beta_{G_2 E} g_2 e$, noting that $w = (e, \boldsymbol{s})$.

2. A model for $(G_1, G_2|W = (E, \boldsymbol{S}))$. For genotype data at a bi-allelic locus, G_j can take three possible values ('g_0=aa', 'g_1=Aa' and 'g_2=AA'). We assume, $\mathrm{P}(G_1 = g_j, G_2 = g'_j|W = w; \boldsymbol{\lambda}) = q_{jj'}(w; \boldsymbol{\lambda}), j, j' = 0, 1, 2$. This specification will require a joint model for multivariate categorical data (trinary for SNP data at a bi-allelic locus). Note that under gene-gene and gene-environment independence, the model can in general be factorized conditional on covariates $\boldsymbol{S}$,

$$\underbrace{\mathrm{P}(G_1 = g_j, G_2 = g'_j|E = e, \boldsymbol{S} = \boldsymbol{s}; \boldsymbol{\lambda}) = q^1_j(\boldsymbol{\lambda}_1|\boldsymbol{S})q^2_{j'}(\boldsymbol{\lambda}_2|\boldsymbol{S})}_{\text{under G−G and G−E independence}}, \qquad \text{for } j, j' = 0, 1, 2.$$

Instead of the above fully non-parametric model, we explore a parametric model for the joint distribution $P(\boldsymbol{G}_1, \boldsymbol{G}_2|W)$. We consider a class of log-linear models with linear by linear structure (Agresti, 2002) for parsimonious modeling of the $(G_1, G_2|W)$ associations,

$$\begin{aligned}
&\log\{\mu(G_1 = g_j, G_2 = g'_j|E = e, \boldsymbol{S} = \boldsymbol{s}; \boldsymbol{\lambda})\} \\
&\quad = \lambda_0 + \lambda_{G_1} g_j + \lambda_{G_2} g'_j + \lambda_E e + \boldsymbol{\lambda}_S^\top \boldsymbol{s} \\
&\quad +\lambda_{G_1G_2} g_j g_{j'} + \lambda_{G_1E} g_j e + \lambda_{G_2E} g_{j'} e + \boldsymbol{\lambda}^\top_{G_1S} g_j \boldsymbol{s} + \boldsymbol{\lambda}^\top_{G_2S} g_{j'} \boldsymbol{s},
\end{aligned} \tag{4.2}$$

where g_j are chosen ordinal scores, typically 0, 1, 2 (Agresti, 2002). This is the common allelic dosage coding under a log-additive genetic susceptibility model. In case of high dimensional $\boldsymbol{G}$, we can further reduce the dimensionality of the problem by assuming common association parameters λ_{GE} and λ_{GS} between similar functional groups of SNPs. As discussed in Agresti (2002), this Poisson log-linear model has a corresponding multinomial representation. Thus, the probability of $P_{G_1,G_2}(g_j, g'_j|\boldsymbol{\lambda}) = P(G_1 = g_j, G_2 = g'_j|E = e, \boldsymbol{S} = \boldsymbol{s})$ can be written in terms of the multinomial probabilities,

$$P_{G_1,G_2}(g_j, g'_j|\boldsymbol{\lambda}) = \frac{\exp(\lambda_{G_1} g_j + \lambda_{G_2} g'_j + \lambda_{G_1G_2} g_j g_{j'} + \lambda_{G_1E} g_j e + \lambda_{G_2E} g'_j e + \boldsymbol{\lambda}^\top_{G_1S} g_j \boldsymbol{s} + \boldsymbol{\lambda}^\top_{G_2S} g'_j \boldsymbol{s})}{\sum_{l=0}^2 \sum_{l'=0}^2 \exp(\lambda_{G_1} g_l + \lambda_{G_2} g'_l + \lambda_{G_1G_2} g_l g'_{l'} + \lambda_{G_1E} g_l e + \lambda_{G_2E} g'_{l'} e + \boldsymbol{\lambda}^\top_{G_1S} g_l \boldsymbol{s} + \boldsymbol{\lambda}^\top_{G_2S} g'_{l'} \boldsymbol{s})}.$$

Note that, gene-gene and gene-environment independence in the above model (2.2) will imply $\lambda_{G_1E} = \lambda_{G_2E} = \lambda_{G_1G_2} = 0$.

3. A MODEL FOR $W = (E, \boldsymbol{S})$. A non-parametric and flexible model for the distribution of W is desired. Recall that W can be a mixed set of quantitative and categorical variables. For the MECC example W is a set of categorical covariates, which will be our primary focus in this Chapter. In Remark 1, we indicate a modeling for the most general case as a kernel-mixture extension of Dunson and Xing (2009) (DX from now on), as presented in Bhattacharya and Dunson (2011). We note that the approach for modeling the joint distribution of a set of categorical variables can be applied to the the joint distribution of the trinary genotype variables $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ in (4.2) as well, but it is not clear how to enforce the gene-gene and gene-environment independence assumptions through direct priors on parameters $\lambda_{G_1E}, \lambda_{G_2E}, \lambda_{G_1G_2}$ as in the log-linear model (4.2). This is the primary reason for using (4.2) for the second component $P(G_1, G_2|W = (E, \boldsymbol{S}))$.

Let $\boldsymbol{W}_u = (E_u, \boldsymbol{S}_u)$, denote the W data corresponding to subject u, $u = 1, \ldots, N$. Here W_u is $p \times 1$ vector of p categorical variables, i.e. $W_u = (w_{u1}, \ldots, w_{up})$ for a subject u. Assume that the j-th component of W can have d_j values $j = 1, \cdots, p$. In order to parsimoniously model this $(d_1 \times d_2 \times \cdots \times d_p)$ joint distribution, DX first note that the joint distribution of two categorical variables can always be expressed as a finite mixture of product-multinomial distributions. Extending this idea DX introduce a latent class index variable $z_u \in \{1, \ldots, k\}$, such that $w_{ur}, w_{ut}, r, t \in \{1, \ldots, p\}, r \neq t$, are conditionally independent given z_u. Then the joint distribution for $\boldsymbol{w}_u$ has this finite mixture representation,

$$\begin{aligned} P_W(w_{u1} = c_1, \cdots, w_{up} = c_p) &= \sum_{h=1}^{k} P(w_{u1} = c_1, \cdots, w_{up} = c_p | z_u = h) P(z_u = h) \\ &= \sum_{h=1}^{k} P(z_u = h) \prod_{j=1}^{p} P(w_{uj} = c_j | z_u = h). \end{aligned} \tag{4.3}$$

For notational convenience, we rewrite (4.3) as

$$\begin{aligned} P_W(w_{u1} = c_1, \cdots, w_{up} = c_p) &= \pi_{c_1 \cdots c_p} \\ &= \sum_{h=1}^{k} \nu_h \prod_{j=1}^{p} \psi_{hc_j}^{(j)}, \qquad \sum_{c_1=1}^{d_1} \cdots \sum_{c_p=1}^{d_p} \pi_{c_1 \cdots c_p} = 1, \end{aligned} \tag{4.4}$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)^\top$ is a probability vector with $\nu_h = P(z_u = h)$ and $\psi_{hc_j}^{(j)} = P(w_{uj} = c_j | z_u = h)$, is a $d_j \times 1$ probability vector i.e., the conditional probability of $w_{uj} = c_j$, given that subject u is in latent class h for $j = 1, \ldots, p$. We will discuss the choice of k through a Dirichlet process prior structure on this latent class probability model in the next section.

Remark 1: While Chatterjee and Chen (2007), Chatterjee and Carroll (2005) use profile likelihood for handling the distribution of W non-parametrically, it has been a challenging task in the Bayesian framework to posit a flexible model for $\boldsymbol{W} = (E, \boldsymbol{S})$ which could be a mixture of categorical and continuous covariates. In this mixed case, Müller et al. (1999) model the joint distribution of the continuous covariates through a Dirichlet Process mixture of normals. Then, conditional on the continuous covariates, the categorical variables have a joint multivariate probit distribution. However, implementation and estimation is cumbersome under this formulation for a large number of categorical predictors and the construction seems artificial without a latent continuous score motivation for the multivariate probit model. A recent paper by Bhattacharya and Dunson (2011) extends the above DX construction for categorical data to handle joint distribution modeling of more complex data, including continuous and discrete data. They extend the conditional independence idea and replace the product-multinomial structure in (4.4) by a product of various kernels, such as Gaussian, Poisson and more complex univariate or multivariate distributional kernel. An additional feature of their paper is a factor analytic representation of the joint distribution that helps with further reducing the dimensionality. The MECC example does not require going beyond the original DX construction, but with continuous E, this is what we would adopt.

4.2.2 Priors

As mentioned before, for this complex retrospective likelihood formulation, we have three sets of parameters from the above three ingredients. For $\boldsymbol{\beta}$ in the disease risk model, we use a Spike and Slab type mixture prior to handle variable selection in a high-dimensional disease risk model with multiple markers. For $\boldsymbol{\lambda}$ in the multivariate gene model, the Bayesian hierarchical approach provides a flexible way to allow for uncertainty around the assumption of gene-gene and gene-environment independence, through prior on $\lambda_{G_1G_2}, \lambda_{G_1E}$, and λ_{G_2E}. When sparsity occurs in a certain configuration of $(\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{W})$ or dimension of $(\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{W})$ grows, the frequentist profile likelihood estimation may become unstable and the log-linear model with shared parameters across gene-sets and the DX latent mixture construction aids with such situations. The presence of missing data, potential sparsity in $\boldsymbol{G}$, $\boldsymbol{W}$, the two-phase data likelihood itself being a missing data likelihood, all make the problem naturally amenable to a unified Bayesian computational treatment. We follow the same sequence to describe the prior structure on the parameters involved in the three ingredients of the likelihood as in the previous section.

1. In the presence of multiple genes in $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$, the logistic disease risk model can potentially have many pairwise and higher order interaction terms. We implement a scalable variable selection framework via Spike and Slab type priors (Mitchell and Beaucamp, 1988; George and McCullogh, 1993) on the parameters $\boldsymbol{\beta}$ in the disease risk model $P(D|\boldsymbol{G}_1, \boldsymbol{G}_2, W; \boldsymbol{\beta})$. We impose mixture prior distributions on each component of $\boldsymbol{\beta}$, say, $(\beta_0, \beta_{G_1}, \beta_{G_2}, \beta_E, \beta_S, \beta_{G_1G_2}, \beta_{G_1E}, \beta_{G_2E})$ for a two-gene model. In general we denote this vector by $\boldsymbol{\beta}_{n_\beta \times 1} = \{\beta_r, r = 1, \ldots, n_\beta\}$. Given a latent variable p_0 representing the mixture

weight on the null regression coefficients, we describe priors as below.

$$\begin{aligned}
\beta_r|f_r,\tau_r &\overset{ind}{\sim} N(0, f_r{\tau_r}^2), \qquad r=1,\ldots,n_\beta \\
f_r|v_0,p_0 &\overset{iid}{\sim} p_0\delta_{v_0}(\cdot)+(1-p_0)\delta_1(\cdot), \\
\tau_r^{-2}|a_1,a_2 &\overset{iid}{\sim} Gamma(a_1,a_2), \\
p_0 &\overset{iid}{\sim} Beta(a,b).
\end{aligned}$$

As discussed in Ishwaran and Rao (2003), v_0 in the above specification is assumed to be a small positive value near 0. This value plays a key role in shrinking the posterior of β_r towards zero when the r-th covariate turns out to be insignificant. A Dirac mass on 1, namely, δ_1 represents non-zero β_r corresponding to a significant/selected covariate. Choosing the tuning parameter v_0 corresponding to $\{f_r\}$ can be done in a data-adaptive way as described through the above hierarchy. For this particular application, we fix $(a_1,a_2,a,b)=(5,50,1,1)$ for hyperparameters in the above prior specification. Instead of componentwise shrinkage, one can also offer a more flexible multivariate shrinkage by using a nonparametric mixture of a point mass at the zero vector (or at a pre-specified vector v_0) and a random probability measure.

2. In the joint log-linear model in (4.2), we typically assume vague normal priors with large variance on the parameters $(\lambda_{G_1},\lambda_{G_2},\lambda_{G_1S},\ \lambda_{G_2S})$. In our data example, we have used a $N(0,10^4)$ prior. On the other hand, for the *G*-*E* pairwise association parameters $(\lambda_{G_1G_2},\lambda_{G_1E},\lambda_{G_2E})$ we reflect a priori information on *G*-*G* or *G*-*E* independence via a normal prior centered at zero but with two different choices for the prior variance. In the first set of priors we reflect the belief that with 95% probability the association parameter lies between log(0.8) and log(1.2). This leads to an approximate SD=0.1 under a normal distribution and thus we assume an informative prior of $N(0,10^{-2})$. In the second choice, following the empirical Bayes estimation of Mukherjee and Chatterjee (2008), we compute association

parameters for G_1-G_2, G_1-E, and G_2-E in the control subjects in the data, say $\hat{\theta}$, and use a data-driven prior $N(0, \hat{\theta}^2)$ on $\lambda_{G_1G_2}$, λ_{G_1E}, and λ_{G_2E}. Possibilities for data-adaptive multivariate shrinkage around the independence assumption can be employed here as well through mixture priors with correlated shrinkage weights.

3. The mixture representation in (4.4) requires determining the number of latent classes k. Following DX, instead of selecting a fixed k, a Bayesian nonparametric approach is carried out through the following Dirichlet process prior specification on $\boldsymbol{\nu}$:

$$
\begin{aligned}
\boldsymbol{\pi} &= \sum_{h=1}^{\infty} \nu_h \boldsymbol{\psi}_h, \qquad \boldsymbol{\psi}_h = \boldsymbol{\psi}_h^{(1)} \otimes \cdots \otimes \boldsymbol{\psi}_h^{(p)}, \qquad h = 1, \ldots, \infty, \\
\boldsymbol{\psi}_h^{(j)} &\sim Dirichlet(a_{j1}, \ldots, a_{jd_j}), \text{ independently for } j = 1, \ldots, p, \\
\nu_h &= \sum_{h=1}^{\infty} V_h \prod_{l<h} (1 - V_l), \qquad V_h \sim Beta(1, \alpha), \\
\alpha &\sim Gamma(a_\alpha, b_\alpha).
\end{aligned}
$$

where $\otimes$ is the outer product. The parameter α is a hyper-parameter that controls the rate of decrease from the stick-breaking process (Sethuraman, 1994). For example, in case of small values of α, ν_h decreases towards zero quickly with increasing h, thus putting most of the weight on first few components, leading to a sparse representation. The hyperprior on α allows one to data-adaptively determine the degree of sparseness or the number of components needed. As discussed in DX (2009), we set $(a_\alpha, b_\alpha) = (1/4, 1/4)$ for a vague prior which implies the probability of the independence assumption in the product multinomial model to be 0.5. We set uniform priors for each category probability $\boldsymbol{\psi}$ by specifying $a_{j1} = \cdots = a_{jd_j} = 1$, for $j = 1, \ldots, p$.

4.2.3 Posterior sampling

In the full likelihood (4.1), we would like to point out that the three components are linked with each other through the sum over each component in the expression for $P(D)$ in the denominator. We denote the two-phase likelihood in (4.1) by L_{TP} which involves the parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{V}, \alpha)$. The full conditionals are not reducible to a simpler closed form and are best represented by the following proportionality relations:

$$\begin{aligned}
\beta_r|\cdot &\propto \mathrm{L}^{\mathrm{TP}} \times \exp(-\frac{{\beta_r}^2}{2f_r{\tau_r}^2}), \qquad r = 1, \ldots, n_\beta, \\
{\tau_r}^{-2}|\cdot &\propto \mathrm{Gamma}(a_1 + 0.5, a_2 + \frac{{\beta_r}^2}{2f_r}), \\
f_r|\cdot &\propto \{I(f_r = v_0)p_0 + I(f_r = 1)(1 - p_0)\} \times \exp(-\frac{1}{2f_r{\tau_r}^2}{\beta_r}^2) \times {f_r}^{-0.5}, \\
p_0|\cdot &\propto \mathrm{Beta}(a + \sum_{r=1}^{n_\beta} I(f_r = v_0), b + \sum_{r=1}^{n_\beta} I(f_r = 1)), \\
\lambda_l|\cdot &\propto \mathrm{L}^{\mathrm{TP}} \times \exp(-\frac{\lambda_l^2}{2\sigma^2}), \qquad l = 1, \ldots, n_\lambda,
\end{aligned}$$

where n_β and n_λ again represent the number of parameters in $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ respectively.

Posterior sampling corresponding to $P(\boldsymbol{W})$:

Let us recapitulate the model structure for $\boldsymbol{W}$ which is essentially a Dirichlet process mixture of discrete Dirichlet kernels. For $u = 1, \cdots, N$ and $j = 1, \cdots, p$,

$$\begin{aligned}
w_{uj} &\sim \mathit{Multinomial}(\{1, \ldots, d_j\}, \psi^j_{z_u,1}, \ldots, \psi^j_{z_u,d_j}), \\
z_u &\sim V_h \prod_{l<h}(1 - V_l)\delta_h, \qquad V_h \sim \mathit{Beta}(1, \alpha), \qquad \alpha \sim \mathit{Gamma}(a_\alpha, b_\alpha).
\end{aligned}$$

DX present an efficient data-augmented Gibbs sampling algorithm by augmenting the likelihood with latent constructs following Walker (2007). The details of the updating steps are contained in Appendix.

Note that while the entire likelihood in DX constituted of W data only, in our problem, $P(\boldsymbol{W})$ is embedded as a component in the joint retrospective likelihood L^{TP} in (4.1). Thus for updating the parameters involved in $P(\boldsymbol{W})$, say $\boldsymbol{\theta}(= \{\boldsymbol{\psi}, \boldsymbol{V}, \alpha\})$, we use the Metropolis Hastings algorithm. Note that only the terms $\prod_u \mathrm{P}(\boldsymbol{W}_u)/\mathrm{P}(D_u)$ from the full likelihood (4.1) involves $\boldsymbol{\theta}$, where $\mathrm{P}(D_u) = \sum_{g_1,g_2} \sum_{\boldsymbol{w}} \mathrm{P}(D_u|g_1, g_2, \boldsymbol{w})\ \mathrm{P}(g_1, g_2|\boldsymbol{w})\mathrm{P}(\boldsymbol{w})$ in the presence of only categorical $\boldsymbol{W}$. We draw $\boldsymbol{\theta}$ following the DX algorithm and for proposal density for $\boldsymbol{\theta}$. Here we consider the implied full conditional $q(\boldsymbol{\theta}^{new}|\boldsymbol{W})$ under this algorithm. Then given $\boldsymbol{\lambda}, \boldsymbol{\beta}$ we repeat the following updates of $\boldsymbol{\theta}$.

- At iteration l, sample a vector $\boldsymbol{\theta}^{new}$ from $q(\boldsymbol{\theta}^{new}|\boldsymbol{W})$ as described in DX (2009) algorithm.
- Compute the acceptance ratio
$$r(\boldsymbol{\theta}^{new}, \boldsymbol{\theta}_l) = \min[1, \frac{\prod_u \mathrm{P}(D_u|\boldsymbol{\theta}_l, \boldsymbol{\lambda}, \boldsymbol{\beta})}{\prod_u \mathrm{P}(D_u|\boldsymbol{\theta}^{new}, \boldsymbol{\lambda}, \boldsymbol{\beta})}].$$
In calculating the acceptance ratio, we note that the numerator and denominator $\prod_u\{\mathrm{P}(W_u|\boldsymbol{\theta}^{new})\}p(\boldsymbol{\theta}^{new})\, q(\boldsymbol{\theta}_l|\boldsymbol{W})/\prod_u\{\mathrm{P}(W_u|\boldsymbol{\theta}_l)\}p(\boldsymbol{\theta}_l)q(\boldsymbol{\theta}^{new}|\boldsymbol{W})$ is canceled out where $p(\boldsymbol{\theta})$ is a prior for $\boldsymbol{\theta}$.
- If $r(\boldsymbol{\theta}^{new}, \boldsymbol{\theta}_l) < U$ where $U \sim unif(0,1)$, we set $\boldsymbol{\theta}_{l+1} = \boldsymbol{\theta}^{new}$. Otherwise, the candidate vector θ^{new} is rejected and $\boldsymbol{\theta}_{l+1} = \boldsymbol{\theta}_l$.
- Repeat the steps until the posterior chains converge to the stationary distributions.

Given the full conditionals, we implement the Gibbs sampler (Geman and Geman, 1984) with Metropolis Hastings updates to sample from respective full conditional distributions. For each parameter, we iterate 50,000 times and discard the first 40,000 iterations as 'burn-in'. We check convergence of the chains using traceplots and the diagnostic statistics 'potential scale reduction factor' (Gelman and Rubin, 1992) using the R package CODA (Plummer et al., 2009). Auto and cross-correlation checks are performed and a thinning of every tenth observation is carried out. Remaining posterior samples are used to construct estimated posterior summaries needed for Bayesian inference.

4.3 The Molecular Epidemiology of Colorectal Cancer Study

In this section, we describe the motivating example from the MECC study in detail and present analysis results. We use data on 1,745 cases and 1,852 controls with completely observed response to the question whether statins were used for more than 5 years. The binary variable 'statin use of at least 5 years' (E), is the environmental factor of interest with 91% "NO" and 9% "YES".

We consider completely observed confounders and precision variables ($\boldsymbol{S}$): age (S_1), gender (S_2), ethnicity (S_3), physical activity (S_4), family history of CRC (S_5), vegetable consumption (S_6), NSAID usage within 3 year (S_7), and Aspirin usage within 3 year (S_8). Age and ethnicity variables were dichotomized as Age $\geq$ or $<$ 50 (94% and 6% respectively), and 'Ashkenazi' and 'Non-Ashkenazi' (68% and 32% respectively). Gender (S_3) was coded as 1 (50%) for male and 0 (50%) for female. The remaining binary factors (S_4, S_5, S_6, S_7, S_8) are classified to 1 or "YES" with the proportions of (0.36, 0.09,0.31, 0.02, 0.20) respectively.

For genotyping at phase II, stratified-sampling based on the disease status (D) and statin use (E) was carried out. All case-control subjects with statin use ("YES") were included at phase II sample. We have 1,200 cases and 1,200 controls at phase II with data available on 294 trinary SNPs $\boldsymbol{G} = (G_1, \ldots, G_{294})$. Genotype data are not completely observed even at phase II due to technical genotyping failures for a limited number of SNPs. Among 2,400 case-control subjects at phase II, 27 subjects and 20 subjects have no genotype information in G_1 and G_2 respectively. We did not have many markers across the genome to successfully impute these missing genotypes, thus we consider a marginalized likelihood as in (4.1).

Among 294 SNPs, we first illustrate our methods with two SNPs on two genes, $RS762551$ on CYP1A2 (G_1) and $RS1056836$ on CYP1B1 (G_2) where both SNPs exhibit significant interactions with statins in a preliminary one at a time, single marker interaction analysis. We illustrate our methods for this simple model as some of our competing methods can only handle single marker interaction analysis. No departure from the Hardy-Weinberg equilibrium was noted ($p = 0.38$ and 0.73 respectively). The raw frequencies of the cross-classification of case-control status (D), statins (E), genotypes G_1 and G_2 are shown in Appendix Table

A.7. Simple logistic regression analysis was carried out to examine G_1-E and G_2-E association among control subjects and yielded odds ratios of 0.78 and 0.94 and corresponding p-values of 0.03 and 0.53 respectively. G_1-G_2 association reveals no dependence (p-value of 0.83) based on Chi-squared test for independence. This implies that the data is suggestive of G_1-E association whereas little evidence for G_2-E or G_1-G_2 association is noted.

We report the results of this analysis in Table 4.1. Along with two-phase full Bayes approach (TPFB) we consider five alternative methods. Unfortunately, none of these existing methods use the data in both phases and make use of the independence constraints. The first three use phase II data only (i) Unconstrained maximum likelihood (UML), a retrospective analysis that does not specify any constraints on $P(G_1, G_2|W)$, (ii) Constrained maximum likelihood (CML), that imposes the Hardy-Weinberg Equilibrium as well as G_1-E/G_1-G_2 independence, (iii) Empirical-Bayes (EB), using data-adaptive 'shrinkage estimation' between the constrained and unconstrained ML estimates. Since methods (ii) and (iii) are developed for single marker analysis, G_2-E independence cannot be enforced in existing software (we used the 'CGEN' package by Bhattacharjee, Chatterjee, and Wheeler, 2011). The above methods completely ignore biased sampling at phase II and may thus lead to overestimation of the main effect of E, especially if differential sampling was carried out in cases and controls at phase II. The next two approaches use information from both phases under a prospective likelihood framework: (iv) a Horvitz-Thompson estimator, typically known as a weighted likelihood (WL) approach (Manski and Lerman, 1977; Breslow and Chatterjee, 1999). This approach uses sampling fractions n_{ij}/N_{ij}, where n_{ij} and N_{ij} are the number of subjects corresponding to $D = i, E = j$ at phase II and phase I respectively. The sampling fraction serves as weights in the likelihood to adjust for biased sampling (we used the *svyglm* function in 'survey' package in R by Lumley, 2011). Finally (v) a pseudo-likelihood (PL) approach which also adjusts for biased sampling probabilities in a likelihood framework (Schill et al., 1993). Briefly, if we denote $P_{ij} = P(D = i|E = j) = \exp(i\alpha_j)/\{1 + \exp(\alpha_j)\}$ where α_j is the log-odds for $D = 1$ when $E = j$, then pseudo-likelihood is defined as $\prod_{i,j} P_{ij}^{N_{ij}} \prod_{i,j,k} p_{ijk}$. Here,

$$p_{ijk} = \frac{n_{ij} \exp\{i(\beta_0 - \alpha_j + s_{ijk}\beta)\}}{n_{0j} + n_{1j} \exp(\beta_0 - \alpha_j + s_{ijk}\beta)}.$$

where s_{ijk} is a covariate for a subject with $D = i$ and $E = j$.

Note that all of these five methods use completely observed phase II data on G_1 and G_2 as opposed to our proposed method that includes partially observed data by marginalization of the likelihood in terms of G_1 and G_2 when needed.

As previously explained, we present our method (TPFB) corresponding to two different priors on the G-E and G-G association parameters in model (4.2). First, we consider informative prior $N(0, 10^{-2})$ that enforces prior belief on independence assumption, we denote this by TPFB. The analysis using an alternative prior where the variance is estimated based on observed association in the data is denoted by TPFB$_{\text{emp}}$. In Table 4.1, variable selection scheme is excluded in the TPFB and TPFB$_{\text{emp}}$ by assuming all $f_r = 1, r = 1, \ldots, n_\beta$ so that all covariates are included across all methods.

Under all methods, note in Table 4.1 that the estimated coefficients corresponding to statin-use suggests strong negative association with CRC status. The estimated effect size varies depending on whether the method accounts for biased sampling and/or gene-environment independence. Note that, in presence of interactions, we cannot really interpret the main effect estimates and need to combine the model results to present estimated subgroup effects. Recall that G_1-E independence does not appear to be supported in the light of this data, thus the CML approach and TPFB yield numerically different estimates of G_1 x E interaction when compared to other methods. For G_2 x E and G_1 x G_2 interaction, the estimates are fairly comparable across methods. Smaller standard errors corresponding to interaction parameters are noted in retrospective methods that explicitly model (G_1, G_2, E)

Table 4.1: Analysis results for the MECC study data with statins (E), G_1 $RS762551$ on CYP1A2 and G_2 $RS1056836$ on CYP1B1. The set of risk factors included are: use of statins (E, 'at least 5 years'=1, 'o.w.'=0), age (S_1, 'over 50'=1, 'o.w.'=0), gender (S_2, male=1, female=0), ethnicity (S_3, Ashkenazi=0, Non-Ashkenazi=1), sports activity (S_4, Yes=1, No=0), vegetable consumption (S_5, High=1, Low=0), family history of CRC (S_6, Yes=1, No=0), the use or non-use of NSAID within 3 years(S_7, Yes=1, No=0), the use or non-use of Aspirin within 3 years (S_8, Yes=1, No=0). Under the TPFB method the 'est.' corresponds to the posterior mean whereas PSD corresponds to posterior standard deviation.

	TPFB$_{\text{emp}}$	TPFB	WL	PL	UML	CML	EB
	est.(PSD)	est.(PSD)	est.(se)	est.(se)	est.(se)	est.(se)	est.(se)
Demographic variables							
Age (over 50)	0.00 (.13)	0.01 (.13)	0.08 (.19)	0.08 (.18)	0.08 (.18)	0.06 (.18)	0.06 (.18)
Gender - Male	0.13 (.06)	0.13 (.06)	0.27 (.09)	0.27 (.09)	0.27 (.09)	0.26 (.09)	0.26 (.09)
Ethnicity	-0.33 (.08)	-0.33 (.08)	-0.34 (.09)	-0.34 (.09)	-0.34 (.09)	-0.34 (.09)	-0.34 (.09)
Exposure variables							
G_1	-0.13 (.09)	-0.10 (.09)	-0.13 (.11)	-0.11 (.09)	-0.14 (.11)	-0.05 (.08)	-0.08 (.10)
G_2	-0.09 (.08)	-0.08 (.08)	-0.10 (.09)	-0.04 (.09)	-0.11 (.09)	-0.09 (.08)	-0.09 (.08)
Statin use	-1.55 (.22)	-1.47 (.21)	-1.66 (.27)	-1.65 (.27)	-1.76 (.27)	-1.55 (.26)	-1.63 (.27)
Sports activity	-0.51 (.07)	-0.51 (.07)	-0.53 (.09)	-0.52 (.09)	-0.52 (.09)	-0.51 (.09)	-0.51 (.09)
Family history of CRC	-0.31 (.10)	-0.31 (.10)	-0.59 (.15)	-0.58 (.16)	-0.58 (.15)	-0.58 (.15)	-0.58 (.15)
Vegetable consumption	-0.22 (.06)	-0.22 (.07)	-0.19 (.09)	-0.19 (.09)	-0.19 (.09)	-0.19 (.09)	-0.19 (.09)
NSAID use	-0.64 (.24)	-0.65 (.24)	-0.58 (.32)	-0.58 (.32)	-0.58(.32)	-0.52 (.31)	-0.52 (.31)
Aspirin use	-0.51 (.10)	-0.51 (.10)	-0.47 (.12)	-0.49 (.11)	-0.49 (.11)	-0.48 (.11)	-0.48 (.11)
G_1 x G_2	-0.01 (.07)	-0.02 (.08)	-0.05 (.09)	-0.14 (.11)	-0.04 (.09)	-0.05 (.06)	-0.06 (.06)
G_1 x Statin use	0.50 (.16)	0.42 (.15)	0.65 (.20)	0.65 (.20)	0.65 (.20)	0.33 (.15)	0.56 (.21)
G_2 x Statin use	0.44 (.15)	0.41 (.14)	0.53 (.18)	0.53 (.19)	0.52 (.19)	0.51 (.18)	0.51 (.18)
Gene-Statin association parameters from $P(G_1, G_2 \vert E, \boldsymbol{S})$							
$\lambda_{G_1 G_2}$	-0.02 (.04)	-0.01 (.04)					
$\lambda_{G_1 E}$	-0.16 (.09)	-0.08 (.06)					
$\lambda_{G_2 E}$	-0.04 (.08)	-0.02 (.07)					

†TPFB, TPFB$_{\text{emp}}$: Two-phase full Bayes (with empirical estimates for prior variances), UML: Unconstrained maximum likelihood, CML: Constrained maximum likelihood, EB: Empirical-Bayes, WL: weighted likelihood, and PL: pseudo-likelihood

dependence structure. The TPFB$_{\text{emp}}$ and TPFB generally present smaller standard errors for the interaction parameter estimates. In general, the results from TPFB approaches are numerically slightly different than other methods as no other method uses the data and assumptions simultaneously as the TPFB method does. All methods suggest evidence in favor of G_1 x E and G_2 x E interaction being present.

To reflect our main interest in sub-group effects of statin across genotype configurations, we report effects of statin across genotype sub-groups of one SNP, holding the other SNP fixed at the common genotype category for that second SNP (coded as 0). It seems that statin effect is strongly modified by genotype of $RS762551$ (G_1). According to TPFB$_{\text{emp}}$ estimates, keeping G_2 genotype fixed at C/C, the benefit of taking statins to reduce the risk of CRC is maximum in the A/A genotype of G_1 with the posterior estimate (and 95% HPD) of the odds-ratios (relative to controls) being 0.21 (0.14, 0.32). The corresponding ORs in genotype category A/C and C/C are 0.35 (0.23, 0.53) and 0.58 (0.32, 1.04) respectively. Figure 4.2 illustrates estimated posterior densities of the odds ratios corresponding to statin-use across each genotype of G_1 (left) or G_2 (right) respectively, while holding the other SNP fixed at the most common category. This figure indicates that the protective effect of statin in CRC are diminishing as the allelic dosage for the minor allele increases in both G_1 and G_2.

Table 4.2: Odds ratio estimates for CRC corresponding to statin users vs non-users across genotype sub-groups. Under all five methods, a model with main effect of G_1, G_2, E controlling for S was fit as in Table 4.1. Common allele in G_1 ($RS762551$ on CYP1A2) and G_2 ($RS1056836$ on CYP1B1) are A and C respectively and minor allele in G_1 and G_2 are C and G respectively.

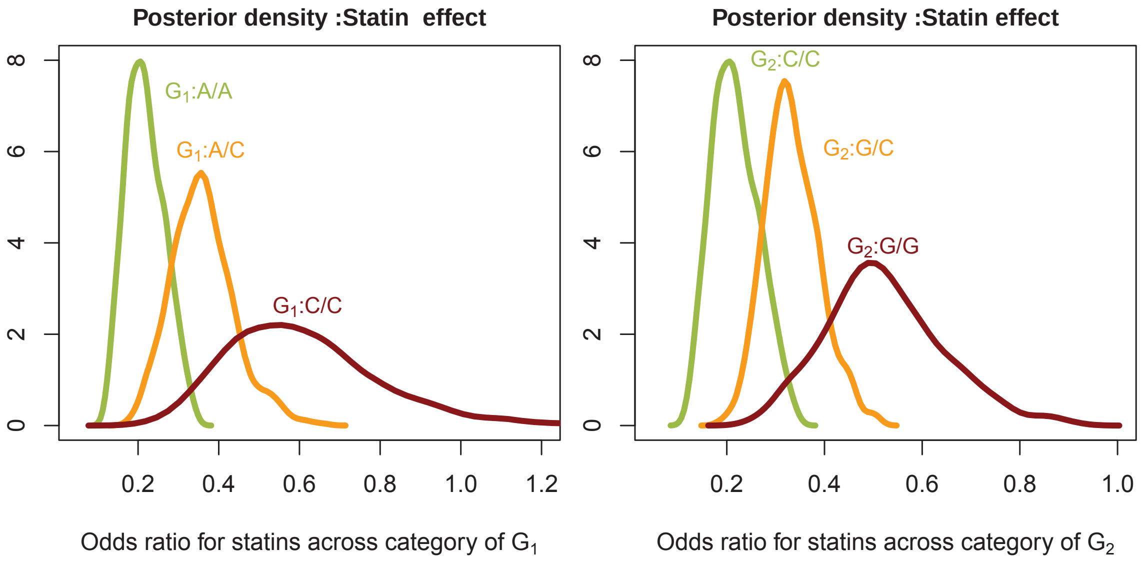
	Statins	Statins	Statins	Statins	Statins
G_1	A/A	A/C	C/C	A/A	A/A
G_2	C/C	C/C	C/C	G/C	G/G
TPFB_{emp}	0.21 (0.14, 0.32)	0.35 (0.23, 0.53)	0.58 (0.32, 1.04)	0.33 (0.24,0.46)	0.51 (0.32, 0.78)
TPFB	0.23 (0.16,0.34)	0.35 (0.24, 0.51)	0.53 (0.33, 0.94)	0.35 (0.26,0.50)	0.53 (0.34, 0.81)
WL	0.19 (0.11, 0.32)	0.37 (0.23,0.58)	0.70 (0.36, 1.37)	0.32 (0.25, 0.53)	0.54 (0.43, 1.13)
PL	0.19 (0.11, 0.33)	0.37 (0.23,0.58)	0.71 (0.36, 1.38)	0.32 (0.25, 0.53)	0.55 (0.42, 1.17)
UML	0.17 (0.10, 0.29)	0.33 (0.21, 0.53)	0.63 (0.32, 1.24)	0.29 (0.20, 0.42)	0.49 (0.29, 0.82)
CML	0.21 (0.13, 0.35)	0.30 (0.19, 0.47)	0.41 (0.23, 0.74)	0.35 (0.25, 0.50)	0.59 (0.36, 0.97)
EB	0.20 (0.12, 0.33)	0.35 (0.21, 0.56)	0.61 (0.29, 1.25)	0.33 (0.22, 0.48)	0.54 (0.32, 0.92)

†TPFB, TPFB_{emp}: Two-phase full Bayes (with empirical estimates for prior variances), UML: Unconstrained maximum likelihood, CML: Constrained maximum likelihood, EB: Empirical-Bayes, WL: weighted likelihood, and PL: pseudo-likelihood

Variable Selection: We explore how variable selection feature performs in this example in TPFB method. Previous research by Ishwaran and Rao (2003) discussed the performance of Spike and Slab prior for variable selection in detail, but not for this particular scenario. We introduce three SNPs (RS5925224, RS10174721, RS1616524) and all possible pairwise G x G and G x E interactions to the previous two SNP model as fit in Table 4.1. The dimension of the disease risk model is 34. None of the main effects and interactions corresponding to these three additional SNPs were significant in a single marker analysis.

We set $f_r = 1$ for S_1 through S_8 to always keep the confounders and precision variables in the model. The tuning parameters v_0 is fixed at 0.0001 for this application. We would like to see if the variable selection can still detect the two significant interactions (G_1 x E, G_2 x E) and recognize the unimportant SNPs and interactions. We tabulate the posterior

Figure 4.2: The left figure shows the posterior densities of the odds ratio estimates of CRC corresponding to statin users versus non-users across three genotypes in RS762551 of CYP1A2(G_1), holding the genotype in RS1056836 of CYP1B1 at the most frequent category, i.e., $(G_2) = (C/C)$. Similarly, The right figure shows the posterior densities of the odds ratio estimates corresponding to statin users versus non-users across three genotypes in RS1056836 of CYP1B1(G_2), holding the genotype in RS762551 of CYP1B1 fixed at the most frequent category, i.e., $(G_1) = (A/A)$



distribution of $\boldsymbol{f} = (f_1, \ldots, f_{n_\beta})$ which indicate 'in-and-out' frequencies of the corresponding parameters. These posterior frequencies of $\boldsymbol{f}$ can be used to define a ranking of important predictors. An alternative is to rank the top models (not just the predictors individually). Before implementing the TPFB, we reduced the dimensionality of parameters in the model $P(\boldsymbol{G}|\boldsymbol{W})$ where $\boldsymbol{G} = (G_1, G_2, G_3, G_4, G_5)$ by assuming common λ_{GG} and λ_{GE} association parameters across all SNPs. We use $N(0, 0.1^2)$ prior on this common parameter. In addition, we further assume a single common parameter λ_{GS} for all G-S associations with a vague normal prior $N(0, 10^4)$.

In Table 4.3, we present numerical results on model and predictor ranking as well as the Bayesian Information Criterion (BIC) corresponding to each model. We only present the top 10 models. According to the result, the model with main effects of E and G_1 x E and G_2 x E interactions seems to be the preferred model (posterior probability 16.7%) followed by the model with E and only G_1 x E interaction (posterior probability 15.9%). Table 4.4 shows

Table 4.3: The top 10 promising models in terms of estimated posterior probabilities of the models. All $\boldsymbol{S}$ adjustment variables are retained in the model and variable selection is performed only on the five genetic and environmental factors and all possible pairwise interactions. Bayesian Information Criterion (BIC) is provided for each model.

Model	posterior probability %	BIC
$[E]$[All S][G_1 x E][G_2 x E]	16.7 %	48747
$[E]$[All S][G_1 x E]	15.9 %	48741
$[E]$[All S]	10.5%	48736
$[E]$[All S][G_2 x E]	5.5 %	48742
$[E]$[All S][G_1 x E][G_2 x E][G_5 x E]	2.6%	48755
$[E]$[All S][G_1 x E][G_5 x E]	2.0 %	48747
$[E]$[All S][G_3][G_1 x E][G_2 x E]	1.6 %	48753
$[E]$[All S][G_3][G_1 x E]	1.3 %	48749
$[E]$[All S][G_1 x E][G_3 x E]	1.2 %	48750
$[E]$[All S][G_3][G_5 x E]	1.2 %	48742

†BIC represents Bayesian Information Criterion

Table 4.4: The estimated posterior probabilities of appearance corresponding to $\boldsymbol{G}$ and $\boldsymbol{E}$ main effects and their interactions are shown under the identical setting as in Table 4.3

Covariates	G_1	G_2	G_3	G_4	G_5	E	E x G_1	E x G_2	E x G_3	E x G_4	E x G_5
Freq.	3.9	4.1	9.5	5.1	3.3	99.2	68.4	48.3	6.8	5.3	13.8

the frequency of retaining a predictor in the model according to the posterior distribution of $\boldsymbol{f}$. The main effect of E appears most of the times (99.2%) with large selection probabilities for G_1 x E and G_2 x E interactions (68.4% and 48.3%) respectively. Overall, non-significant interactions/main effects are well filtered under this variable selection scheme.

4.4 Simulation study

In this section, we assess the performance of the proposed method by conducting a simulation study. We mainly consider two aspects (i) varying gene-gene/gene-environment association structure and (ii) when phase II sampling is differential between cases and controls. We compare our method with the five alternative methods mentioned before: WL, PL, UML, CML, and EB in terms of the average bias and mean squared errors (MSE), based on 200 simulated datasets.

We first describe the data generation procedure. We consider two genes G_1 and G_2, and, one environment factor E, with disease status D, all binary. We generate data from the following log-linear model (Li and Conti, 2009):

$$
\begin{aligned}
\log(\mu|D, G_1, G_2, E) &= \gamma_0 + \gamma_{G_1}G_1 + \gamma_{G_2}G_2 + \gamma_E E + \gamma_D D \\
&+ \lambda_{G_1E}G_1E + \lambda_{G_2E}G_2E + \lambda_{G_1G_2}G_1G_2 \\
&+ \beta_{G_1}G_1D + \beta_{G_2}G_2D + \beta_E ED \\
&+ \beta_{G_1E}G_1ED + \beta_{G_2E}G_2ED + \beta_{G_1G_2}G_1G_2D, \qquad (4.5)
\end{aligned}
$$

where μ denotes expected cell counts corresponding to the (D, G_1, G_2, E) configuration. Under this model, we are capable of manipulating G_1-E, G_2-E, and G_1-G_2 association under controls by adjusting $\lambda_{G_1E}, \lambda_{G_2E}$, and $\lambda_{G_1G_2}$ respectively where these parameters are approximately equivalent to those in model $P(G_1, G_2|W)$ (4.2) when the disease is rare. Similarly, we can set $\beta_{G_1E}, \beta_{G_2E}$ or $\beta_{G_1G_2}$, corresponding to the G x E or G x G interactions in the disease risk model. The parameters $(\gamma_0, \gamma_{G_1}, \gamma_{G_2}, \gamma_E)$ controls the marginal frequencies of G_1, G_2 and E in controls. Note that γ_D needs to be set at a large negative value for the disease to be rare.

For the model parameters in (4.5), we fixed $(\gamma_0, \gamma_{G_1}, \gamma_{G_2}, \gamma_E, \gamma_D) = (-6, -0.5, -0.5, -2.0, -4.5)$ that produces approximately 2.5% of cases, frequency of $G_1 = 1$ and $G_2 = 1$ both at 45% while the prevalence of $E = 1$ is 15%. We assign $(\beta_{G_1E}, \beta_{G_2E}, \beta_{G_1G_2}) = (0, \log(2), \log(2))$ in (4.5). For setting parameters corresponding to G-E/G-G association, we set $(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (\log(2), 0, \log(1.5))$ to reflect G_1-G_2 and G_2-E dependence, and $(0, 0, 0)$ for the independence scenario.

Now we turn our attention to the sampling design. We randomly generate $1,000$ cases and $1,000$ controls with complete (D, G_1, G_2, E) data. We then carry out (D, E)-stratified sampling as follows. We select 600 cases and 600 controls in phase II. We consider two scenarios regarding this the stratified sampling stratgey: (a) all subjects with a positive $E(= 1)$, in cases and controls, are automatically included in phase II; (b) all subjects with a positive E $(= 1)$ in cases are included in phase II, however, 600 controls for phase II are randomly selected regardless of E status. Finally, information on G_1 and G_2 from phase I subjects, that is, 400 cases and 400 controls, is treated as missing. We iterate this step to generate 200 replicate datasets under each sampling scheme.

Table 4.5 displays the simulation results. We follow the convention that $\perp$ and $\sim$ represent the independence and dependence respectively. Under $G_1 \perp E, G_2 \perp E$, and $G_1 \perp G_2$ the CML method yields the smallest MSE with respect to G_1 x E and G_1 x G_2 interaction followed by TPFB, TPFB$_{\text{emp}}$, and EB while WL, PL, and UML present relatively larger MSE. Here we need to note that the current implementation of CML and EB can only use G_1-E and G_1-G_2 independence, but not G_2-E independence. As phase II sampling becomes differential between cases and controls from scenario (a) to (b), we notice the substantial increase in the bias for estimating the main effect of E from CML, UML, and EB as expected while WL, PL, TPFB, and TPFB$_{\text{emp}}$ provide relatively less biased estimates. This trend remains present in case where $G_1 \perp E, G_2 \sim E$, and $G_1 \sim G_2$. Beyond the bias in E estimation from CML, UML, and EB, we note that under the departure from the independence assumption, namely, $G_1 \perp G_2$, there is a dramatic increase in the bias corresponding to the G_1 x G_2 interaction under CML and to some extent in TPFB. TPFB$_{\text{emp}}$ and EB are more robust to this assumption. Both TPFB show gain in efficiency for interaction estimation compared to PL and WL. Overall, our proposed methods, especially TPFB$_{\text{emp}}$, yield obvious gain in efficiency compared to PL and WL in terms of the G x E or G x G interactions in the presence of independence. On the other hand, TPFB$_{\text{emp}}$ provides less biased estimates of the E effect compared to UML, CML, and EB where the bias introduced by ignoring the two-phase design is inevitable. When sub-sampling ratio is 80%, the pattern remains same as seen in Appendix Table A.8.

Table 4.5: Simulation results under two scenarios 1) $G_1 \perp E$, $G_1 \perp G_2$, and $G_2 \perp E$ association, 2) $G_1 \perp E$, $G_1 \sim G_2$, and $G_2 \sim E$ with different strategies for stratified sampling. The results are based on 200 replicated datasets, each with 1,000 cases and 1,000 controls in phase I and 600 cases and 600 controls in phase II. The approaches listed, TPFB, TPFB$_{\text{emp}}$, WL, PL, UML, CML, and EB where each represents Two-phase full Bayes (with empirically obtained prior variance), Weighted likelihood, Pseudolikeliohod, Unconstrained Maximum likelihood, Constrained Maximum Likelihood, and Empirical Bayes respectively. The CML imposes G_1-E and G_1-G_2 independence, however, no constraints on G_2-E association. We set $(\beta_E, \beta_{G_1G_2}, \beta_{G_1E}, \beta_{G_2E}) = (-1.5, 0, \log(2), \log(2))$ for all scenarios. In terms of MCMC chains from both TPFBs, the posterior results based on 1,000 samples after 'burn-in' of 40,000 among 50,000 iterations and a thinning of every 10 samples.

		$G_1 \perp E$, $G_1 \perp G_2$, $G_2 \perp E$				$G_1 \perp E$, $G_1 \sim G_2$, $G_2 \sim E$			
Stratified sampling (a)†		E	G_1 x G_2	G_1 x E	G_2 x E	E	G_1 x G_2	G_1 x E	G_2 x E
		$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (0, 0, 0)$				$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (\log(2), 0, \log(1.5))$			
TPFB	Bias	-0.049	0.004	-0.015	-0.034	-0.119	0.288	-0.041	0.194
	MSE	0.189	0.074	0.182	0.214	0.178	0.173	0.162	0.222
TPFB$_{\text{emp}}$	Bias	0.026	0.029	-0.114	-0.097	0.005	0.092	-0.086	0.047
	MSE	0.182	0.058	0.251	0.179	0.203	0.103	0.150	0.223
WL	Bias	-0.081	-0.003	0.042	0.036	-0.060	0.030	0.007	0.059
	MSE	0.223	0.115	0.279	0.294	0.184	0.130	0.202	0.269
PL	Bias	-0.082	-0.006	0.043	0.037	-0.062	0.022	0.008	0.060
	MSE	0.223	0.112	0.279	0.294	0.184	0.130	0.202	0.269
UML	Bias	-0.138	-0.006	0.043	0.037	-0.120	0.022	0.008	0.060
	MSE	0.244	0.112	0.279	0.294	0.201	0.130	0.202	0.269
CML	Bias	-0.137	-0.022	0.035	0.033	-0.133	0.707	0.036	0.054
	MSE	0.221	0.054	0.170	0.287	0.195	0.557	0.126	0.258
EB	Bias	-0.136	-0.020	0.034	0.034	-0.131	0.157	0.025	0.055
	MSE	0.222	0.068	0.201	0.287	0.194	0.163	0.144	0.258
		$G_1 \perp E$, $G_1 \perp G_2$, $G_2 \perp E$				$G_1 \perp E$, $G_1 \sim G_2$, $G_2 \sim E$			
Stratified sampling (b)‡		E	G_1 x G_2	G_1 x E	G_2 x E	E	G_1 x G_2	G_1 x E	G_2 x E
		$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (0, 0, 0)$				$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (\log(2), 0, \log(1.5))$			
TPFB	Bias	-0.032	-0.014	-0.031	-0.020	-0.173	0.308	0.036	0.220
	MSE	0.165	0.098	0.261	0.210	0.249	0.205	0.222	0.252
TPFB$_{\text{emp}}$	Bias	0.014	-0.016	-0.091	-0.073	-0.041	0.044	0.078	0.021
	MSE	0.173	0.066	0.294	0.238	0.194	0.089	0.249	0.211
WL	Bias	-0.051	0.006	0.070	0.077	-0.079	0.009	0.127	-0.011
	MSE	0.458	0.153	0.577	0.445	0.287	0.137	0.467	0.386
PL	Bias	-0.076	0.003	0.074	0.079	-0.060	0.017	0.138	-0.031
	MSE	0.261	0.150	0.554	0.422	0.268	0.134	0.449	0.384
UML	Bias	0.492	0.003	0.074	0.079	0.507	0.017	0.138	-0.031
	MSE	0.534	0.150	0.554	0.422	0.550	0.134	0.449	0.384
CML	Bias	0.488	-0.002	0.018	0.085	0.508	0.691	0.079	-0.018
	MSE	0.483	0.060	0.232	0.386	0.516	0.530	0.157	0.352
EB	Bias	0.490	0.002	0.050	0.083	0.503	0.171	0.121	-0.024
	MSE	0.491	0.087	0.355	0.388	0.516	0.167	0.276	0.356

†All subjects with $E = 1$ in case and control are sub-sampled for phase II.

‡All cases with $E = 1$ are included in phase II, however, control are randomly selected for phase II.

TPFB uses the informative prior $N(0, 10^{-2})$ on G-G and G-E associations in the model (4.2)

TPFB$_{\text{emp}}$ uses the prior $N(0, \hat{\theta}^2)$ on G-G and G-E associations in the model (4.2) where $\hat{\theta}^2$ is empirically estimated G-G or G-E association parameter under controls.

4.5 Discussion

We presented a flexible Bayesian approach to estimate gene-gene (G x G) and/or gene-environment (G x E) interactions under two-phase sampling. The proposed approach can handle multiple genetic and environmental factors. The method can trade off between bias and efficiency by incorporating uncertainty around gene-environment independence through the hierarchical structure in a data-adaptive way. The underlying ingredients of this hierarchy are the disease risk model, the multivariate gene model, and the joint model for the environment factors/covariates respectively. Our method can also handle potential missingness in genetic information due to technical inconsistency, or due to merging different studies or cohorts, leading to non-monotone missing data structure at phase II sub-sample.

We compared our method to simpler alternatives such as UML, CML, and EB that use gene-environment independence but only based on phase II data, ignoring biased sampling. We also considered methods that account for biased sampling at phase II: weighted likelihood and pseudo likelihood, but do not leverage the independence assumption. Our method provides a framework that integrates both of these features. In a clinical study like the MECC example, where interest lies in estimating the differential effect of statin use across genetic sub-groups for devising targeted prevention strategies, estimates of main effects as well as gene-environment interaction are equally important, thus both estimates need to be assessed. This Chapter is the first Bayesian paper with retrospective modeling for G x E studies under two-phase sampling that can handle multiple markers.

There are some limitations of the current Chapter that need to be expanded and explored in future studies. First, we do not fully address the performance of our method in the presence of a truly high-dimensional gene model through simulation studies. The method is scalable to handle up to 294 SNPs and pairwise interactions in our data example, but we have not carried out a simulation study due to computation time. We also need to deal with exponentially increasing number of G x E and G x G interactions in the disease risk model as well as G-E/G-G/G-S associations in the multivariate gene model, as we add more G-variables in the model. We address this by Bayesian variable selection and assuming a common parameter for G-E/G-G/G-S association on genes in the same pathway in the multivariate gene model. The latter is a rather ad-hoc strategy for reducing the dimension. Bias in parameter estimates is expected to arise under departures from this assumption. Calculation of $P(D)$ in the denominator of the likelihood could also pose challenges with truly high-dimensional data. Second, we have not tested the Dunson and Bhattacharya (2011) algorithm for mixed set of discrete and continuous covariates in W. Future research will focus on the higher-dimensional G and E settings, more general structure of the W vector as well as possibility of capturing higher order interactions, not just pairwise interactions.

CHAPTER V

A Spatio-Temporal Point Process Model for Analyzing Diarrheal Case Patterns under a serial Case-Control study

5.1 Introduction

Diarrhea is one of the leading causes of pediatric death. According to the World Health Organization (http://www.who.int/mediacentre/factsheets/fs330/en/index.html), diarrheal deaths exceed the combined death toll due to AIDs, tuberculosis, and malaria, largely due to the high death rate in developing countries. Compared to approximately 2.5 million deaths each year in developing countries (Kosek *et al.*, 2003), hundreds of millions of diarrheal cases and thousands of resulting deaths are reported annually in developed countries such as the United States (Herikstad *et al.*, 2002). Previous epidemiological studies (Curtis and Cairncross, 2003; Checkley *et al.*, 2004; Barreto *et al.*, 2007) attest that this high prevalence is largely attributable to individual risks factors such as individual hygiene, food contamination, and socio-economic status as well as community-associated factors such as inferior water quality and sanitation systems. Although early work has identified a variety of risk factors for diarrhea, the debate continues on the limitations of previous studies. Eisenberg *et al.* (2006) argued that these risks do not reflect changes in community or ecological determinants. Accordingly, current epidemiologic research is moving towards a system-based approach to understand community level factors, household factors, and individual level factors that may underlie the biological or social causes of diarrhea.

The ECODESS (Ecologia, Desarrollo, Salud, y Sociedad) study is designed to further the understanding of the underlying causal process of diarrheal prevalence and transmission rates involving social and ecological factors such as road construction, social networks, sanitation, and other confounding factors (http://www.sph.umich.edu/scr/ecodess/home.php). Eisenberg *et al.* (2006) selected 21 communities in the Esmeraldas province of Ecuador for this study. Their work suggests an association between remoteness from the most populated city, Borbón, and all-causes of diarrhea as well as an association between diarrhea and three pathogens (E. coli, Giardia, Rota virus). They argued that new road construction results in deforestation, changes in sanitation, hygiene, and changes in social life, all of which contribute to disease transmission. In a follow-up study, Bates *et al.* (2007) investigated the role of social networks on disease transmission and found an association between social networks and disease prevalence. These findings are consistent with Gushulak and MacPherson (2004) who showed that remote communities, known to have lower immigration and emigration rates, have lower transmission rates. Levy *et al.* (2009) focused on the impact of seasonal changes in water quality induced by rainfall on E. coli and discovered a negative association between water quality and E. coli counts.

With advances in Geographical Information System (GIS), collection of spatially referenced data is becoming more prevalent. Accordingly, spatial inference ranging from association studies between geographically distant covariates and outcomes to the interpolation or prediction of unobserved values at desired locations is of substantial interest. For example, spatial research in various fields such as marketing, ecology, and environment have received much recent attention (Choi *et al.*, 2008; Gelfand and Barber, 2007; Cowles and Zimmerman, 2003).

The ECODESS study is unique as it includes longitudinal and spatial data on the sampled communities. The data include complete enumeration and locations of disease cases, spatially and temporally referenced covariates like remoteness and social networks, and temporal covariates such as temperature and precipitation. These features provide researchers an opportunity to investigate spatial heterogeneity and spatial covariates that may affect diarrheal prevalence.

Since the seminal papers on stationary point processes (Ripley, 1976, 1977), there has been a substantial amount of research on point process models, as well as their applications. Noting that homogeneous point process models are too simple to capture spatial inhomogeneity, Diggle and Elliot (1995) proposed an inhomogeneous Poisson process to account for non-uniformly distributed point patterns of specific diseases under the assumption that the disease is intrinsically non-infectious. More complex approaches became available along with the developments in computational techniques. Heikkinen and Arjas (1998) presented a non-parametric Bayesian model of the intensity function of a spatial Poisson process based on a step function approximation using a Markov random field prior. Baddely *et al.* (2000) discussed semi-parametric and non-parametric estimation of spatial interactions under an inhomogeneous point process. Møller and Waagepetersen (2003) integrated state-of-the-art spatial point process models in their book and provided mathematical theories and examples of miscellaneous applications. Hossain and Lawson (2009) compared commonly used Bayesian point process models in the presence of a putative hazard source. Several R (R Development Core Team, 2011) packages have been written that estimate various parameters of rather simple spatial point process models. Two of the more useful packages are Spatstat (Baddeley and Turner, 2005) and DCluster (Gómez-Rubio *et al.*, 2010).

In infectious diseases, clustering or aggregation of cases typically occurs. Waller and Gotway (2004) demonstrated that case patterns frequently display clustering in space. To deal with clustering, various cluster point process models have been proposed. These include, the Markov point process (Van Lieshout, 2000), the shot noise Cox process (SNCP) (Brix, 1999; Brix and Kendall, 2002), and the log Gaussian Cox process (LGCP) (Møller *et al.*, 1998). Among these, the LGCP is popular due to its flexibility, simplicity, and mathematical tractability. With respect to the LGCP, Waagepetersen (2004) proved that the expectation of the approximate posterior from discretized LGCPs converges to the exact posterior expectation as the cell size of the grid goes to zero. Beneš *et al.* (2005) used the LGCP to investigate the association between tick-born encephalitis and spatially varying covariates of vegetation and altitude. LGCP modeling has also been extended to the spatio-temporal setting. Brix and Diggle (2001) developed a class of space-time LGCP models where they used moment-based parameter estimation with space-time correlation structure for the intensity. Brix and Møller (2001) proposed a space-time point process based on a bivariate log Gaussian Cox birth process in modeling two types of weeds that monotonically propagate over time. Diggle *et al.* (2005) used LGCP with a specific intensity function multiplied by the background intensity obtained by a kernel intensity surface estimation. They also presented a multiplicative decomposition of the spatio-temporal intensity in an ad hoc fashion. Recently, Liang *et al.* (2009) implemented a marked LGCP for differentiating colorectal cancer types and incorporated non-spatial individual level covariates.

A popular goal in spatial inference is prediction (Gelfand *et al.*, 2001, 2003). When spatial prediction is required on a different spatial scale than the originally observed scale, spatial misalignment, also called “change of support”, typically occurs. Since Krige (1951) initially proposed, what is now referred to as, ordinary Kriging, many versions of Kriging have been introduced. Gelfand *et al.* (2001) proposed general types of Bayesian approaches to handle the spatial misalignment problem (SMP) based on a Gaussian process. These include points to points, points to blocks, blocks to points, and blocks to blocks. They also extended their approach to the spatio-temporal setting, illustrating it on ozone measurement data.

In the ECODESS study, spatially and temporally referenced covariates along with diarrheal case coordinates are available at 21 of the 158 communities in the Esmeraldas province. Data were collected approximately every twelve months for six years. The researchers wish to explain spatial and temporal variation in disease patterns within the 21 communities and predict the number of diarrheal cases at unsampled communities. To achieve both goals, we propose a Bayesian two-stage spatial point process model accounting for spatial misalignment of the measured covariates. Spatial inhomogeneity, within communities, is accounted for by a spatially varying covariate, called the social network covariate, and temporal heterogeneity is accounted for by temperature and precipitation fluctuations. A second spatially varying covariate, remoteness from Borbón, is also observed. However, this covariate is spatially misaligned with the social network covariate. Whereas the social network is measured within communities, remoteness is measured from the center of each community to Borbón. In this Chapter, we model case patterns in relation to these spatial and temporal covariates.

To the best of our knowledge, there is no literature on modeling serial case patterns involving spatially and temporally referenced covariates. Park and Kim (2004) considered case-control studies for diarrhea with longitudinal data without spatial information. Later, Diggle *et al.* (2007) proposed spatial point process modeling with two distinct intensities corresponding to cases and controls without a temporal component. Liang *et al.* (2009) also considered two cancer patterns without a longitudinal component. In this respect, our approach provides a unique aspect involving spatio-temporal case pattern data analysis as well as prediction.

This Chapter is organized as follows. In Section 5.2, we describe the motivating example in detail. In Section 5.3, we propose a Bayesian two-stage model that accounts for spatial misalignment. The model includes inference at the sampled communities, and prediction of cases at unsampled communities. Its direct application to the ECODESS study follows in Section 5.4. Results from simulation studies are given in Section 5.5. Finally, we conclude with remarks on the current work and possible extensions.

5.2 ECODESS Study : Sampling Design

The ECODESS study covers the northern coastal Ecuadorian province of Esmeraldas (Figure 5.1, left panel). In the ECODESS study, 21 communities were randomly selected among 158 communities within Esmeraldas excluding Borbón: the most populated city located at the confluence of the three rivers. Communities were selected by a block randomized design using location, size, population, and the relative distance to Borbón. Within these 21 communities, all households were enrolled in the study and 98% of the residents participated. The right panel of Figure 5.1 depicts the locations of all 158 communities relative to Borbón.

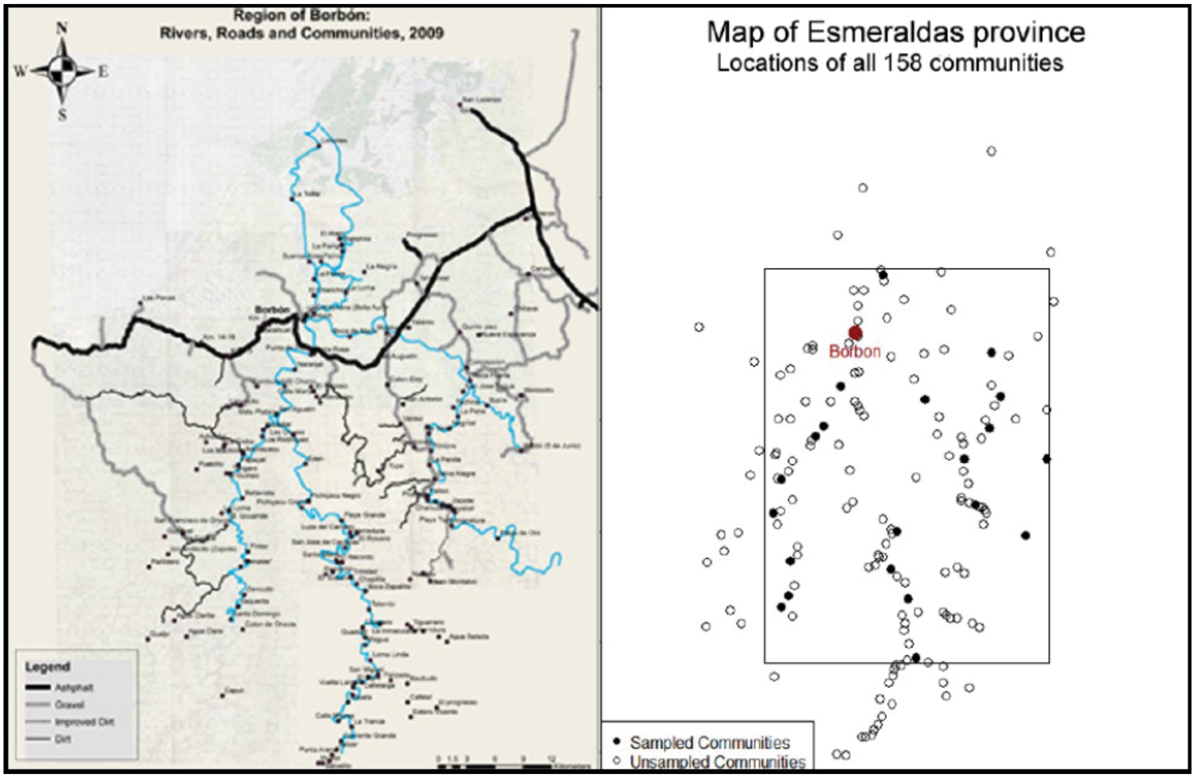


Figure 5.1: The left panel displays a map of the Esmeraldas province in Equador. The blue lines represent three main rivers; the Cayapas, the Santiago, and the Onzole. Each point on the map indicates the location of a community. The right panel shows whether a community was sampled in the ECODESS study. The 21 solid dots represent sampled communities and the remaining empty dots represent unsampled communities.

The ECODESS research team visited each sampled community annually or semi-annually, on a rotating basis from the beginning of August 2003 to March 2008, for a total of 7 cycles. Each visit lasted 15 days. The researchers interviewed each household every morning and identified all diarrheal cases. A case was defined as an individual having three or more loose stools in a 24-hour period. Cases, as well as GIS coordinates of the household, were recorded. Demographic data such as age, gender, and sanitation were also collected.

One goal of the ECODESS study was to identify the association between diseases prevalence and a remoteness metric defined by the travel time and total cost of travel to Borbón (Eisenberg *et al.*, 2006). The remoteness metric, R_c, corresponding to community $c, c = 1, \ldots, 21$, is defined as $R_c = C_c / \sum_c C_c + T_c / \sum_c T_c$ where T_c and C_c are time and cost of travel to Borbón from community c. Travel time and cost was determined only for the 21 sampled communities. There is a highly significant linear association between the remoteness metric and distance from Borbón with $R^2 = 0.83$; thus we feel that distance is a justifiable surrogate for the remoteness metric. The number of cases in each community and the corresponding the distance are summarized in Table 5.1. Figure 5.2 shows case locations at four of the seven cycles.

According to Bates *et al.* (2007), the social network covariate is defined through multiple factors obtained from sociometric surveys. In their study, spatial index, the harmonic mean of distances from one house to all other houses, shows a strong linear relationship with social network. Large values of the spatial index corresponds to low density. Bates *et al.* (2007) demonstrated that areas with lower values of spatial index, or greater density, have a greater chance of being exposed to disease transmission, which leads to increases in disease prevalence. Note that under a discretized setting as our study, this spatial index needs to be defined at each cell regardless of whether there is a house in that cell. To do this, we compute the harmonic mean of the distances between each cell and the cells containing at least one house.

Table 5.1: The number of cases and population reported in the 21 sampled communities across 7 cycles in the ECODESS study. For each community the first row is the number of cases and the second row is population. We report the remoteness metric measured in 2007 and the corresponding distance from Borbón. The last column presents grid sizes where the length of each side of a cell is approximately 11 meters.

				Cycle					Distance	
Community	1	2	3	4	5	6	7	Remoteness	(Km)	Grid size
San Agustin	9	10	8	8	9	7	7	0.012	9.27	154×135
	280	335	335	342	343	324	324			
Colon Eloy	12	6	10	19	27	16	18	0.015	12.21	92× 43
	815	884	891	868	889	950	961			
Naranjal	0	3	4	1	1	0	0	0.022	6.87	151×215
	88	88	92	86	85	88	88			
Timbire[†]	11	10	5	8	14	6	10	0.027	20.59	83 × 50
	469	528	522	544	571	596	601			
Roca Fuerte[†]	7	2	3	7	4	5	4	0.039	19.99	25×18
	156	166	165	188	179	181	172			
La Loma	16	5	4	2	2	2	2	0.040	4.74	268 × 219
	149	165	158	158	154	138	139			
Ranchito	1	1	6	0	1	0	0	0.040	7.49	100× 102
	51	68	81	65	61	65	65			
Quinto piso	3	1	4	1	3	3	5	0.049	16.51	221 × 278
	65	81	88	92	95	74	77			
La Pena	1	6	1	1	2	0	2	0.049	20.67	25× 12
	98	111	99	89	91	96	88			
Las Cruces	4	3	0	2	2	1	0	0.061	13.42	88×91
	102	114	107	98	97	124	107			
Tangare	0	0	0	2	2	2	3	0.080	20.26	100×21
	101	101	101	101	105	98	99			
El Rosario	0	2	0	3	1	1	0	0.113	24.86	43×52
	129	128	129	131	126	132	129			
Guayabal	9	2	0	2	3	1	6	0.122	25.68	14×35
	146	148	145	144	146	139	149			
Arenales	0	1	3	1	7	0	2	0.140	28.84	175×216
	137	156	163	112	126	137	126			
Wimbi	6	5	16	10	10	2	3	0.152	28.53	20×32
	321	350	369	335	342	339	344			
Playa de Oro	4	6	0	3	4	3	3	0.155	32.63	17×23
	231	253	255	257	255	265	267			
Trinidad	2	0	1	3	2	3	0	0.158	29.02	15×11
	98	105	106	102	103	117	105			
Telembi	4	8	8	14	5	5	1	0.165	32.76	43×24
	283	352	348	330	308	403	397			
Vaquerita[†]	0	1	3	2	3	0	0	0.173	33.01	15×46
	35	33	37	33	38	35	35			
Santo Domingo[†]	10	4	8	10	10	4	11	0.190	34.05	89×29
	479	505	512	486	481	465	473			
San Miguel	2	1	3	1	2	0	2	0.198	39.72	15×20
	123	148	154	142	143	145	159			
Case Total	101	77	87	100	114	61	79			
Population Total	4356	4819	4857	4703	4738	4911	4905			

[†] The communities with bold font are used for internal prediction

Given temperature and precipitation at three weather stations located in San Miguel, Borbón, and Playa de Oro, we interpolate temperature and precipitation at each of the remaining 19 sampled communities at each cycle using ordinary Kriging (R package *GSTAT*; Pebesma, 2004). Covariate information is summarized in Table 5.2.

5.3 Proposed Method

We propose a Bayesian two-stage spatio-temporal point process model that involves a community level analysis followed by prediction of the number of cases at unsampled communities.

We denote the set of the 21 sampled community study windows by $\boldsymbol{S} = \{S_c\}_{c=1}^{21}$. Each rectangular-shaped S_c is the smallest rectangle such that each side is a multiple of 11 meters and S_c contains all houses within community c. We partition each community window $S_c \in \boldsymbol{S}$ into disjoint square cells of size 11×11 meters, $S_{c,m}$, $m = 1, \ldots, K_{1c} \times K_{2c}$. Here K_{1c} and K_{2c} are determined by the area of each community. The grid size of each community is summarized in Table 5.1. Furthermore, we denote the set of disjoint study windows for unsampled communities by $\boldsymbol{U} = \{U_l\}_{l=1}^{106}$. Note that, for prediction, we excluded 31 communities located outside of the rectangle that contains the 21 sampled communities to avoid extrapolation (see the rectangle in the right panel of Figure 5.1).

We develop a two-stage model. In stage I, we adopt a log Gaussian Cox process (LGCP, refer to Appendix A.4.1) model for each sampled community. In stage II, we make use of the LGCP model results to predict the number of diarrheal cases at unsampled communities. Let $\boldsymbol{x}_{S,t} = \{\boldsymbol{x}_{S_c,t}\}_{c=1}^{21}$ denote the set of case coordinates corresponding to sampled community c at cycle t. Similarly, let $\boldsymbol{x}_{U,t} = \{\boldsymbol{x}_{U_l,t}\}_{l=1}^{106}$ denote the set of case coordinates corresponding to unsampled communities at cycle t.

Stage I: Estimation of the intensity at each sampled community.

We consider a LGCP involving community c at cycle t. Denote the intensity function by $\Lambda_{c,t}$. A pair of temporal covariates, $(Temp_{c,t}, Prec_{c,t})$, represent temperature and precipitation, respectively, at community c and cycle t. Additionally, we have a single spatially-referenced covariate, the spatial index or SI, where SI_c is SI at community c and $\boldsymbol{SI} = \{SI_c(s)\}_{c=1}^{21}$, $s \in S_c$. Covariate summaries are tabulated in Table 5.2.

The underlying Gaussian random field (GRF) for community c at cycle t is $Y_{c,t}$ with mean $\mu_{c,t} = \mu_c^{com} + \mu_t^{time}$, where μ_c^{com} is a community specific random effect that follows $N(\mu_{com}, \sigma^2_{com})$. This random effect captures the community-specific uncertainty not explained by current covariates but attributable to community-level characteristics such as average individual age and household sanitation. Furthermore, μ_t^{time}, $t = 2, \dots, 7$, ($\mu_1^{time} = 0$ for reference) is the time-specific mean offset. When σ_t^2 is the marginal variance at time t, the covariance of the GRF is $Cov(Y_{c,t}(s), Y_{c,t}(s')) = \sigma_t^2 r(s, s') = \sigma_t^2 \exp(-k \| s - s' \|^{\alpha})$ where $s, s' \in S_c$ and $\| \cdot \|$ denotes the Euclidean norm.

Here, k and α in the correlation function are estimated via the minimum contrast estimation method (Møller *et al.*, 1998) and are considered known constants, $(k, \alpha) = (0.26, 20)$, across communities and cycles in our model.

Then, the intensity process for community c at cycle t is

$$\Lambda_{c,t}(s) \quad = \quad \pi_{c,t} \exp(Y_{c,t}(s) + \eta SI_c(s) + \beta_1 Temp_{c,t} + \beta_2 Prec_{c,t}), \qquad s \in S_c,$$

where $\pi_{c,t}$ is the known population density for community c at cycle t. The vector of parameters in the first stage is denoted by $\mathbf{\Omega}_1 = (\boldsymbol{\mu}^{com}, \boldsymbol{\mu}^{time}, \mu_{com}, \sigma^2_{com}, \eta, \beta_1, \beta_2, \boldsymbol{\sigma}^2)^\top$ where $\boldsymbol{\mu}^{com} = (\mu_1^{com}, \dots, \mu_{21}^{com})$, $\boldsymbol{\mu}^{time} = (\mu_2^{time}, \dots, \mu_7^{time})$, and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_7^2)$.

Given that case coordinates $\boldsymbol{x}_{S_c,t}$ are recorded for community c at cycle t, the likelihood corresponding to cycle t is expressed as

$$(5.1) \qquad f(\boldsymbol{x}_{S,t} | \mathbf{\Omega}_1) \propto \prod_{c=1}^{21} \left[\exp\left\{ -\int_{S_c} \Lambda_{c,t}(s) ds \right\} \prod_{\xi \in \boldsymbol{x}_{S_c,t}} \Lambda_{c,t}(\xi) \right],$$

where the likelihood for all cycles is $\prod_{t=1}^{7} f(\boldsymbol{x}_{S,t} | \mathbf{\Omega}_1)$.

However, computationally we cannot work directly with the GRFs, $Y_{c,t}$, $c = 1, \dots, 21$, $t = 1, \dots, 7$, as they are infinite dimensional. Therefore, we approximate $Y_{c,t}$ through its realized value on a discretized grid, say $\widetilde{Y_{c,t}}$. Note that the approximated value is constant within each cell in S_c. Following Møller *et al.* (1998), we can expand $\widetilde{Y_{c,t}}$ on an extended grid $\boldsymbol{S}_c^{ext}$,

$$\widetilde{Y_{c,t}^{ext}} \;=\; \sigma_t^2 \Sigma_{c,t}^{ext\,1/2} \Gamma_{c,t}^{ext} + \mu_{c,t}^{ext}, \qquad \Gamma_{c,t}^{ext} \sim N_{2K_{1c}\times 2K_{2c}}(0, I),$$

where superscript ext represents values on $\boldsymbol{S}_c^{ext}$. Here, $\Sigma_{c,t}^{ext}$ is the correlation matrix, $\Gamma_{c,t}^{ext}$ is a random vector following $N(0, I)$, and $\mu_{c,t}^{ext}$ is a mean vector defined on $\boldsymbol{S}_c^{ext}$, respectively. Then, $\widetilde{Y_{c,t}}$ is obtained as the appropriate marginal distribution of $\widetilde{Y_{c,t}^{ext}}$.

Now we describe independent priors for $\boldsymbol{\Omega}_1$ and $\Gamma_{c,t}^{ext}$. With little knowledge about the parameters, we propose vague independent priors:

$$\begin{aligned}
\log \sigma_t^2 &\overset{iid}{\sim} N(0, 10^4), \qquad t = 1, \ldots, 7, \\
\mu_c^{com} &\overset{iid}{\sim} N(\mu_{com}, \sigma_{com}^2), \qquad c = 1, \ldots, 21, \qquad \mu_t^{time} \overset{iid}{\sim} N(0, 10^4), \qquad t = 2, \ldots, 7, \\
\eta &\sim N(0, 10^4), \qquad \beta_1 \sim N(0, 10^4), \qquad \beta_2 \sim N(0, 10^4), \\
\Gamma_{c,t}^{ext} &\overset{iid}{\sim} N_{2K_{1c}\times 2K_{2c}}(0, I), \qquad t = 1, \ldots, 7,\ c = 1, \ldots, 21.
\end{aligned}$$

Hyper-priors for the population mean μ_{com} and variance σ_{com}^2 are

$$\mu_{com} \;\sim\; N(0, 2 \times 10^5), \qquad \sigma_{com}^2 \sim IG(0.1, 0.1),$$

where $IG(\alpha, \beta)$ indicates the inverse gamma distribution with mean $\beta/(\alpha - 1)$. We provide details regarding full conditionals and our posterior sampling strategy for stage I parameters in Appendices A.4.2 and A.4.3.

STAGE II: PREDICTS THE MISSING NUMBER OF CASES AND ESTIMATING THE MEAN INTENSITIES AT UNSAMPLED COMMUNITIES.

We can predict the number of cases at unsampled communities under the Poisson distribution if we know the intensity and the population. The expected number of a homogeneous Poisson process is $population/area \times intensity \times area = population \times intensity$. Since the populations at unsampled communities $\boldsymbol{U}$ are known, we need only estimate the corresponding intensities.

The log mean intensities corresponding to sampled communities are determined by the posterior samples of $\mathbf{\Omega}_1$ from stage I. We denote the log mean intensity at community c and cycle t by $\overline{I^t_{Sc}(\mathbf{\Omega}_1)}$, or simply I^t_{Sc}, which is equivalent to $\log(\overline{\Lambda_{c,t}}/\pi_{c,t})$ where $\overline{\Lambda_{c,t}}$ is the posterior mean of $\Lambda_{c,t}$. Let $\boldsymbol{I}^t_S = (I^t_{S_1}, \ldots, I^t_{S_{21}})^\top$. At community c, the log mean intensity $I^t_{S_c}$ at cycle t satisfies

$$\begin{aligned} & \sum_{m=1}^{K_{1c}\times K_{2c}} \pi_{c,t}\exp(I^t_{c,m})|S_{c,m}| \\ = \;& 1/(K_{1c}\times K_{2c})|S_c| \sum_{m=1}^{K_{1c}\times K_{2c}} \pi_{c,t}\exp(I^t_{c,m}) \\ = \;& \pi_{c,t}\exp(I^t_{S_c})|S_c|, \end{aligned}$$

where $I^t_{c,m}$ represents the posterior mean log intensity in cell $S_{c,m}$ at cycle t and $|\cdot|$ represents area. Then $I^t_{S_c} = \log\{1/(K_{1c}\times K_{2c})\sum_{m=1}^{K_{1c}\times K_{2c}} \exp(I^t_{c,m})\}$.

At cycle t, let $I^t_{U_l}$ be the log mean intensity at unsampled community l and let $\boldsymbol{I}^t_U = (I^t_{U_1}, \ldots, I^t_{U_{106}})^\top$. In terms of the number of cases, $N(x_{U_l,t})$ denotes the number of cases at unsampled community l and cycle t and let $N(\boldsymbol{x}_{U,t}) = (N(x_{U_1,t}), \ldots, N(x_{U_{106},t}))^\top$. Similarly, $N(x_{S_c,t})$ denotes the number of diarrheal cases at sampled community c and cycle t. Let $N(\boldsymbol{x}_{S,t}) = (N(x_{S_1,t}), \ldots, N(x_{S_{21},t}))^\top$. We also denote a vector of hyper-parameters, to be defined later, associated with log mean intensities from sampled and unsampled communities by $\mathbf{\Omega}_2$.

If we denote the joint likelihood of $N(\boldsymbol{x}_{U,t}) = n(\boldsymbol{x}_{U,t}), N(\boldsymbol{x}_{S,t}) = n(\boldsymbol{x}_{S,t})$ given corresponding log mean intensities $\boldsymbol{I}^t_U, \boldsymbol{I}^t_S$ by $f(n(\boldsymbol{x}_{U,t}), n(\boldsymbol{x}_{S,t})|\boldsymbol{I}^t_U, \boldsymbol{I}^t_S)$, then the full joint likelihood is

$$\begin{aligned} & \prod_{t=1}^{7} f(n(\boldsymbol{x}_{U,t}), n(\boldsymbol{x}_{S,t})|\boldsymbol{I}^t_U, \boldsymbol{I}^t_S)\pi(\boldsymbol{I}^t_U, \boldsymbol{I}^t_S|\mathbf{\Omega}_2)\pi(\mathbf{\Omega}_2), \\ = & \prod_{t=1}^{7} f(n(\boldsymbol{x}_{U,t})|\boldsymbol{I}^t_U) f(n(\boldsymbol{x}_{S,t})|\boldsymbol{I}^t_S)\pi(\boldsymbol{I}^t_U, \boldsymbol{I}^t_S|\mathbf{\Omega}_2)\pi(\mathbf{\Omega}_2), \end{aligned} \tag{5.2}$$

where $\pi(\boldsymbol{I}^t_U, \boldsymbol{I}^t_S|\mathbf{\Omega}_2)$ is the joint distribution of the unknown $\boldsymbol{I}^t_U$ and the known $\boldsymbol{I}^t_S$ given $\mathbf{\Omega}_2$. The hyper-prior distribution of $\mathbf{\Omega}_2$ is $\pi(\mathbf{\Omega}_2)$.

The first two terms in (5.2), $[N(\boldsymbol{x}_{U,t})|\boldsymbol{I}_U^t]$ and $[N(\boldsymbol{x}_{S,t})|\boldsymbol{I}_S^t]$ follow the product of independent Poisson distributions, that is, $f(n(\boldsymbol{x}_{U,t})|\boldsymbol{I}_U^t) = \prod_{l=1}^{106} f(n(\boldsymbol{x}_{U_l,t})|\boldsymbol{I}_{U_l}^t)$ and $f(n(\boldsymbol{x}_{S,t})|\boldsymbol{I}_S^t) = \prod_{c=1}^{21} f(n(\boldsymbol{x}_{S_c,t})|\boldsymbol{I}_{S_c}^t)$, where

$$\begin{aligned} f(n(\boldsymbol{x}_{U_l,t})|\boldsymbol{I}_{U_l}^t) &\propto \exp\left\{-pop_{l,t}^n \exp(\boldsymbol{I}_{U_l}^t)\right\}\{pop_{l,t}^n \exp(\boldsymbol{I}_{U_l}^t)\}^{n(x_{U_l,t})}, \\ f(n(\boldsymbol{x}_{S_c,t})|\boldsymbol{I}_{S_c}^t) &\propto \exp\left\{-pop_{c,t}^n \exp(\boldsymbol{I}_{S_c}^t)\right\}\{pop_{c,t}^n \exp(\boldsymbol{I}_{S_c}^t)\}^{n(x_{S_c,t})}, \end{aligned}$$

where $pop_{l,t}^n$ and $pop_{c,t}^n$ represent the populations of unsampled community l and sampled community c at cycle t, respectively. Also, $n(x_{U_l,t})$ and $n(x_{S_c,t})$ are imputed and realized values for the number of cases at unsampled community l and sampled community c, respectively.

We assume the joint distribution of $\boldsymbol{I}_U^t$ and $\boldsymbol{I}_S^t$ follows the multivariate normal distribution,

$$\left[\left(\begin{array}{c} \boldsymbol{I}_S^t \\ \boldsymbol{I}_U^t \end{array}\right)\middle|\boldsymbol{\Omega}_2\right] \sim N_{21+106}\left[\left(\begin{array}{c} \boldsymbol{\mu}_S^t(\boldsymbol{\tau}) \\ \boldsymbol{\mu}_U^t(\boldsymbol{\tau}) \end{array}\right), \left(\begin{array}{cc} \boldsymbol{\Phi}_S^t(\rho_t^2,\phi) & \boldsymbol{\Phi}_{S,U}^t(\rho_t^2,\phi) \\ {\boldsymbol{\Phi}_{S,U}^t}^\top(\rho_t^2,\phi) & \boldsymbol{\Phi}_U^t(\rho_t^2,\phi) \end{array}\right)\right].$$

Here $(\boldsymbol{\mu}_S^t(\boldsymbol{\tau}))_c = \tau_{0t}+\tau_1 R(S_c)$ and $(\boldsymbol{\mu}_U^t(\boldsymbol{\tau}))_l = \tau_{0t}+\tau_1 R(U_l)$ where τ_{0t} is a time t specific offset and the parameter τ_1 corresponds to the distance between Borbón and the center of community S_c and U_l, respectively. In terms of the covariance, $(\boldsymbol{\Phi}_S^t(\rho_t^2,\phi))_{c'c''} = \rho_t^2 \exp(-\phi\|c' - c''\|^2)$ where c' and c'' represent two centers of $\{S_c\}_{c=1}^{21}$. Likewise, the $(\boldsymbol{\Phi}_{S,U}^t(\rho_t^2,\phi))_{c'l'} = \rho_t^2 \exp(-\phi\|c' - l'\|^2)$ where c' and l' represent the center of S_c and U_l, respectively and $(\boldsymbol{\Phi}_U^t(\rho_t^2,\phi))_{l'l''} = \rho_t^2 \exp(-\phi\|l' - l''\|^2)$ where l' and l'' represent two centers of $\{U_l\}_{l=1}^{106}$, respectively. Let $\boldsymbol{\Omega}_2 = (\boldsymbol{\tau}_0, \tau_1, \phi, \boldsymbol{\rho}^2)^\top$ where $\boldsymbol{\tau}_0 = (\tau_{01}, \ldots, \tau_{07})$ and $\boldsymbol{\rho}^2 = (\rho_1^2, \ldots, \rho_7^2)$. Then, $[\boldsymbol{I}_U^t|\boldsymbol{I}^t{}_S, \boldsymbol{\Omega}_2]$ follows a conditional multivariate normal distribution.

Independent vague hyper-priors on stage II parameters $\mathbf{\Omega}_2$ follow:

$$\begin{aligned} \tau_{0t} &\overset{iid}{\sim} N(0, 10^4), \qquad t = 1, \ldots, 7, \\ \tau_1 &\sim N(0, 10^4), \\ \log(\rho_t^2) &\overset{iid}{\sim} N(0, 10^4), \qquad t = 1, \ldots, 7, \\ \log(\phi) &\sim N(0, 10^4). \end{aligned}$$

Full conditionals and posterior sampling strategy concerning these parameters are given in Appendices A.4.2 and A.4.3.

5.4 Data example

In this section, we analyze the ECODESS data with our model. We ran both first and second stage algorithms for 100,000 iterations, discarding the first 50,000 samples as burn-in. The resulting chain was thinned by saving every 50-th iteration.

Marginal posterior density estimates for parameters β_1 (temperature), β_2 (precipitation), η (spatial index), and τ_1 (remoteness) are shown in Figure 5.3. The 95% highest posterior density (HPD) intervals for temperature and precipitation cover zero, indicating temperature and precipitation do not have a significant association with diarrheal cases. However, both spatial index and remoteness exhibit a negative association with cases. After adjusting for population density, larger numbers of cases appear in areas of dense housing, consistent with previous reports (Bates *et al.*, 2007). Furthermore, the more remote the community, the fewer cases of diarrhea are observed. As argued by Eisenberg *et al.* (2006), more remote communities have less migration which possibly lowers disease transmission. Figure 5.4 shows marginal estimates of μ_{com} and σ^2_{com}: the population level mean and variance of the random intercepts μ_c^{com}, as well as the time specific mean contributions, μ_t^{time}. The population level mean contribution is $\mu_{com} = -2.05$ with variance of $\sigma^2_{com} = 0.33$, exhibiting some variation in cases across communities. Regarding temporal changes to the mean of the intensity, substantive changes occur at cycles 2 and 6, with contribution to the mean of -0.71 and -0.65 respectively, indicating a decrease in cases at these two time points.

Next we considered the predictive performance of our model. First, we selected four communities, out of 21, at which we assess internal prediction performance. The four communities chosen are Timbire, Roca Fuerte, Vaquerita, and Santo Domingo, as highlighted in Table 5.1. These four communities were chosen based on their remoteness to Borbón and their population sizes (Timbire and Roca Fuerte are not remote whereas Vaquerita and Santo Domingo are remote. Timbire and Santo Domingo have large populations whereas Roca Fuerte and Vaquerita have small populations). We then fitted our model on the remaining 17 communities and compared the predicted number of cases with the observed number of cases at the remaining four communities. As a baseline comparison, we also fitted a Poisson spatial Kriging model or PSK (Christensen and Ribeiro, 2002) to predict the number of cases at these four communities, given data from the remaining 17 communities. In the Poisson spatial Kriging model, we used remoteness as the sole predictor. The PSK was fitted at each cycle independently of the other cycles. For the PSK, the same Gaussian correlation function was used (i.e., we plugged in the posterior estimates of $\hat{\rho_t^2}$ and $\hat{\phi}$ from our model). Results from both methods are reported in Table 5.3. Both methods seem to perform fairly well in that, for the most part, 95% HPD intervals of the predicted counts at these four communities at all cycles cover the observed number of cases. Compared to our method, the PSK provides wider HPD intervals. This is possibly because the PSK estimates parameters at each cycle independently whilst μ_c^{com} and τ_1 in our model are shared across all cycles. Note that our model may underestimate the uncertainty as we use the stage I results as observed data.

We computed a discrepancy measure, the integrated relative mean squared prediction error (IRMSPE) in terms of the counts, $\text{IRMSPE}_C = \sum_{t=1}^{7}\sum_{c'=1}^{4}(\widehat{N_{c',t}} - n_{c',t})^2/n_{c',t}{}^2$. Here

$n_{c',t}$ is the observed count at community c' at cycle t and $\widehat{N_{c',t}}$ is the corresponding predicted count. The discrepancy measure IRMSPE$_C$ values for our method and for the PSK are 3.37 and 5.67, respectively. This suggests that our proposed model gives more accurate internal prediction.

The ECODESS study is ongoing. At cycle 8, 24 communities participated. In this cycle, 8 new communities were introduced and five original communities from the first 7 cycles were excluded. We ran our model to perform external validation at cycle 8 data. This includes prediction of cycle 8 data for the four communities used for internal prediction plus four new communities that were unsampled in the first 7 cycles (Yalares, Valdez, Loma Linda, and El Progreso). We use $\hat{\tau_0}$, the averaged value of $\hat{\tau_{0t}}, t = 1, \ldots, 7$, in the second stage of our method. For the PSK, the average number of cases at each community over 7 cycles are used. Results are given in Table 5.4. Both approaches predict relatively poorly at new communities, compared to good performance at the original four communities. We also obtained IRMSPE$_C$ statistics, 2.32 and 3.59 for our method and the PSK respectively. According to IRMSPE$_C$, our method shows better predictive performance compared with the PSK approach.

Remark: Figure 5.5 displays predicted counts at the 106 unsampled communities obtained from our model and from the PSK at cycles 5 and 6. Without the truth, the best we can do is note that both models give similar results.

5.5 Simulation study

In this section, we report on a simulation study that examines the performance of our proposed model. We investigate both the community-level spatial effect and global-level spatial effect on the accuracy of prediction. We generate simulated datasets based on the following setups.

We consider a 256×256 grid with cells $\boldsymbol{B} = (B)_{i,j}$, $i, j = 1, \ldots, 256$ on a global study window $S = [0,1]^2$. In $\boldsymbol{B}$, we assign nine boxes, each containing 64×64 cells, which will represent communities (Figure 5.6). These nine boxes cover 56% of the global study window. The remaining area is considered non-resident area as in the ECODESS study. We assume 3 cycles. Then, we generate points on cells ($\boldsymbol{B}$) in the nine boxes based on an LGCP with an approximated intensity $\widetilde{\Lambda_{c,t}}(s) = \exp(\widetilde{Y_{c,t}}(s) + \delta_1 \widetilde{S_{1c}}(s) + \delta_2 \widetilde{S_{2c}})$, $s \in \boldsymbol{B}$ at box c, $c = 1, \ldots, 9$, and time t, $t = 1, \ldots, 3$, in which a homogeneous process is assumed within each cell of $\boldsymbol{B}$. Here $\widetilde{Y_{c,t}}(s)$ is an approximated GRF on box c at time t. The mean of $\widetilde{Y_{c,t}}(s)$ is $\mu_c^{com} + \mu^{time} t$ where we assign $\mu_c^{com} \overset{iid}{\sim} N(\mu_{com}, 0.1)$, $c = 1, \ldots, 9$, to allow some community-wise variation and $\mu^{time} = -0.5$ for a decreasing linear trend over time in the mean intensity. We assume the same Gaussian correlation function for the covariance of $\widetilde{Y_{c,t}}(s)$ across all c: $\sigma_t^2 \exp(-\alpha \|d\|^2)$ where ${\sigma_t}^2 = 1$ and $\alpha = 1024$. With this correlation function, the correlation between two neighboring cells in $\boldsymbol{B}$ (d=0.004) is 0.985 whilst the correlation between two neighboring boxes (d=0.125) is approximately zero. The within-community spatial covariate $\widetilde{S_{1c}}(s)$ for $c = 1, \ldots, 9$ is defined by the Euclidean distance between the left top corner of box c and all cells in box c, respectively. On the other hand, the global-level spatial covariate $\widetilde{S_{2c}}$ is defined as the Euclidean distance between the center of each box c and the $(78, 136)$-th cell in $\boldsymbol{B}$. Thus, a SMP exists between $\widetilde{S_{1c}}(s)$ and $\widetilde{S_{2c}}$. One-hundred such datasets were simulated.

We assume that the data from 8 boxes are observed and used to fit the model. Furthermore, we assume points from one remaining box are unobserved in which we perform prediction.

Four scenarios are considered: 1) $(\delta_1, \delta_2, \mu_{com}) = (1, 0, 8)$, 2) $(\delta_1, \delta_2, \mu_{com}) = (1, -1.5, 8.65)$, 3) $(\delta_1, \delta_2, \mu_{com}) = (0, 0, 8.25)$, and 4) $(\delta_1, \delta_2, \mu_{com}) = (0, -1.5, 8.85)$. Each scenario depends on whether there exists a community-level spatial effect and whether there exists a global-level spatial effect. We adjust the hyper prior mean, μ_{com}, to give approximately 3,000 points at time $t = 1$. Note that $\widetilde{S_{1c}}(s)$ is considered in stage I whereas $\widetilde{S_{2c}}$ is only considered in stage II of our model. We also present the result from the PSK approach which uses count data on the eight sampled boxes at each cycle and predicts counts on the unsampled box using the covariate $\widetilde{S_{2c}}$. For the PSK, we assume the covariance is known.

We evaluate each parameter estimate via bias and mean squared error of the posterior median. We also quantify the discrepancy between observed counts and estimated or predicted counts by calculating the integrated relative mean squared error, IRMSE$_C$, and the integrated relative mean squared prediction error, IRMSPE$_C$:

$$\begin{aligned}\text{IRMSE}_C &= \sum_{t=1}^{3} \sum_{x \in sampled} \frac{(\widehat{N_{x,t}} - n_{x,t})^2}{n_{x,t}^2},\\ \text{IRMSPE}_C &= \sum_{t=1}^{3} \sum_{x \in unsampled} \frac{(\widehat{N_{x,t}} - n_{x,t})^2}{n_{x,t}^2}.\end{aligned}$$

Similarly, we evaluate the intensity function estimate by calculating the integrated relative mean squared error, IRMSE$_I$, and the predictive intensity by calculating the integrated relative mean squared prediction error, IRMSPE$_I$:

$$\text{IRMSE}_I = \sum_{t=1}^{3} \sum_{x \in sampled} \frac{(\widehat{\Lambda_{x,t}} - \lambda_{x,t})^2}{\lambda_{x,t}^2},$$

$$\text{IRMSPE}_I = \sum_{t=1}^{3} \sum_{x \in unsampled} \frac{(\widehat{\Lambda_{x,t}} - \lambda_{x,t})^2}{\lambda_{x,t}^2}.$$

Here $n_{x,t}$ is the true number of points from the box x at cycle t, $\widehat{N_{x,t}}$ is the estimated counts for the sampled box or the predicted counts for the unsampled box. Similarly, $\lambda_{x,t}$ is the averaged true intensity used for generating points in box x at cycle t and $\widehat{\Lambda_{x,t}}$ is the corresponding estimated, or predicted, intensity. Small value of these four statistics indicate better estimation and prediction.

Under each scenario, we ran a half million iterations. The first half of the posterior samples were discarded and the chain was thinned by saving every tenth iteration. The remaining 25,000 samples were used to calculate simulation statistics.

We present the simulation results in Table 5.5. Under all scenarios, parameter estimation of μ^{time} and δ_1 have small biases while σ_t^2 has a somewhat larger bias (Table 5.5.(a)). When we compare the results from scenarios 1 and 2 and scenarios 3 and 4, the presence of a global spatial effect δ_2 does not affect the bias of μ_{com}. However, uncertainty in μ_c^{com} increases by including a non-null δ_2. Table 5.5.(b) shows the results of IRMSE$_I$ and IRMSE$_C$. These discrepancy measures do not depend upon the presence of a global-level spatial effect. However, the accuracy of estimation increases when a community-level spatial effect is introduced. According to the estimated prediction measures IRMPSE$_C$ and IRMPSE$_I$, we can see that the prediction accuracy increases, as we introduce a global-level spatial effect, however, this accuracy does not change with the inclusion of a community-level spatial effect. Compared with the results from the PSK approach, our proposed model yields lower discrepancy measures indicating more accurate prediction.

5.6 Discussion

In this Chapter, we proposed a Bayesian two-stage spatio-temporal point process approach based on a log Gaussian Cox process to model case patterns from a serial case-control study of diarrheal disease in a rural province of Ecuador. In stage I we adopted the LGCP model that allows us to build a parametric model for the intensity function which can accommodate geographically referenced spatial indices and temporal covariates such as precipitation and temperature over 21 widespread communities over seven time cycles. Compared to previous spatial studies, in this analysis we used a relatively high resolution equipped with a spatially referenced covariate in order to assess its spatial effect in detail at individual communities with varying sizes. In stage II, we predict the number of diarrheal cases at unsampled communities conditional on the estimated intensities of 21 communities obtained in stage I. The Bayesian inferential framework is a natural choice offering flexible modeling and efficient computational algorithms. Consequently, the proposed method provides practical spatial and temporal perspectives on diarrheal disease prevalence. In accordance with previous findings from the ECODESS study, our results show a negative association between spatial index and the number of cases. Regarding temporal covariates, precipitation and temperature do not appear to be related to case patterns. In addition, more remote communities have fewer cases after adjusting for population differences.

There are several limitations of our current analysis. First, we ignored correlation across cycles. If strong temporal correlation exists, we can employ the modeling approach of Brix and Diggle (2001) to include a spatio-temporal correlation structure. In this context, the degree of gain in accuracy depends upon the strength of the temporal dependence as well as the appropriate correlation structure as discussed in Brix and Diggle (2001). Second, we ignored the uncertainty in the stage I posterior estimates using only point estimates for for prediction in stage II. A joint likelihood would alleviate this limitation.

As the original study design is a matched case-control study, in the future we will consider a marked point process model to differentiate the two types of outcomes: cases and controls. In addition, with community level factors studied herein, incorporating individual risk factors, typically non-spatial covariates, into the intensity model will be investigated in the future. As suggested by Liang *et al.* (2009), we can use continuous risk factors such as age and sanitation level by introducing interaction terms with spatially or temporally varying covariates. To do this, the study domain needs to be extended to the product space $S \times T \times V$ where S denotes space, T denotes time, and V denotes individual risk factors.

Table 5.2: Covariate summaries by community.

Community	Precipitation(mm)/Temperature(°C)							Spatial Index (m)
	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6	Cycle 7	median (min, max)
San Agustin	0.4/25.8	0.2/27.0	0.3/26.8	0.2/26.6	0.1/26.1	0.5/26.1	0.0/25.5	4.3 (1.7, 5.0)
Colon Eloy	0.3/26.1	0.2/26.8	0.3/26.7	0.2/26.5	0.2/25.9	0.5/26.1	0.0/25.4	3.3 (1.7, 4.1)
Naranjal	0.4/26.6	0.1/26.4	0.3/26.2	0.1/26.5	0.3/25.8	0.4/26.1	0.0/26.1	4.4 (-0.1, 5.3)
Timbire	0.1/25.8	0.5/26.2	0.3/25.4	0.2/25.9	0.3/25.8	0.6/26.1	0.2/25.1	3.4 (1.9, 4.3)
Roca Fuerte	0.4/26.0	0.3/25.5	0.3/26.4	0.1/26.2	0.2/25.8	0.5/26.1	0.0/25.5	2.1 (0.2, 3.0)
La Loma	0.3/26.8	0.1/26.6	0.3/26.4	0.1/26.6	0.2/25.8	0.3/26.1	0.0/26.1	4.8 (0.2, 5.4)
Ranchito	0.4/26.6	0.1/26.4	0.3/26.2	0.1/26.5	0.2/25.8	0.4/26.1	0.0/26.1	3.8 (0.1, 4.5)
Quinto piso	0.3/26.0	0.3/26.6	0.4/26.6	0.2/26.4	0.2/25.8	0.6/26.1	0.0/25.4	4.8 (0.7, 5.3)
La Pena	0.4/26.0	0.4/25.4	0.3/26.3	0.1/26.1	0.2/25.8	0.6/26.1	0.0/25.4	1.9 (0.0, 2.8)
Las Cruces	0.4/26.3	0.1/25.8	0.5/27.1	0.1/26.3	0.4/25.8	0.5/26.1	0.0/26.0	3.5 (0.1, 4.3)
Tangare	0.3/25.6	0.1/25.6	1.3/26.0	0.1/26.1	0.5/25.8	0.3/26.6	0.0/26.0	3.2 (1.0, 4.1)
El Rosario	0.3/25.2	0.1/25.2	0.6/26.8	0.1/25.9	0.5/25.8	0.5/26.1	0.0/25.7	2.9 (-3.1, 3.7)
Guayabal	0.5/25.3	0.6/25.8	0.3/25.2	0.2/25.7	0.3/25.6	0.7/26.1	0.2/25.0	2.2 (0.2, 3.1)
Arenales	0.4/26.0	0.4/25.6	0.3/26.1	0.7/25.7	0.3/25.8	0.2/25.8	0.0/25.9	4.4 (1.2, 5.2)
Wimbi	0.5/25.8	0.4/25.2	0.3/26.0	0.1/25.9	0.2/25.8	0.7/26.1	0.0/25.2	2.4 (0.5, 3.2)
Playa de Oro	0.6/24.5	0.6/25.0	0.2/24.7	0.1/25.3	0.3/25.2	0.9/26.1	0.1/24.5	2.0 (1.0, 2.8)
Trinidad	0.4/25.3	0.1/25.2	0.6/26.7	0.8/25.4	0.5/25.8	0.5/26.1	0.0/25.6	1.6 (0.2, 2.3)
Telembi	0.4/25.3	0.4/25.3	0.7/26.7	0.8/25.4	0.6/25.8	0.5/26.1	0.0/25.6	2.6 (1.2, 3.5)
Vaquerita	0.4/26.0	0.4/25.6	0.3/26.0	0.7/25.7	0.4/25.8	0.3/25.8	0.0/26.5	2.4 (-0.2, 3.1)
Santo Domingo	0.5/26.0	0.4/25.5	0.3/26.0	0.7/25.7	0.4/25.8	0.4/26.4	0.0/25.9	3.8 (1.2, 4.5)
San Miguel	0.6/25.7	0.4/25.3	0.9/26.7	0.8/25.4	0.7/25.8	0.5/26.1	0.0/25.7	1.9 (-0.4, 2.6)

Table 5.3: Prediction results for four selected communities. The numbers in the parenthesis are the 95% highest posterior density (HPD) intervals.

Community	Method	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6	Cycle 7
	True	11	10	5	11	17	6	10
Timbire	Our model	9 (2, 14)	8 (2, 12)	9 (3, 14)	10 (4, 17)	12 (6, 20)	7 (2, 13)	8 (3, 14)
	PSK	11 (1, 27)	10 (1, 23)	9 (1, 22)	17 (2, 42)	18 (2, 46)	9 (1, 23)	10 (1, 23)
	True	7	2	3	7	4	5	4
Roca Fuerte	Our model	3 (0, 6)	2 (0, 5)	3 (0, 6)	3 (0, 7)	4 (0, 8)	2 (0, 5)	2 (0, 5)
	PSK	4 (1, 10)	4 (1, 9)	4 (0, 8)	4 (0, 8)	5 (1, 10)	2 (0, 5)	3 (0, 8)
	True	0	1	3	2	3	0	0
Vaquerita†	Our model	0 (0, 2)	0 (0, 2)	0 (0, 2)	0 (0, 2)	1 (0, 3)	0 (0, 2)	1 (0, 3)
	PSK	1 (0, 2)	1 (0, 2)	1 (0, 2)	0 (0, 1)	1 (0, 2)	0 (0, 1)	1 (0, 2)
	True	10	4	8	10	10	4	11
Santo Domingo	Our model	8 (2, 18)	8 (2, 17)	8 (1, 17)	9 (1, 18)	10 (1, 21)	5 (0, 11)	6 (1, 13)
	PSK	9 (1, 22)	8 (1, 19)	9 (1, 24)	16 (1, 41)	16 (2, 42)	7 (1, 17)	6 (1, 16)

* Rounded predicted numbers are shown.

†Vaquerita is excluded in calculating IRMSPE$_C$ to avoid division by zero.

Table 5.4: The number of predicted cases from eight communities out of 24 communities at cycle 8 where four communities are sampled in the previous cycles and four communities are newly sampled. The numbers in the parenthesis are 95% HPD credible intervals.

					Predicted number of cases		
	Community	Distance(km)	Size	Population	Observed	Our model	PSK
Original	Timbire	20.59	83 × 50	517	9	10 (8, 12)	8 (4, 11)
	Roca Fuerte	19.99	25×18	141	2	3 (2, 4)	5 (2, 9)
	Telembi	32.76	43×24	388	5	9 (7, 11)	5 (3, 7)
	Santo Domingo	34.05	89×29	420	7	6 (4, 7)	5 (3, 8)
New	Yalares	13.13	166 × 104	146	12	4 (2, 5)	5 (2, 10)
	Valdez	16.14	173 × 122	336	9	9 (7, 11)	8 (4, 14)
	Loma Linda	38.66	126 × 98	262	17	7 (5, 8)	4 (2, 7)
	El Progreso	41.12	50 × 126	123	9	2 (1, 3)	4 (2, 7)

* Rounded predicted numbers are shown.

Table 5.5: Simulation results.

Scenario 1 ($\delta_2 = 0$)	μ_C	σ_C^2	σ_1^2	σ_2^2	σ_3^2	β_t	δ_1
True value	8.0	0.01	1.0	1.0	1.0	-0.5	1.0
bias	0.05	0.06	-0.08	-0.010	-0.13	-0.02	-0.05
MSE	0.03	0.00	0.01	0.01	0.02	0.00	0.49
Scenario 2 ($\delta_2 \approx -1.5$)	μ_C	σ_C^2	σ_1^2	σ_2^2	σ_3^2	β_t	δ_1
True value	8.65	0.01	1.0	1.0	1.0	-0.5	1.0
bias	0.03	0.13	-0.08	-0.10	-0.11	-0.01	-0.02
MSE	0.02	0.02	0.01	0.01	0.02	0.00	0.50
Scenario 3 ($\delta_2 = 0$)	μ_C	σ_C^2	σ_1^2	σ_2^2	σ_3^2	β_t	δ_1
True value	8.25	0.01	1.0	1.0	1.0	-0.5	0.0
bias	0.01	0.06	-0.08	-0.09	-0.11	-0.01	-0.03
MSE	0.04	0.00	0.01	0.01	0.02	0.00	0.63
Scenario 4 ($\delta_2 \approx -1.5$)	μ_C	σ_C^2	σ_1^2	σ_2^2	σ_3^2	β_t	δ_1
True value	8.85	0.01	1.0	1.0	1.0	-0.5	0.0
bias	-0.02	0.13	-0.07	-0.10	-0.11	-0.02	0.01
MSE	0.02	0.02	0.01	0.01	0.02	0.00	0.50

(a)

Scenario	$IRMSE_C$	$IRMSPE_C$	$IRMSPE_{PSK}^{\dagger}$	$IRMSE_I$	$IRMSPE_I$
1	0.0158	0.0027	0.0085	1.175	0.211
2	0.0157	0.0021	0.0086	1.156	0.198
3	0.0165	0.0031	0.0091	1.137	0.201
4	0.0162	0.0022	0.0093	1.148	0.184

†PSK : Poisson spatial Kriging.

(b)

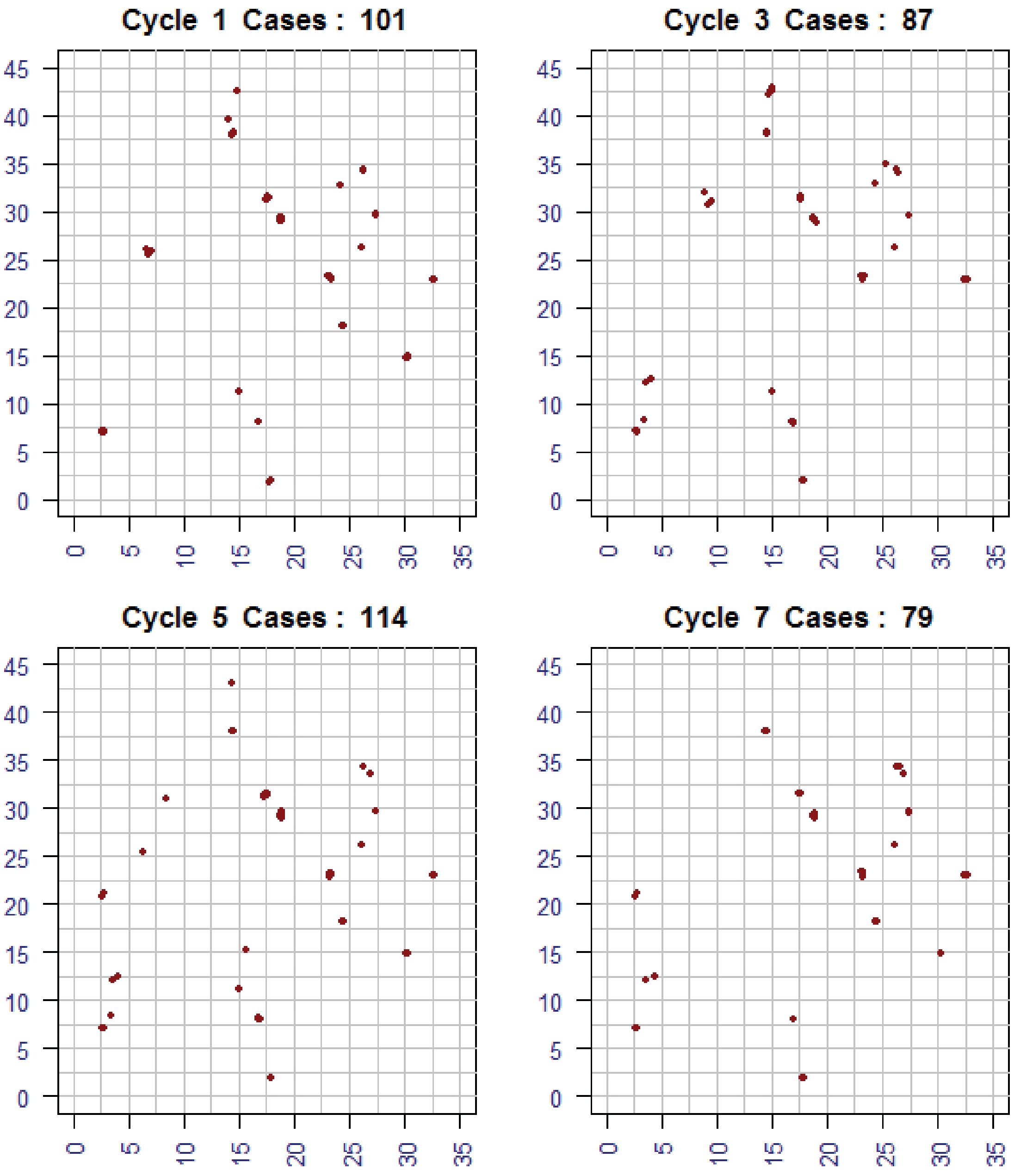


Figure 5.2: Spatial and temporal changes in diarrheal cases in Esmeraldas province across 4 selected cycles. One unit on the 35 × 45 grid corresponds to 1.1km.

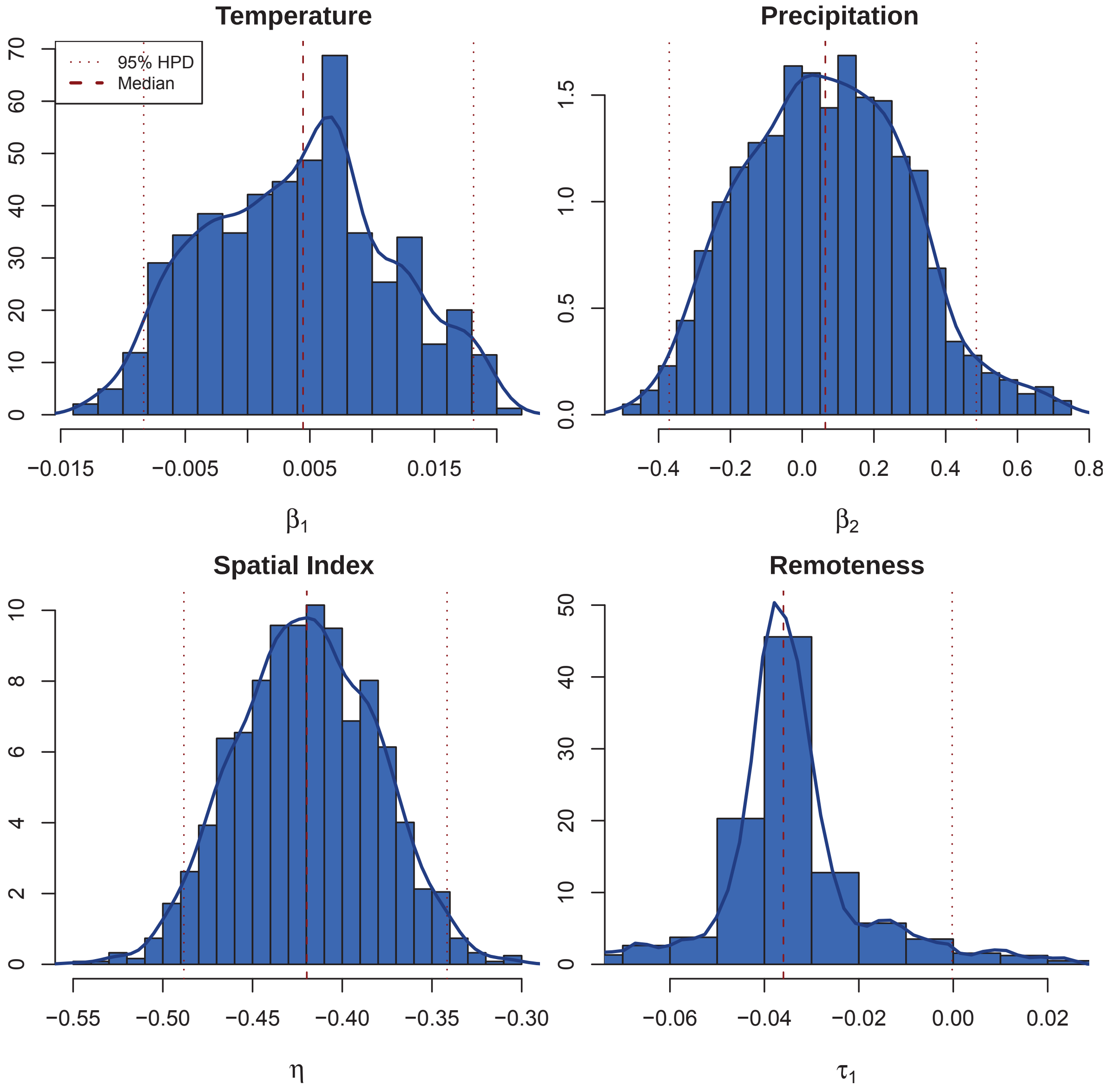


Figure 5.3: Marginal posterior densities.

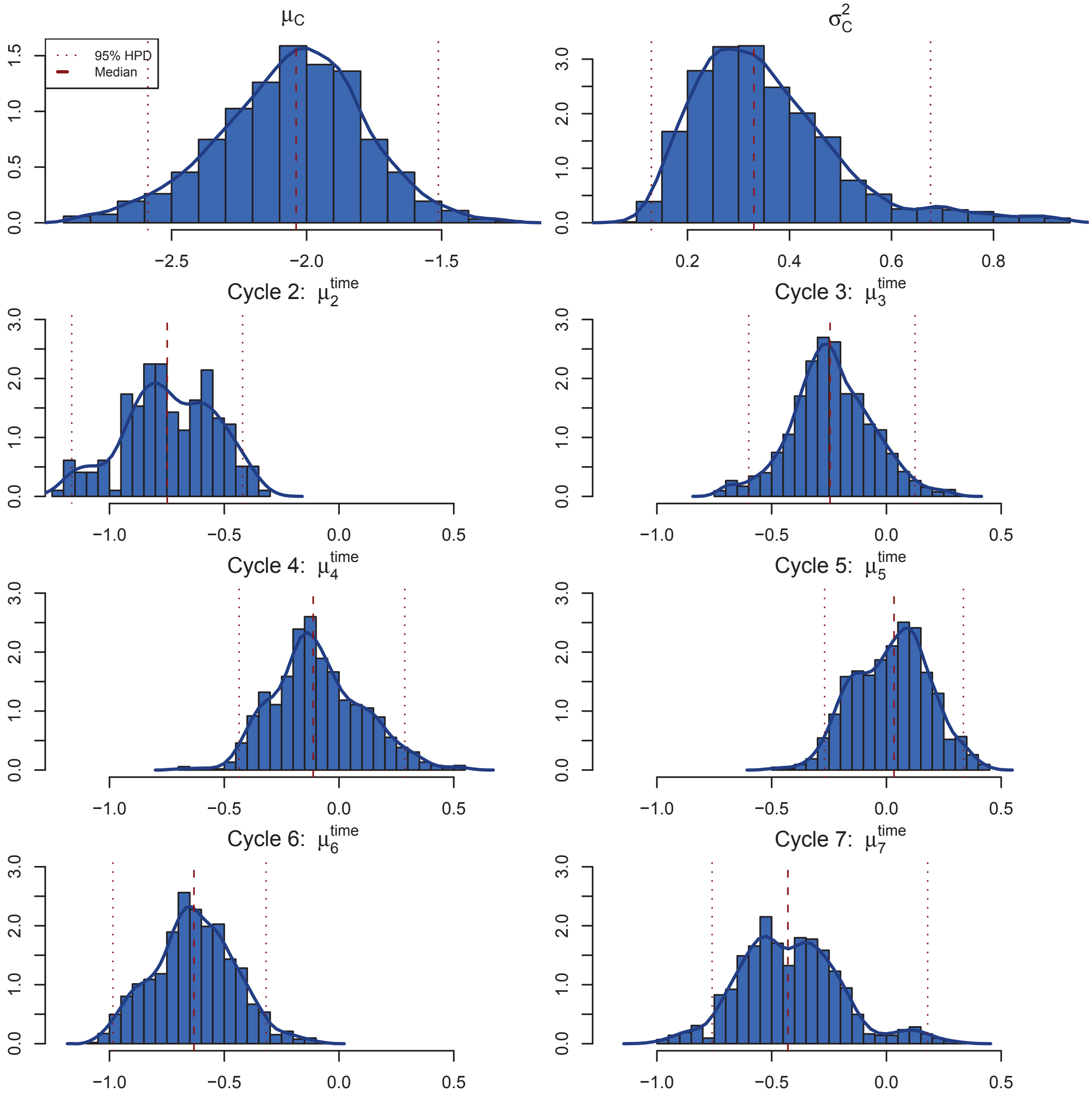


Figure 5.4: Marginal posterior densities.

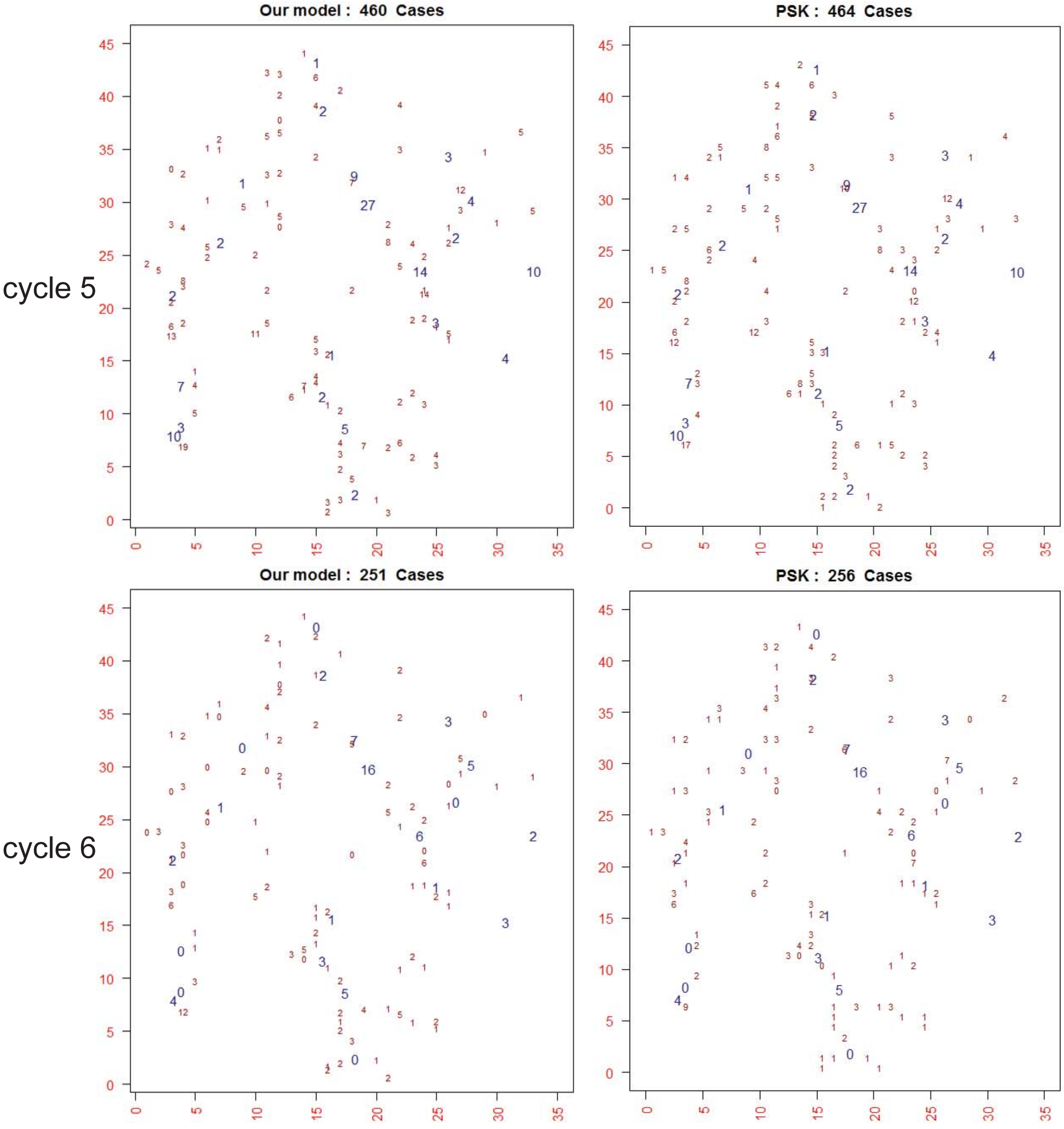


Figure 5.5: Predicted number of cases at unsampled communities (red) at cycles 5 and 6. Observed number of cases at sampled communities are shown in blue. Left column: results from our proposed model. Right Column: results from the PSK. One unit on each axis corresponds to 1.1km.

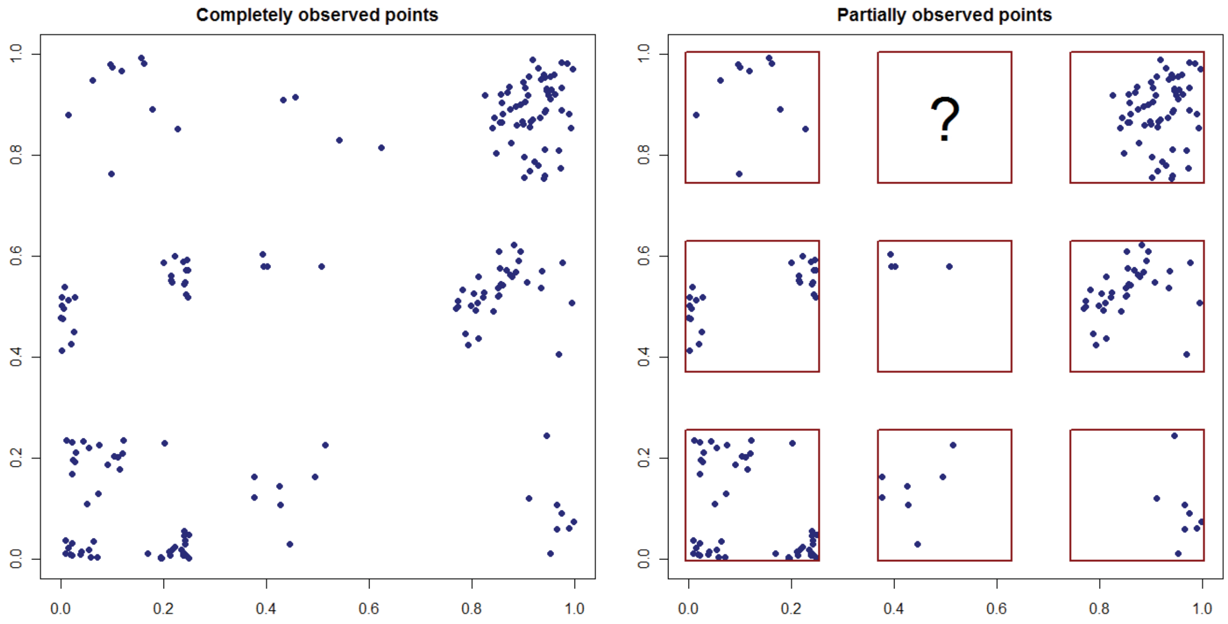


Figure 5.6: Left panel shows point patterns on the global study window. Right panel shows point patterns on 9 rectangular boxes including eight training boxes and one validation box. The region outside the boxes are considered unpopulated areas where no 'case' can occur.

CHAPTER VI

Discussion

6.1 Discussion

Case-control studies have attracted research attention of epidemiologists as well as statisticians for decades. In this dissertation, we developed methods to deal with three novel problems under case control sampling scheme. In Chapter II, we proposed a Bayesian inference under the *stereotype* regression model for cancer subtypes in matched case control study. Following the work in Chapter II, in Chapter III we presented expectation/conditional maximization algorithm as well as a full Bayes procedure with data augmentation to handle non-ignorable missingness in covariates under the stereotype regression model. We proposed a flexible Bayesian approach to estimate gene-gene (G x G) and gene-environment (G x E) interactions under two-phase sampling with potential missingness in genetic covariates in Chapter IV. Finally, we proposed a Bayesian two stage spatio-temporal point process model to analyze a serial case-control study of diarrheal disease carried out in the developing country Ecuador. The proposed approach involves community-wise association study with spatial/temporal predictors and prediction of disease prevalence at unsampled communities.

Since most of the modeling techniques are fairly new in this domain, there are many pending issues in relation to this dissertation and possible extensions beyond this dissertation. Though the work on Chapter II and III are fairly complete, there are possible extensions of the stereotype regression model in the Bayesian framework that need further explorations. One can consider a more flexible semi-parametric model for the exposure distribution compared to the models belong to exponential family and the missingness mechanism. For example, modeling for mixture of continuous and categorical exposures using the kernel-mixture extension discussed in Bhattacharya and Dunson (2011) will be an interesting alternative. Also, handling missingness with correlated or clustered observations as in a longitudinal cohort study under this class of models is of interest. In the same context, a random effect approach on the stratum effects as pointed out by Rice (2008) is also a plausible alternative rather than using conditional likelihood. This will reduce the bias under data missing at random for complete-case analysis. One can also explore how to further reduce computational complexity due to non-linearity, lack of identifiability in the parameters, and missing data under the frequentist framework.

Apart from the theoretical and methodological aspects in Chapter IV, implementation of the hierarchical Bayesian methods with a rather complex posterior space can pose potential problems. In this respect, our method in the presence of a truly high-dimensional gene model > 50 needs to be addressed. This includes handling not only high-dimensional pairwise G x E interactions in the disease risk model but also G-E associations in the association model and calculation of $P(D)$ in the denominator of the retrospective likelihood. For modeling general types of predictors, we can consider mixed set of discrete and continuous covariates following Bhattacharya and Dunson (2011).

In relevance to Bayesian spatio-temporal modeling in Chapter V, there are a gamut of issues that need to be further studied. With the original study design in mind, we can consider a marked-point process model to differentiate two types of outcomes, case and controls, as attempted in Liang *et al.* (2009). In addition, we can introduce individual risk factors, such as age, into the intensity model by allowing interaction terms with spatially or temporally varying covariates. We can potentially consider adapting non-separable time-spatial covariance structure by Gneiting (2002) into modeling intensity. The two stage approach we currently present underestimate the estimator uncertainty. A joint likelihood that integrates the two stages will be developed as a part of our future work.

To conclude, the current thesis is expected to generate new ideas for further propagation of Bayesian thoughts under complex and novel sampling/data structure that are related to variations of case-control sampling designs.

APPENDICES

APPENDIX A

Appendix

A.1 Chapter II : Supplementary Material

A.1.1 Testing of hypotheses in stereotype regression model

As indicated in the main text, the multiplicative nature of the stereotype model poses some issues for testing the null hypothesis of independence $H_0 : \boldsymbol{\beta} = 0$ in the likelihood based framework. Under this global null hypothesis, the score parameters $\{\phi_k\}$ are not identifiable. McCullagh (1984) pointed out in the discussion of Anderson (1984) that the approximate null distribution of the likelihood-ratio statistic is that of the largest eigenvalue of a Wishart matrix (Haberman, 1981). The testing problem under such non-regular conditions remains to be explored in the frequentist domain for this class of models. Theories developed for modified partial likelihood ratio test (Hanfelt and Liang, 1995; Chen *et al.*, 2001) could be useful in deriving a suitable test under this class of models.

The Bayesian paradigm provides a natural alternative to bypass the testing dilemma under a non-identifiable parameter setting. One natural approach is to examine the HPD confidence intervals for a set of plausible values of $\boldsymbol{\beta}$ supported by the data. One can also compare exact posterior probabilities of the null model and the unrestricted model by comparing Bayes factors (Berger, 1985), without relying on asymptotic theory for the LR statistic.

Bayes factors are typically used to compare two competing models M_1 and M_2 by calculating the ratio of the marginal likelihood under the two models:

$$BF(M_2; M_1) = \frac{\int p(\theta_2|M_2)p(y|\theta_2, M_2)d\theta_2}{\int p(\theta_1|M_1)p(y|\theta_1, M_1)d\theta_1} = \frac{p(M_2|y)/p(M_2)}{p(M_1|y)/p(M_1)}.$$

For the purpose of testing our hypotheses of interest which are often of the form $H_0 : \boldsymbol{\theta} \in \Theta_0$, we consider M_1 as the space $\boldsymbol{\theta} \in \Theta_0$ and M_2 as the space defined by the alternative, H_a : $\boldsymbol{\theta} \in \Theta_1$. In many situations $\Theta_1 = \Theta_0^c$ and thus, $p(M_2|y) = 1 - p(M_1|y)$. In this Chapter, we use proper priors on the parameters and avoid testing a point null hypothesis by defining Θ_0 to be an interval or ball around the standard point null of interest, that has positive prior probability.

Another hypotheses of interest in the stereotype regression model is to test whether the categories k and l are indistinguishable in terms of the predictors by testing the simple hypotheses $H_0 : \phi_k = \phi_l$. Tests for ordering property of $\{\phi_k\}$ or for the hypothesis of equal spacing of $\{\phi_k\}$ can also be easily carried out as the choice between models of different levels of complexity is made within the same hierarchy of models. Such tests of ordering as well as indistinguishability among categories can indicate whether the data support stochastic ordering of outcomes or collapsing across categories. The model also allows one to calculate a measure of how different category k vs l are by simply estimating $|\phi_k - \phi_l|$. Empirically, one would choose the simplest model which will best fit the data.

Under $\boldsymbol{\beta} \neq \mathbf{0}$, one can continue testing the indistinguishability hypotheses for a subset of s categories $s < K$, as well as $H_0 : \phi_k = \phi_l$ for $s = 2$. The likelihood ratio statistic has a standard chi-squared distribution with df=$s - 1$. One can compare the general model to the model with equally spaced fixed scores or with the polytomous logit model, also by using likelihood ratio chi-squared statistic. Comparison between the ordered and unordered model can be carried out by comparing the deviance statistic or through other popular model selection criteria. We illustrate the hypotheses testing issues further through our real data example.

A.1.2 Hypotheses testing in the Flint Men's Health Study

We performed evaluation of three hypotheses of common interest for illustrative purposes. Namely: (i) the indistinguishability hypotheses that the scores ϕ_1 and ϕ_2 are not considerably different. (ii) the hypotheses of global independence that age and PSA have no association with the stages of cancer and (iii) the hypotheses that after controlling for PSA, age and stages of cancer have no conditional association. In the Bayesian framework, the first hypotheses in (i) can be tested by comparing the posterior odds of $|\phi_1 - \phi_2| \in (0, \epsilon)$ for a small value of ϵ, normalized by prior odds. For (ii) one can evaluate the posterior odds of the joint probability $\beta_1 \in (-\epsilon_1, \epsilon_1)$ and $\beta_2 \in (-\epsilon_2, \epsilon_2)$ whereas (iii) warrants evaluation of the probability $\beta_1 \in (-\epsilon_1, \epsilon_1)$. All the hypotheses are evaluated for small but *arbitrary* values of ϵ_l, $l = 1, 2$.

The generated sequence of posterior observations on each parameter allow us to obtain the posterior probabilities of each hypotheses under consideration and the prior odds can be evaluated under the assumed prior choices, finally providing us with a Bayes factor for H_0 against H_a. We recognize that the Bayes factor is sensitive to the choice of ϵ and the prior odds. With $\epsilon = 0.1$ and $\epsilon_1 = \epsilon_2 = 0.5$, the Bayes Factors for hypotheses (i), (ii) and (iii) are respectively 4.45, ≈ 0.0, 4.86 for the ordered model and 4.33, ≈ 0.0, 4.76 for the unordered model. The tentative results indicate evidence in favor of (i) and (iii) but overwhelming evidence against (ii) due to large effect of β_2 corresponding to PSA. As described before, there is no appropriate test for the global null hypothesis (ii) in the frequentist setting due to non-regular conditions, but under $\boldsymbol{\beta} \neq \mathbf{0}$, we can test the indistinguishability hypotheses (i) $H_0 : \phi_1 = \phi_2$ under the ML approach. The likelihood ratio statistic 1.625 has a standard chi-squared distribution with df=1, with a corresponding p-value of 0.20. We also test the null hypothesis of association between age and stage after controlling for PSA $H_0 : \beta_1 = 0$

via the LR test, and the likelihood ratio statistic has value 4.33 with a p-value of 0.04, a conclusion different from the one obtained by Bayesian methods.

We repeat the evaluation of hypotheses (i)-(iii) as discussed above with the matched dataset. With identical values for ϵ and ϵ_l, $l = 1, 2$. The Bayes factors in favor of the null model are respectively 0.43, $\approx$ 0.0, and 0.86 for unordered stereotype model and 0.55, $\approx$ 0.0, 0.66 for the ordered stereotype model. The evidence in favor of the indistinguishability hypotheses (i) and the hypotheses of no conditional association of age and stage after controlling for PSA is weaker, falling in the zone of indecision according to the Bayes Factor values. However, the evidence against (ii), the global null hypotheses of independence is again astronomic. Under $\boldsymbol{\beta} \neq \mathbf{0}$, we can test of $H_0 : \phi_1 = \phi_2$ in a ML framework. The likelihood ratio chi-square statistic value for (i) is 0.33 and p-value of 0.56, illustrating the plausibility of the indistinguishability hypotheses. Likewise, the likelihood ratio statistics for testing $H_0 : \beta_1 = 0$ is 1.80 with p-value of 0.18, which shows evidence in favor of (iii). The conditional multinomial logit model does not converge for estimation of log (OR) corresponding to Stage 3 versus controls.

In order to test the simple hypotheses regarding the partial effect of each predictor, namely age and transformed PSA one can also examine whether the null hypothesized value of zero lies within the 95% HPD interval. Table A.1 with unmatched data illustrate that transformed PSA is strongly associated with cancer stage (95% HPD: (3.39,5.15)) whereas age is not associated (95% HPD: (-2.31,1.99)). Similar findings are noted in Table A.2 with analysis of matched data.

We also investigated the evidence in favor of ordering in the $\boldsymbol{\phi}$ via calculating the posterior probability that the $\boldsymbol{\phi}$ are monotone, i.e. $\pi((0 = \phi_0 < \phi_1 < \phi_2 < \phi_3 = 1|Y, X, Z, S)$, while fitting the unordered Bayes model. These posterior probabilities for unmatched and matched datasets are 0.892 and 0.722 respectively, suggesting that the ordered model may be preferred in both instances, a finding that is consistent with the point estimates and the DICs noted in Tables A.1 and A.2 .

Finally, in order to assess the reflection of the non-identifiability under $\beta = 0$, issue in the Bayesian context we generated data from the null model $\beta = 0$ and noted that the posterior dependence between β and ϕ is very weak, for any choice of ϕ.

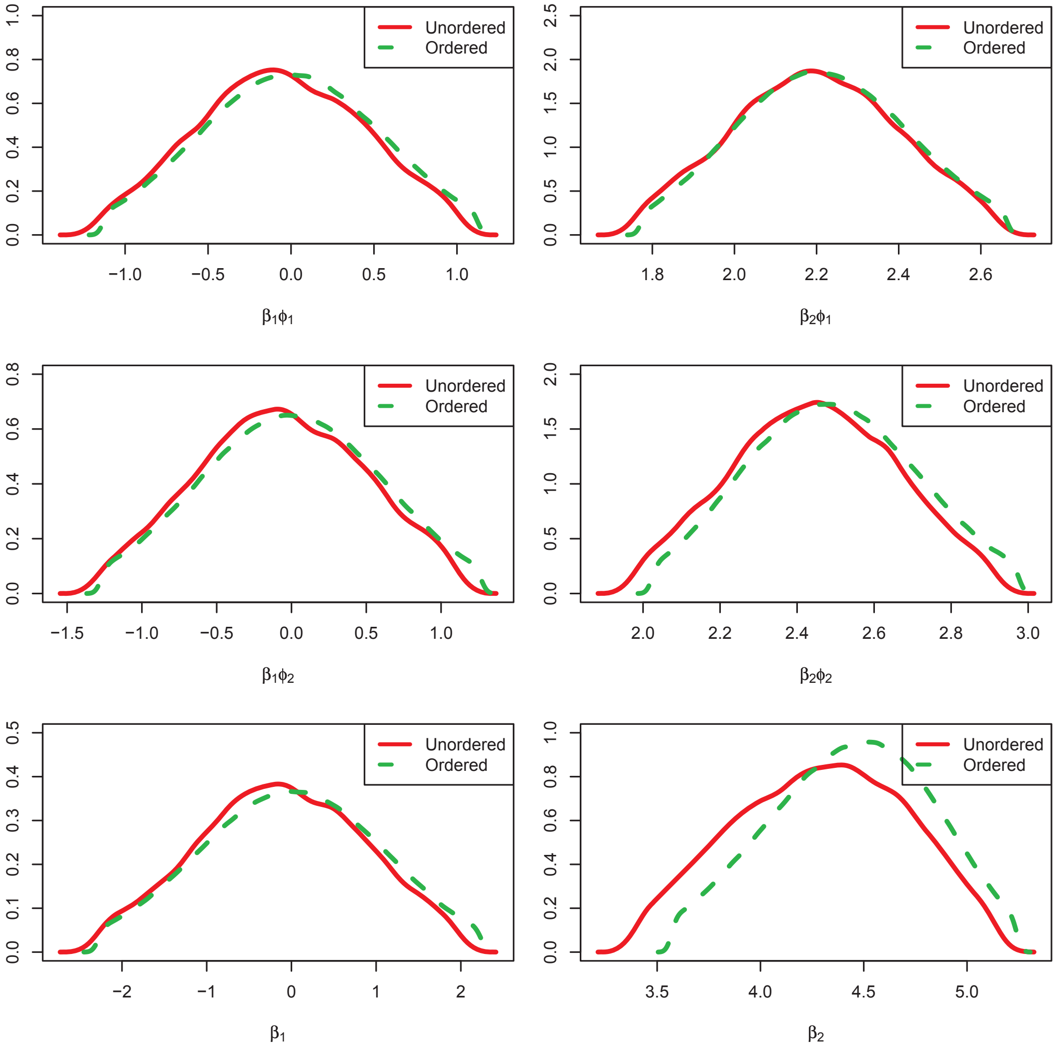


Figure A.1: Posterior density estimates for the log odds-ratio parameters in the stereotype model for unmatched FMHS data with numerical summaries as presented in Appendix Table A.1. The log(OR) parameters of category k vs category 0 with covariates X_r, are represented by, $\phi_k\beta_j$, $k = 1, 2, 3, 4$, $r = 1, 2$ with X_1 corresponding to the scaled age variable, and X_2 corresponding to log(1+PSA). The results are based on 10,000 observations generated from the posterior distribution of each parameter. The solid line corresponds to the unordered model, whereas the dashed line corresponds to the ordered model.

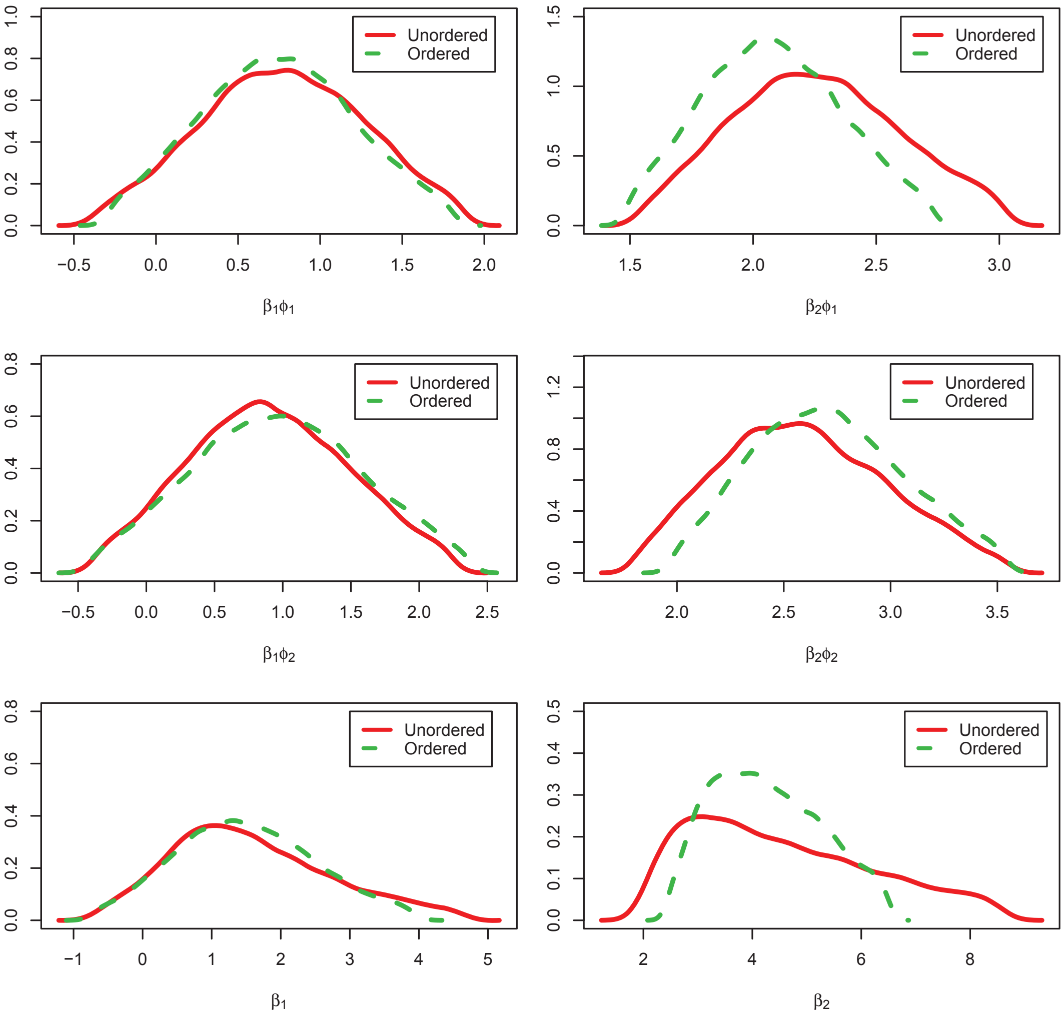


Figure A.2: Posterior density estimates for the log odds-ratio parameters in the stereotype model for 1:3 matched FMHS data with numerical summaries as presented in Appendix Table A.2. The log(OR) parameters of category k vs category 0 with covariates X_r, are represented by, $\phi_k\beta_r$, $k = 1, 2, 3, 4$, $r = 1, 2$ with X_1 corresponding to the scaled age variable, and X_2 corresponding to log(1+PSA). The results are based on 10,000 observations generated from the posterior distribution of each parameter. The solid line corresponds to the unordered model, whereas the dashed line corresponds to the ordered model.

Table A.1: Results corresponding to the log odds-ratios of category k vs category 0 with covariates X_r, namely, $\phi_k\beta_r$, $k = 1, 2, 3$, $r = 1, 2$ using unmatched FMHS data with corresponding parameter summaries as presented in Table 2.1 of main text. Under the Bayesian methods the 'Mean' corresponds to the posterior mean whereas PSD and HPD correspond to posterior standard deviation and highest posterior density intervals respectively. For the MLE, CI corresponds to the Wald-type large sample confidence intervals with variance for the log(OR) derived using multivariate delta theorem.

		X_1(=Transformed Age)			X_2(=log(1+PSA))		
		0:1	0:2	0:3	0:1	0:2	0:3
Unordered Stereotype	Mean	-0.06	-0.07	-0.12	2.21	2.44	4.29
Bayes	95% HPD	(-0.99,0.88)	(1.10, 0.97)	(-1.97, 1.71)	(1.78, 2.60)	(2.05, 2.85)	(3.39, 5.15)
Unordered Stereotype	Est.	-0.68	-0.75	-1.92	3.08	3.38	8.72
MLE	95% CI	(-1.31, -0.05)	(-1.44, -0.06)	(-3.64, -0.19)	(2.16, 3.99)	(2.40, 4.36)	(7.11, 10.33)
Ordered Stereotype	Mean	-0.01	-0.01	-0.02	2.21	2.49	4.44
Bayes	95% HPD	(-1.13, 1.12)	(-1.27, 1.27)	(-2.28, 2.24)	(1.78, 2.66)	(2.02, 2.98)	(3.57, 5.24)

Table A.2: Results corresponding to the log odds-ratios of category k vs category 0 with covariates X_r, namely, $\phi_k\beta_r$, $k = 1, 2, 3$, $r = 1, 2$ using 1:3 matched FMHS data with corresponding parameter summaries as presented in Table 2.2 of main text. Under the Bayesian methods the 'Mean' corresponds to the posterior mean whereas PSD and HPD correspond to posterior standard deviation and highest posterior density intervals respectively. For the MLE, CI corresponds to the Wald-type large sample confidence intervals with variance for the log(OR) derived using multivariate delta theorem.

		X_1(=Transformed Age)			X_2(=log(1+PSA))		
		0:1	0:2	0:3	0:1	0:2	0:3
Unordered Stereotype	Mean	0.77	0.90	1.58	2.29	2.59	4.56
Bayes	95% HPD	(-0.18, 1.71)	(0.26, 1.97)	(-0.50, 3.90)	(1.60, 3.01)	(1.89,3.34)	(1.97, 7.78)
Unordered Stereotype	Est.	0.78	0.90	1.58	2.22	2.56	4.47
MLE	95% CI	(0.15, 1.42)	(0.17 , 1.63)	(0.34, 2.81)	(-0.24, 4.69)	(-0.27, 5.38)	(0.98, 7.79)
Ordered Stereotype	Mean	0.75	1.00	1.52	2.09	2.72	4.26
Bayes	95% HPD	(-0.10, 1.64)	(-0.19, 2.14)	(-0.40, 3.46)	(1.49, 2.64)	(2.08, 3.41)	(2.60, 6.10)

Table A.3: Results of the simulation study corresponding to the log(OR) parameters of category k vs category 0 with covariates X_r, namely, $\phi_k\beta_r$, $k = 1, 2, 3, 4$, $r = 1, 2$ under unmatched case-control sampling design with corresponding parameter summaries as presented in Table 2.3 of main text. The results are based on 500 simulated datasets. For each parameter we report estimated bias and mean squared error based on the 500 replications. The outcome variable Y has five categories from 0 to 4. The true values for the parameters are: $\beta_1 = 1.0$, $\beta_2 = 2.0$; for unordered setting $\phi_1 = 0.6$, $\phi_2 = 0.9$, $\phi_3 = 0.6$ and for ordered setting $\phi_1 = 0.25$, $\phi_2 = 0.5$ and $\phi_3 = 0.75$.

Unordered data		X_1				X_2			
		0:1	0:2	0:3	0:4	0:1	0:2	0:3	0:4
Unordered Stereotype	Bias	-0.01	-0.03	0.01	-0.02	0.01	-0.03	0.08	-0.06
Bayes	MSE	0.11	0.23	0.12	0.25	0.05	0.07	0.07	0.08
Unordered Stereotype	Bias	0.09	0.08	0.07	0.20	0.13	0.07	0.13	0.14
MLE	MSE	0.17	0.32	0.15	0.38	0.09	0.12	0.10	0.96
Ordered Stereotype	Bias	-0.08	-0.19	0.16	0.05	-0.09	-0.32	0.40	0.08
Bayes	MSE	0.10	0.20	0.21	0.33	0.05	0.13	0.19	0.08
Ordered data		X_1				X_2			
		0:1	0:2	0:3	0:4	0:1	0:2	0:3	0:4
Unordered Stereotype	Bias	-0.05	-0.08	-0.13	-0.12	-0.02	-0.05	-0.07	0.14
Bayes	MSE	0.03	0.09	0.18	0.30	0.04	0.04	0.05	0.12
Unordered Stereotype	Bias	-0.01	0.01	-0.01	0.12	0.01	0.01	0.01	0.13
MLE	MSE	0.04	0.15	0.25	0.59	0.04	0.07	0.09	0.25
Ordered Stereotype	Bias	-0.04	-0.08	-0.11	-0.08	-0.03	-0.04	-0.07	-0.08
Bayes	MSE	0.02	0.08	0.17	0.29	0.02	0.02	0.03	0.06

Table A.4: Results of the simulation study under matched case-control sampling design with log(OR) parameters of category k vs category 0 with covariates X_r, namely, $\phi_k\beta_r$, $k = 1, 2, 3, 4$ $r = 1, 2$ with corresponding parameter summaries as presented in Table 2.4 of main text. The results are based on 500 simulated datasets. For each parameter we report estimated bias and mean squared error based on the 500 replications. The outcome variable Y has five categories from 0 to 4. The true values for the parameters are: $\beta_1 = 1.0$, $\beta_2 = 2.0$; for unordered setting $\phi_1 = 0.6$, $\phi_2 = 0.9$, $\phi_3 = 0.6$ and for ordered setting $\phi_1 = 0.25$, $\phi_2 = 0.5$ and $\phi_3 = 0.75$.

Unordered Setting		X_1				X_2			
		0:1	0:2	0:3	0:4	0:1	0:2	0:3	0:4
Unordered Stereotype	Bias	0.01	0.05	0.07	0.00	0.05	0.12	0.18	0.01
Bayes	MSE	0.12	0.33	0.13	0.38	0.10	0.34	0.16	0.30
Unordered Stereotype	Bias	0.01	0.03	0.06	0.15	0.05	0.09	0.15	0.30
MLE	MSE	0.13	0.34	0.13	0.77	0.10	0.33	0.15	1.02
Ordered Stereotype	Bias	-0.07	-0.13	0.29	0.22	-0.13	-0.28	0.57	0.43
Bayes	MSE	0.10	0.20	0.22	0.46	0.06	0.15	0.29	0.34
Ordered Setting		X_1				X_2			
		0:1	0:2	0:3	0:4	0:1	0:2	0:3	0:4
Unordered Stereotype	Bias	0.01	0.03	0.09	0.01	0.03	0.04	0.12	-0.01
Bayes	MSE	0.05	0.19	0.40	0.39	0.06	0.18	0.19	0.27
Unordered Stereotype	Bias	0.01	0.03	0.08	0.13	0.13	0.03	0.10	0.23
MLE	MSE	0.05	0.21	0.41	0.56	0.06	0.18	0.18	0.39
Ordered Stereotype	Bias	-0.01	0.00	0.05	0.12	-0.01	-0.00	0.05	0.17
Bayes	MSE	0.03	0.11	0.28	0.43	0.03	0.04	0.06	0.17

A.2 Chapter III : Supplementary Material

A.2.1 Generalization to $C_i : M_i$ case:control matching

Our results could be directly generalized to the setting of a more general $C_i : M_i$ case:control matching ratio. Let $N_i = C_i + M_i$ be the total number of observations in stratum i. Let C_{ki} be the number of cases in stratum i having response $Y_{ij} = k$, i.e., $C_{ki} = \sum_{j=1}^{N_i} I(Y_{ij} = k)$, $k = 1, \cdots, K$. The conditioning statistic in this case is $\{C_{ki}, k = 1, \cdots, K\}$, the number of cases of each type in stratum i. The conditional likelihood in stratum i is

$$L_i = \frac{\exp(\sum_{k=1}^K \phi_k \boldsymbol{\beta}^T \boldsymbol{S}_{ki})}{\sum_{\{\boldsymbol{q}_{1i}, \cdots, \boldsymbol{q}_{Ki}\} \in \Omega_i} \exp(\sum_{k=1}^K \phi_k \boldsymbol{\beta}^T \boldsymbol{q}_{ki})},$$

where $\boldsymbol{S}_{ki} = \sum_{j=1}^{N_i} \boldsymbol{X}_{ij} I(Y_{ij} = k)$, is the sum of covariate vectors corresponding to the cases of subtype k; and Ω_i is the collection of all possible sum vectors $(q_{1i}, \cdots, q_{Ki})$ of the form $q_{ki} = \boldsymbol{X}_{il_1^k} + \ldots + \boldsymbol{X}_{il_{C_{ki}}^k}$. The indices $(l_1^1, \cdots, l_{C_{1i}}^1, \cdots, l_1^K, \cdots, l_{C_{Ki}}^K)$ is a subset of size $C_i = \sum_{k=1}^K C_{ki}$ chosen from N_i elements to ensure that there are exactly C_{ki} distinct entries contributing to q_{ki}, $k = 1, \cdots, K$. Therefore, the cardinality of Ω_i is $\binom{N_i}{C_{1i}\; C_{2i} \cdots, C_{Ki}}$, the number of ways to assign exactly C_{ki} elements to the k-th category (after this assignment, the remaining $M_i = N_i - C_i$ are automatically controls). Thus, we partition N_i elements into subsets of length $\{C_{1i}, \ldots, C_{Ki}, M_i\}$. The conditional likelihood L_i is simply the probability of having the observed data, given that there are exactly C_{ki} cases having $Y_{ij} = k$ in stratum i and remaining are controls. In a finite population sampling framework, L_i could also be interpreted as the probability that K random samples of size C_{ki} are selected without replacement from the finite population of size N_i with probability proportional to $\exp(\phi_k \boldsymbol{\beta}^\top \boldsymbol{q}_{ki})$ where $\boldsymbol{q}_{ki}$ is the sum of the $\boldsymbol{X}_{ij}$'s selected in sample $k = 1, \cdots, K$.

A.2.2 Lemma 1 and 2: Statement in the general case

Let us denote the prospective model for the disease risk as given in 3.3 of the main text by $P(Y_{ij} = k|\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, S_i)/P(Y_{ij} = 0|\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, S_i) = \rho_k(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, S_i)$, where $\rho_k(\cdot)$ is as described in the RHS of (3.3) through the stereotype model, but could be a non-negative real-valued function, in general. Then Lemmas follow by simply using the definition of conditional probability and algebraic simplification of collected terms.

Lemma 1.

$$f(X_{ij}|\boldsymbol{Z}_{ij}, Y_{ij} = k, S_i) = \frac{f(X_{ij}|\boldsymbol{Z}_{ij}, Y_{ij} = 0, S_i)\rho_k(X_{ij}, \boldsymbol{Z}_{ij}, S_i)}{\int \rho(X_{ij}, \boldsymbol{Z}_{ij}, S_i) f(X_{ij}|\boldsymbol{Z}_{ij}, Y_{ij} = 0, S_i) dX_{ij}}.$$

Lemma 2.

$$\frac{P(Y_{ij} = k|\boldsymbol{Z}_{ij}, S_i)}{P(Y_{ij} = 0|\boldsymbol{Z}_{ij}, S_i)} = \int \rho_k(X_{ij}, \boldsymbol{Z}_{ij}, S_i) f(X_{ij}|\boldsymbol{Z}_{ij}, Y_{ij} = 0, S_i) dX_{ij}.$$

A.2.3 Lemma 1 and 2: Specific version for the exponential family of distributions

Assume that the exposure distribution in the controls belongs to the exponential family, i.e.,

$$f(X_{ij}|S_i, \mathbf{Z}_{ij}, Y_{ij} = 0) = \exp[\xi_{ij}\{\theta_{ij}X_{ij} - b(\theta_{ij})\} + c(\xi_{ij}, X_{ij})].$$

The canonical parameters θ_{ij} are modeled as a regression function of the completely observed covariates $\boldsymbol{Z}_{ij}$ and S_i, namely, $\theta_{ij} = \kappa_0 + \boldsymbol{\kappa}_1^\top \mathbf{Z}_{ij} + \kappa_2 S_i$.

Lemma 1. The distribution of the exposure variable conditional on a disease subclass, $Y = k$, namely, $f(X_{ij}|S_i, \mathbf{Z}_{ij}, Y_{ij} = k)$ is also of general exponential form with scale parameter ξ_{ij} and canonical parameter $\theta^*_{ijk} = \theta_{ij} + \xi_{ij}^{-1}\phi_k\beta_1$ for $k = 1, \cdots, K$.

Lemma 2. Under the same set of assumptions, the marginal odds can be expressed as:

$$\frac{P(Y_{ij} = k|S_i, \mathbf{Z}_{ij})}{P(Y_{ij} = 0|S_i, \mathbf{Z}_{ij})} = \exp\{\beta_{0k}(S_i) + \phi_k \boldsymbol{\beta_2}^\top \mathbf{Z}_{ij}\} \times \exp[\xi_{ij}\{b(\theta^*_{ijk}) - b(\theta_{ij})\}].$$

This Lemma can be adapted to simplify the conditional likelihood specified in (3.7) in the main text.

A.2.4 ECM and Full conditionals when the missing exposure X is assumed to follow the Bernoulli distribution.

We turn our attention to the situation when exposure distribution in the controls arises from a Bernoulli distribution with $p_{ij} = H(\theta_{ij})$ where $\theta_{ij} = \kappa_0 + \boldsymbol{\kappa}_1^\top \boldsymbol{Z}_{ij} + \kappa_2 S_i$, namely,

$$f(X_{ij}|Y_{ij} = 0, \mathbf{Z}_{ij}, S_i) = \exp\left[\theta_{ij} X_{ij} - \log\left\{1 + \exp(\theta_{ij})\right\}\right].$$

We can then express the exposure distribution within sub-types of cases as $f(X_{ij}|Y_{ij} = k_i, \boldsymbol{Z}_{ij}, S_i) = \exp\left[\theta^*_{ij} X_{ij} - \log\left\{1 + \exp(\theta^*_{ij})\right\}\right]$ where $\theta^*_{ij} = \theta_{ij} + \phi_{k_i}\beta_1$ based on Lemma 1 of A.2.3. Using these expressions in (3.7) in the main context, we have, for the Bernoulli case,

$$\begin{aligned}
L_{cm}^{comp} &= \prod_{i=1}^{N}\prod_{j=1}^{M+1}\Big[H(\delta_0 + \delta_1 X_{ij} + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij})^{R_{ij}} \\
&\times \left\{1 - H(\delta_0 + \delta_1 X_{ij} + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij})\right\}^{1-R_{ij}}\Big] \\
&\times \prod_{i:Y_{i1}=k_i}^{N} \exp\Big(\left[\{R_{ij}X_{ij} + (1-R_{ij})X_{ij}\}(\theta_{ij} + \phi_{k_i}\beta_1) - \log\left\{1 + \exp(\theta_{ij} + \phi_{k_i}\beta_1)\right\}\right]\Big) \\
&\times \prod_{i=1}^{N}\prod_{j=2}^{M+1} \exp\Big(\{R_{ij}X_{ij} + (1-R_{ij})X_{ij}\}\theta_{ij} - \log\left\{1 + \exp(\theta_{ij})\right\}\Big) \\
\text{(A.1)} \quad &\times \prod_{i=1}^{N} \frac{\exp\left[\phi_{k_i}\beta_2^\top \boldsymbol{Z}_{i1} + \log\left\{\frac{1+\exp(\theta_{i1}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{i1})}\right\}\right]}{\sum_{j=1}^{M+1}\exp\left[\phi_{k_i}\beta_2^\top \boldsymbol{Z}_{ij} + \log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]}.
\end{aligned}$$

Both the ECM and FB inference are developed based on the above complete data likelihood. Similar expressions for X from a Normal distribution is presented in A.2.5

The ECM:

We can easily calculate the expected complete data log-likelihood in A.2.4 based on the fact that $p(X_{ij} = 1|Y_{ij}, \boldsymbol{Z}_{ij}, S_i, R_{ij} = 0)$ has a Bernoulli distribution with known structure with a mean $H\{\psi_{ij}(\theta)\} = p(X_{ij} = 1|Y_{ij}, \boldsymbol{Z}_{ij}, S_i, R_{ij} = 0)$ where $\psi_{ij}(\theta) = \theta_{ij} + I(Y_{ij} = k_i)\phi_{k_i}\beta_1 + \log\{\bar{\pi}_{ij}(1)/\bar{\pi}_{ij}(0)\}$. Here $\pi_{ij}(s)$ denotes $H(\delta_0 + \delta_1 s + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij})$ and $\bar{\pi}_{ij}(s) = 1 - \pi_{ij}(s)$. We can now express the three terms in (A.1) as:

$$\begin{aligned}
L_1(\Theta^{(t+1)}) &= \sum_{(i,j):R_{ij}=1} \big[X_{ij}\{\theta_{ij} + I(Y_{ij} = k_i)\phi_{k_i}\beta_1\} \\
&- \log\left[1 + \exp\{\theta_{ij} + I(Y_{ij} = k_i)\phi_{k_i}\beta_1\}\right] + \log\{\pi_{ij}(X_{ij})\}\big], \\
E\{L_2(\Theta^{(t+1)})\} &= \sum_{(i,j):R_{ij}=0} (H\{\psi_{ij}(\theta)\}[\theta_{ij} + I(Y_{ij} = k_i)\phi_{k_i}\beta_1 + \log\{\bar{\pi}_{ij}(1)\}]) \\
&+ \sum_{(i,j):R_{ij}=0} \Big([1 - H\{\psi_{ij}(\theta)\}]\log\{\bar{\pi}_{ij}(0)\} - \log\left[1 + \exp\{\theta_{ij} + I(Y_{ij} = k_i)\phi_{k_i}\beta_1\}\right]\Big),
\end{aligned}$$

$$\begin{aligned}
L_3(\Theta^{(t+1)}) &= \sum_{i=1}^{N} \Bigg\{\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{i1} + \log\left\{\frac{1 + \exp(\theta_{i1} + \phi_{k_i}\beta_1)}{1 + \exp(\theta_{i1})}\right\} \\
&- \log\Bigg(\sum_{j=1}^{M+1} \exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \log\left\{\frac{1 + \exp(\theta_{ij} + \phi_{k_i}\beta_1)}{1 + \exp(\theta_{ij})}\right\}\right]\Bigg)\Bigg\}.
\end{aligned}$$

We then follow the ECM steps outlined in Section 3.3.1.

The Bayesian Route:

We can obtain the following full conditional distributions of the model parameters, given the augmented data, by using the likelihood in (A.1) and the prior structure as in (3.12).

$$\begin{aligned}
\pi(\beta_1|\cdot) &\propto \frac{\exp\left(-\frac{1}{2\sigma_{\beta_1}^2}\left[\beta_1 - \mu_{\beta_1} - \sigma_{\beta_1}^2 \sum_{i=1}^N \phi_{k_i}\left\{R_{i1}X_{i1} + (1 - R_{i1})X_{i1}\right\}\right]^2\right)}{\prod_{i=1}^N \sum_{j=1}^{M+1} \exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]}, \\
\pi(\beta_{2r}|\cdot) &\propto \frac{\exp\left\{-\frac{1}{2\sigma_{\beta_2}^2}(\beta_{2r} - \mu_{\beta_{2r}} - \sigma_{\beta_2}^2 \sum_{i=1}^N \phi_{k_i} Z_{i1r})^2\right\}}{\prod_{i=1}^N \sum_{j=1}^{M+1} \exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]}, \\
\pi(\phi_k|\cdot) &\propto \frac{\exp\left\{-\frac{\left(\phi_k - \mu_{\phi_k} - 2\sigma_{\phi_k}^2 \sum_{i=1}^N I(Y_{i1}=k)\left[\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{i1} + \beta_1\{R_{i1}X_{i1} + (1-R_{i1})X_{i1}\}\right]\right)^2}{2\sigma_{\phi_k}^2}\right\}}{\prod_{i=1}^N \sum_{j=1}^{M+1} \exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]},
\end{aligned}$$

$$\begin{aligned}
\pi(\delta_q|\cdot) \quad &\propto \quad \frac{\exp\left\{-\frac{1}{2\sigma_\delta^2}(\delta_q - \mu_{\delta_q} - \sigma_\delta^2 \sum_{i=1}^N \sum_{j=1}^{M+1} R_{ij}V_{ijq})^2\right\}}{\prod_{i=1}^N \prod_{j=1}^{M+1}\left[1+\exp\left\{\delta_0 + \delta_1(R_{ij}X_{ij} + (1-R_{ij})X_{ij}) + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij}\right\}\right]}, \\
\text{where} \quad & \quad V_{ij0} = 1,\ V_{ij1} = Y_{ij},\ V_{ij2} = X_{ij},\ V_{ij3} = S_i. \\
\pi(\delta_{4r}|\cdot) \quad &\propto \quad \frac{\exp\left\{-\frac{1}{2\sigma_\delta^2}(\delta_{4r} - \mu_{\delta_{4r}} - \sigma_\delta^2 \sum_{i=1}^N \sum_{j=1}^{M+1} R_{ij}Z_{ijr})^2\right\}}{\prod_{i=1}^N \prod_{j=1}^{M+1}\left[1+\exp\left\{\delta_0 + \delta_1(R_{ij}X_{ij} + (1-R_{ij})X_{ij}) + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij}\right\}\right]}, \\
\pi(\kappa_0|\cdot) \quad &\propto \quad \frac{\exp\left(-\frac{1}{2\sigma_\kappa^2}\left[\kappa_0 - \mu_{\kappa_0} - \sigma_\kappa^2 \sum_{i=1}^N \sum_{j=1}^{M+1}\left\{R_{ij}X_{ij} + (1-R_{ij})X_{ij}\right\}\right]^2\right)}{\prod_{i=1}^N \sum_{j=1}^{M+1} \exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]} \\
&\times \quad \prod_{i=1}^N \prod_{j=1}^{M+1} \frac{1}{1+\exp(\theta_{ij})}, \\
\pi(\kappa_{1r}|\cdot) \quad &\propto \quad \frac{\exp\left(-\frac{1}{2\sigma_\kappa^2}\left[\kappa_{1p} - \mu_{\kappa_{1p}} - \sigma_\kappa^2 \sum_{i=1}^N \sum_{j=1}^{M+1}\left\{R_{ij}X_{ij} + (1-R_{ij})X_{ij}\right\} Z_{ijr}\right]^2\right)}{\prod_{i=1}^N \sum_{j=1}^{M+1} \exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]} \\
&\times \quad \prod_{i=1}^N \prod_{j=1}^{M+1} \frac{1}{1+\exp(\theta_{ij})},\ r = 1, \ldots, p, \\
\pi(\kappa_2|\cdot) \quad &\propto \quad \frac{\exp\left(-\frac{1}{2\sigma_\kappa^2}\left[\kappa_2 - \mu_{\kappa_2} - \sigma_\kappa^2 \sum_{i=1}^N \sum_{j=1}^{M+1}\left\{R_{ij}X_{ij} + (1-R_{ij})X_{ij}\right\} S_i\right]^2\right)}{\prod_{i=1}^N \sum_{j=1}^{M+1} \exp\left[\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \log\left\{\frac{1+\exp(\theta_{ij}+\phi_{k_i}\beta_1)}{1+\exp(\theta_{ij})}\right\}\right]} \\
&\times \quad \prod_{i=1}^N \prod_{j=1}^{M+1} \frac{1}{1+\exp(\theta_{ij})},
\end{aligned}$$

where $r = 1, \ldots, p,\ k = 1, \ldots, K-1$ and $q = 0, \ldots, 3$. Conditional on the current value of the sampled Θ, we sample X_{ij} from the conditional distribution $p(X_{ij} = 1|Y_{ij}, \boldsymbol{Z}_{ij}, S_i, R_{ij} = 0) = H\{\psi_{ij}(\theta)\}$, where $\psi_{ij}(\theta)$, is exactly as defined in the ECM approach. The Bayesian iterative computation scheme as described in Section 3.3.2 of the main context is then followed.

A.2.5 ECM and Full conditionals when the missing exposure X is assumed to follow a normal distribution.

For the purpose of illustration, in addition to the binomial setting presented above, we extend the ECM and the full Bayesian approaches to the case where the exposure distribution in controls follows a Normal distribution with the mean θ_{ij} and unknown variance σ^2. The canonical mean parameter modeled as $\theta_{ij} = \kappa_0 + \boldsymbol{\kappa}_1^\top \boldsymbol{Z}_{ij} + \kappa_2 S_{ij}$. Thus,

$$f(X_{ij}|S_i, \boldsymbol{Z}_{ij}, Y_{ij} = 0) = \exp\left\{ \frac{(\theta_{ij}X_{ij} - \frac{\theta_{ij}^2}{2})}{\sigma^2} - \frac{X_{ij}^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right\}.$$

Based on Lemma 1 of A.2.3, we can obtain $\theta_{ij}^* = \theta_{ij} + \sigma^2\phi_{k_i}\beta_1$ for the distribution of X_{ij} in disease category $Y_{ij} = k_i$. Now we can develop the complete likelihood in an identical manner as for binary X in the main text, namely,

$$\begin{aligned}
L_{cm}^{comp} &\propto \prod_{i=1}^{N}\prod_{j=1}^{M+1} \Big[H(\delta_0 + \delta_1 X_{ij} + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij})^{R_{ij}} \\
&\times \Big\{1 - H(\delta_0 + \delta_1 X_{ij} + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij})\Big\}^{1-R_{ij}} \Big] \\
&\times \prod_{i:Y_{i1}=k_i}^{N} \exp\left\{ -\frac{R_{ij}(X_{ij} - \theta_{ij} - \sigma^2\beta_1\phi_{k_i})^2}{2\sigma^2} - \frac{(1-R_{ij})(X_{ij} - \theta_{ij} - \sigma^2\beta_1\phi_{k_i})^2}{2\sigma^2} \right\} \\
&\times \prod_{i=1}^{N}\prod_{j=2}^{M+1} \exp\left\{ -\frac{R_{ij}(X_{ij} - \theta_{ij})^2}{2\sigma^2} - \frac{(1-R_{ij})(X_{ij} - \theta_{ij})^2}{2\sigma^2} \right\} \times \exp\left\{ -\frac{N(M+1)}{2}\log(\sigma^2) \right\} \\
&\times \prod_{i=1}^{N} \frac{\exp\left\{\phi_{k_i}\left(\beta_2{}^\top \boldsymbol{Z}_{i1} + \beta_1\theta_{i1}\right)\right\}}{\sum_{j=1}^{M+1}\exp\left\{\phi_{k_i}\left(\beta_2{}^\top \boldsymbol{Z}_{ij} + \beta_1\theta_{ij}\right)\right\}}. \qquad \text{(A.2)}
\end{aligned}$$

The ECM:

Applying logarithm to the complete likelihood (A.2), we now derive $l_{cm}^{comp}(\Theta)$. Here $L_1(\Theta^{(t+1)})$ and $L_3(\Theta^{(t+1)})$ have closed forms as following,

$$\begin{aligned}
L_1(\Theta^{(t+1)}) &= \sum_{(i,j):R_{ij}=1} \left[-\frac{\{X_{ij} - \theta_{ij} - I(Y_{ij} = k_i)\sigma^2\phi_{k_i}\beta_1\}^2}{2\sigma^2} + \log H(\delta_0 + \delta_1 X_{ij} + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij}) \right] \\
L_3(\Theta^{(t+1)}) &= \sum_{i=1}^{N} \left[\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{i1} + \phi_{k_i}\beta_1\theta_{i1} - \log\left\{ \sum_{j=1}^{M+1} (\phi_{k_i}\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \phi_{k_i}\beta_1\theta_{ij}) \right\} \right]. \qquad \text{(A.3)}
\end{aligned}$$

For computing $E\{L_2(\Theta^{(t+1)})\}$, the conditional distribution of unobserved X_{ij} conditioning on $Y_{ij}, \boldsymbol{Z}_{ij}, S_i, R_{ij} = 0$ has no standard closed form with

$$\begin{aligned}
p(X_{ij}|Y_{ij}, \boldsymbol{Z}_{ij}, S_i, R_{ij} = 0) &\propto \left\{1 - H(\delta_0 + \delta_1 X_{ij} + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij})\right\} \\
&\times \exp\left[\frac{\{X_{ij} - \theta_{ij} - I(Y_{ij} = k_i)\sigma^2\phi_{k_i}\beta_1\}^2}{2\sigma^2} \right].
\end{aligned}$$

Due to difficulty in direct sampling from this distribution, we employ the Metropolis-Hasting algorithm to first generate X_{ij} and then evaluate the Monte Carlo average of $E\{L_2(\Theta^{(t+1)})\}$. To proceed, let $X_{ij}^{(l)}$, $l = 1, \dots, L$ be L random samples generated from the aforementioned distribution via the Metropolis-Hasting algorithm, then the resulting expectation $E\{L_2(\Theta^{(t+1)})\}$ is evaluated as

$$E\{L_2(\Theta^{(t+1)})\} = L^{-1} \sum_{(i,j):R_{ij}=0} \sum_{l=1}^{L} \left\{ \log p(X_{ij}^{(l)} | Y_{ij}, \boldsymbol{Z}_{ij}, S_i) + \log p(R_{ij} = 0 | X_{ij}^{(l)}, Y_{ij}, \boldsymbol{Z}_{ij}, S_i) \right\}.$$

At each step of the CM, we first maximize (3.11) as outlined in the main text. Note that in the CM step, here we need to resort to the Monte Carlo approximation for calculating $E\{L_2(\Theta^{(t+1)})\}$ whereas the closed form calculation is available for the binary exposure.

The Full Bayesian approach:

We assume identical normal prior distributions on $\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\kappa}$ as in the main text and an additional prior on σ for the variance of the X as follows.

$$\sigma^2 \overset{iid}{\sim} IG(a, b),$$

The factorization scheme on the complete likelihood (3.11) allows us to obtain the following full conditionals,

$$\begin{aligned}
\pi(\beta_1|\cdot) \quad &\propto \quad \frac{\exp\left\{-\frac{(\beta_1 - \mu_{\beta_1} - \sigma_{\beta_1}^2 \sum_{i=1}^N \theta_{i1}\phi_{k_i})^2}{2\sigma_{\beta_1}^2}\right\}}{\prod_{i=1}^N \sum_{j=1}^{M+1} \exp\left\{\phi_{k_i}(\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \beta_1\theta_{ij})\right\}} \\
&\times \quad \exp\left(-\frac{\sum_{i=1}^N \sigma^2\beta_1\phi_{k_i}\left[\sigma^2\beta_1\phi_{k_i} + 2\theta_{i1} - 2\{R_{i1}X_{i1} + (1 - R_{i1})X_{i1}\}\right]}{2}\right), \\
\pi(\beta_{2r}|\cdot) \quad &\propto \quad \frac{\exp\left\{-\frac{(\beta_{2r} - \mu_{\beta_{2r}} - \sigma_{\beta_2}^2 \sum_{i=1}^N \phi_{k_i} Z_{i1r})^2}{2\sigma_{\beta_2}^2}\right\}}{\prod_{i=1}^N \sum_{j=1}^{M+1} \exp\left\{\phi_{k_i}(\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij} + \beta_1\theta_{ij})\right\}}, \\
\pi(\delta_q|\cdot) \quad &\propto \quad \frac{\exp\left\{-\frac{\left(\delta_q - \mu_{\delta_q} - \sigma_\delta^2 \sum_{i=1}^N \sum_{j=1}^{M+1} R_{ij} V_{ijq}\right)^2}{2\sigma_\delta^2}\right\}}{\prod_{i=1}^N \prod_{j=1}^{M+1}\left[1 + \exp\left\{\delta_0 + \delta_1(R_{ij}X_{ij} + (1 - R_{ij})X_{ij}) + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij}\right\}\right]}, \\
\text{where} \quad & \quad V_{ij0} = 1,\ V_{ij1} = Y_{ij},\ V_{ij2} = X_{ij},\ V_{ij3} = S_i, \\
\pi(\delta_{4r}|\cdot) \quad &\propto \quad \frac{\exp\left\{-\frac{\left(\delta_{4r} - \mu_{\delta_{4r}} - \sigma_\delta^2 \sum_{i=1}^N \sum_{j=1}^{M+1} R_{ij} Z_{ijr}\right)^2}{2\sigma_\delta^2}\right\}}{\prod_{i=1}^N \prod_{j=1}^{M+1}\left[1 + \exp\left\{\delta_0 + \delta_1(R_{ij}X_{ij} + (1 - R_{ij})X_{ij}) + \delta_2 Y_{ij} + \delta_3 S_i + \delta_4^\top \boldsymbol{Z}_{ij}\right\}\right]},
\end{aligned}$$

$$
\begin{aligned}
\pi(\phi_k|\cdot) &\times \frac{\exp\left[-\frac{\{\phi_k-\mu_{\phi_k}-\sigma^2_{\phi_k}\sum_{i=1}^{N} I(Y_{i1}=k)\beta_2{}^\top \boldsymbol{Z}_{i1}+\beta_1\theta_{i1}\}}{2\sigma^2_{\phi_k}}\right]}{\prod_{i=1}^{N}\sum_{j=1}^{M+1}\exp\left\{\phi_{k_i}(\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij}+\beta_1\theta_{ij})\right\}} \\
&\times \exp\left(-\frac{\sum_{Y_{i1}=k}\phi_k\sigma^2\beta_1\left[\sigma^2\beta_1\phi_k+2\theta_{i1}-2\left\{R_{i1}X_{i1}+(1-R_{i1})X_{i1}\right\}\right]}{2\sigma^2}\right), \\
\pi(\kappa_0|\cdot) &\propto \exp\left\{-\frac{(\kappa_0-\mu_{\kappa_0}-\sigma^2_\kappa\sum_{i=1}^{N}\phi_{k_i}\beta_1)^2}{2\sigma^2_\kappa}\right\}\times\frac{\exp\left[-\frac{\sum_{i=1}^{N}\sum_{j=1}^{M+1}\kappa_0\left\{R_{ij}X_{ij}+(1-R_{ij})X_{ij}\right\}}{\sigma^2}\right]}{\prod_{i=1}^{N}\sum_{j=1}^{M+1}\exp\left\{\phi_{k_i}(\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij}+\beta_1\theta_{ij})\right\}} \\
&\times \exp\left[-\frac{\sum_{i=1}^{N}\sum_{j=1}^{M+1}\kappa_0\left\{\kappa_0+2\boldsymbol{\kappa}_1^\top\boldsymbol{Z}_{ij}+2\kappa_2S_i+2I(Y_{ij}=k_i)\sigma^2\beta_1\phi_{k_i}\right\}}{2\sigma^2}\right], \\
\pi(\kappa_{1r}|\cdot) &\propto \exp\left\{-\frac{(\kappa_{1r}-\mu_{\kappa_{1r}}-\sigma^2_\kappa\sum_{i=1}^{N}\phi_{k_i}\beta_1Z_{i1r})^2}{2\sigma^2_\kappa}\right\}\times\frac{\exp\left[-\frac{\sum_{i=1}^{N}\sum_{j=1}^{M+1}\kappa_{1r}Z_{ijr}\left\{R_{ij}X_{ij}+(1-R_{ij})X_{ij}\right\}}{\sigma^2}\right]}{\prod_{i=1}^{N}\sum_{j=1}^{M+1}\exp\left\{\phi_{k_i}(\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij}+\beta_1\theta_{ij})\right\}} \\
&\times \exp\left[-\frac{\sum_{i=1}^{N}\sum_{j=1}^{M+1}\kappa_{1r}Z_{ijr}\left\{\kappa_{1r}Z_{ijr}+2\kappa_o++2\kappa_2S_i+2I(Y_{ij}=k)\sigma^2\beta_1\phi_{k_i}\right\}}{2\sigma^2}\right], \\
\pi(\kappa_2|\cdot) &\propto \exp\left\{-\frac{(\kappa_2-\mu_{\kappa_2}-\sigma^2_\kappa\sum_{i=1}^{N}\phi_{k_i}\beta_1S_i)^2}{2\sigma^2_\kappa}\right\}\times\frac{\exp\left[-\frac{\sum_{i=1}^{N}\sum_{j=1}^{M+1}\kappa_2S_i\left\{R_{ij}X_{ij}+(1-R_{ij})X_{ij}\right\}}{\sigma^2}\right]}{\prod_{i=1}^{N}\sum_{j=1}^{M+1}\exp\left\{\phi_{k_i}(\boldsymbol{\beta}_2^\top \boldsymbol{Z}_{ij}+\beta_1\theta_{ij})\right\}} \\
&\times \exp\left[-\frac{\sum_{i=1}^{N}\sum_{j=1}^{M+1}\kappa_2S_i\left\{\kappa_2S_i+2\kappa_0+2\boldsymbol{\kappa}_1^\top\boldsymbol{Z}_{ij}+2I(Y_{ij}=k_i)\sigma^2\beta_1\phi_{k_i}\right\}}{2\sigma^2}\right]
\end{aligned}
$$

where $r = 1, \ldots, p,\ k = 1, \ldots, K-1$ and $q = 0, \ldots, 3$. The additional scale parameter has a full conditional distribution, $\pi(\sigma^2|\cdot) \propto IG(a^*, b^*)$, where

$$
\begin{aligned}
a^* &= a + \frac{N(M+1)}{2}, \\
b^* &= b + \frac{\sum_{i=1}^{N}\left\{R_{i1}(X_{i1}-\theta^*_{i1})^2+(1-R_{i1})(X_{i1}-\theta^*_{i1})^2\right\}}{2} \\
&+ \frac{\sum_{i=1}^{N}\sum_{j=2}^{M+1}\left\{R_{ij}(X_{ij}-\theta_{ij})^2+(1-R_{ij})(X_{ij}-\theta_{ij})^2\right\}}{2}.
\end{aligned}
$$

Conditional on the current values of the sampled Θ, we first sample X_{ij} from the conditional distribution (A.3) via the Metropolis-Hasting algorithm. Given the augmented data, we generate samples of $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\kappa}, \boldsymbol{\delta}, \sigma)$ from the aforementioned full conditionals. The remaining Bayesian iterative computation scheme as described in Section 3.3.2 of the main text is then followed.

Table A.5: Fitted model parameters and different information criteria for a model comparison between the polytomous logistic model and the stereotype model with complete case data† from the MECC Study. Here we used case-control status with cancer Stages recorded for the cases as response with 4 categories. We use physical activity, use of statins, use of NSAID, vegetable consumption and family history of colorectal cancer as the set of explanatory variables. AIC, BIC, and DIC represent the Akaike Information Criterion, the Bayes Information Criterion, and the Deviance Information Criterion respectively. The AIC and the BIC results are calculated when both models are fitted under maximum likelihood framework and the DIC results correspond to a Bayesian estimation of the parameters in both models.

		Polytomous logistic model			Stereotype model		
		Estimates(SE/PSD ‡)			Estimates(SE/PSD ‡)		
Type	Coefficient	0:1	0:2	0:3	0:1	0:2	0:3
Maximum Likelihood Approach	Intercept	-1.55(0.13)	-0.41(0.09)	-0.37(0.09)	-1.47(0.12)	-0.39(0.09)	-0.39(0.09)
	β_{sports}	-0.02(0.16)	-0.46(0.12)	-0.26(0.12)		-0.35(0.10)	
	$\beta_{statins}$	-0.27(0.27)	-0.88(0.25)	-0.48(0.21)		-0.64(0.18)	
	$\beta_{history}$	0.69(0.22)	0.19(0.20)	0.28(0.19)		0.21(0.16)	
	β_{NSAID}	-0.04(0.19)	-0.47(0.15)	-0.44(0.15)		-0.44(0.12)	
	$\beta_{vegetable}$	-0.17(0.09)	-0.22(0.07)	-0.35(0.07)		-0.27(0.07)	
	(ϕ_1, ϕ_2, ϕ_3)	-	-	-	0.40(0.19)	1.06(0.19)	1.00
Goodness of fit	AIC		5442.9			5432.0	
	BIC		5546.0			5499.2	
Bayesian Approach	Intercept	-1.54(0.13)	-0.41(0.09)	-0.37(0.09)	-1.46(0.11)	-0.44(0.09)	-0.35(0.09)
	β_{sports}	-0.03(0.16)	-0.46(0.12)	-0.26(0.11)		-0.36(0.10)	
	$\beta_{statins}$	-0.28(0.27)	-0.91(0.25)	-0.50(0.21)		-0.64(0.18)	
	$\beta_{history}$	0.68(0.22)	0.19(0.20)	0.28(0.19)		0.23(0.17)	
	β_{NSAID}	-0.07(0.19)	-0.49(0.14)	-0.45(0.16)		-0.47(0.12)	
	$\beta_{vegetable}$	-0.18(0.10)	-0.22(0.06)	-0.36(0.06)		-0.31(0.06)	
	(ϕ_1, ϕ_2, ϕ_3)	-	-	-	0.40(0.14)	0.89(0.12)	1.00
Goodness of fit	DIC		5442.4			5431.3	

†The complete data consisted of 1,134 matched pairs with 2,268 subjects.

‡For maximum likelihood estimates, asymptotic standard error (SE) based on inverse of the observed Fisher information is provided. For Bayesian approaches, posterior standard deviation (PSD) is provided.

Table A.6: Coverage probabilities of the Wald-type confidence intervals (for CMLE and ECM) and the HPD intervals for FB. The nominal coverage probabilities should be 95%. The simulation settings are identical to that of Table 3.2 in the main text. Here, binary exposure $X|Z,S$ is generated from $P(X=1|Z,S) = H(0.3+0.3Z-1.5S)$. The CMLE, the ECM and the FB methods are considered. The results are based on 200 simulated datasets, each with 1,000 cases and 1,000 controls. For each parameter of interest in the disease risk model, we report the percentage of intervals including the true parameter value out of the 200 replications. The true values for the parameters of interest are: $\beta_1 = -0.3$, $\beta_2 = -0.7$, $\phi_1 = 0.8$, and $\phi_2 = 1.7$.

	Method		
	CMLE	ECM	FB
Parameter	Coverage	Coverage	Coverage
MM1. $logit\{p(R_{ij}=1|Y_{ij},X_{ij},Z_{ij},S_i)\}$=*0.8*			
β_1	0.945	0.945	0.935
β_2	0.930	0.940	0.940
ϕ_1	0.925	0.955	0.945
ϕ_2	0.910	0.965	0.935
MM4. $logit\{p(R_{ij}=1|Y_{ij},X_{ij},Z_{ij},S_i)\}$=$0.5X_{ij}+0.5Y_{ij}+0.5$			
β_1	0.855	0.925	0.915
β_2	0.910	0.930	0.940
ϕ_1	0.950	0.940	0.945
ϕ_2	0.925	0.965	0.945
MM5. $logit\{p(R_{ij}=1|Y_{ij},X_{ij},Z_{ij},S_i)\}$=$X_{ij}Z_{ij}+Y_{ij}X_{ij}+1$			
β_1	0.725	0.685	0.710
β_2	0.900	0.910	0.895
ϕ_1	0.890	0.945	0.910
ϕ_2	0.935	0.835	0.850

A.3 Chapter IV : Supplementary Material

Table A.7: The contingency table contains frequency with respect to two SNPs on two genes, $RS762551$ on CYP1A2 (G_1) and $RS1056836$ on CYP1B1 (G_2), statins (E), and disease status. The frequencies are based on completely observed 2,354 subjects at phase II

	CYP1A2(RS762551)						CYP1B1(RS1056836)				
	Control		Case				Control		Case		
	E=0	E=1	E=0	E=1	Total		E=0	E=1	E=0	E=1	Total
G=0(A/A)	410	118	513	38	1089	(C/C)	282	75	345	20	744
G=1(A/C)	441	86	473	50	1036	(G/C)	460	100	528	53	1128
G=2(C/C)	99	21	92	13	229	(G/G)	208	50	205	28	483
Total	950	225	1078	101	2354		950	225	1078	101	2354

A.3.1 Dunson and Xing (2007) Update

We describe the posterior sampling steps in relation to parameters in $P(\boldsymbol{W}|\boldsymbol{\theta})$, $\boldsymbol{\theta} = \{\boldsymbol{\psi}, \boldsymbol{V}, \alpha\}$, by following Dunson and Xing (2009). They introduce a vector of latent variables $\boldsymbol{u} = \{u_1, \ldots, u_N\}$, $u_u > 0$. The joint distribution of $\boldsymbol{u}, \boldsymbol{w}|\boldsymbol{V}, \boldsymbol{\psi}, \alpha$ is defined as,

$$\prod_{u=1}^{N} \left\{ \sum_{h \in A_{u\nu}} \prod_{j=1}^{p} \prod_{l=1}^{d_j} \psi_{hl}^{(j)\,I(w_{uj}=l)} \right\}, \tag{A.4}$$

where $A_{u\nu} = \{h : \nu_h > z_u\}$ and $\nu_h = V_h \prod_{l<h}(1 - V_l)$. The joint posterior is then the product of the augmented data likelihood (A.4) and respective priors on $\boldsymbol{V}, \boldsymbol{\psi}, \alpha$. The augmented data Gibbs sampling is based on the following steps.

(a) We start with the simplest one. Update u_u for $u = 1, \ldots, N$, by sampling from $U(0, \nu_{z_u})$

(b) Next step is regarding $\boldsymbol{\psi}_h^{(j)}$. Note that we can find $h^* = max\{z_1, \ldots, z_N\}$ such that we do not need to compute any conditionals $h > h^*$ later on. And we notice that

$$\pi(\boldsymbol{\psi}_h^{(j)}|\cdot) \propto Dirichlet(a_{j1}, \ldots, a_{jd_j}) \times \prod_{u=1}^{N} \prod_{l=1}^{d_j} \psi_{hl}^{(j)\,I(w_{uj}=l)}.$$

Due to the conjugate prior, the posterior conditional for j-th response when $z_u = h$ is given as

$$Dirichlet\left(a_{j1} + \sum_{u:z_u=h} I(w_{uj} = 1), \ldots, a_{jd_j} + \sum_{u:z_u=h} I(w_{uj} = dj) \right).$$

(c) The conditionals with respect to V_h is

$$\pi(V_h|\cdot) \propto (1-V_h)^{\alpha-1} \times \prod_{u=1}^{N} I(u_u < V_h) \prod_{l<h}(1-V_l).$$

If we focus on the latter part, we can obtain a beta$(1, \alpha)$ distribution truncated at

$$\left[max_{u:z_u=h}\left\{\frac{u_u}{\prod_{l<h}(1-V_l)}\right\}, 1 - max_{u:z_u>h}\left\{\frac{u_u}{V_{z_u}\prod_{l<z_u, l\neq h}(1-V_l)}\right\}\right].$$

(d) Update z_u from the multinomial full conditional as given,

$$Pr(z_u = h|\cdot) = \frac{I(\nu_h > u_u)\prod_{j=1}^{p}\psi_{hw_{uj}}^{(j)}}{\sum_{l\in A_{u\nu}}\prod_{j=1}^{p}\psi_{lw_{uj}}^{(j)}}, \qquad u = 1, \ldots, N.$$

As discussed in Walker (2007), we will be in trouble without latent variables $\boldsymbol{u}$ in that the choice of z_u can be infinite. Since the number of subjects in the dataset itself is finite, the cardinality of the set $A_{i\nu}$ is finite. To validate this argument, we need to find the smallest k^* such that $\sum_{h=1}^{k^*}\nu_h > 1 - \min\{u_1, \ldots, u_N\}$. Noticing that $\sum_{h=1}^{\infty}\nu_h$ is monotonically increasing and $\sum_{h=1}^{\infty}\nu_h = 1$, we can compute the desired k^*.

(e) In the last step, we update α from

$$Gamma\left(a_\alpha + h^*, b_\alpha - \sum_{h=1}^{h^*}\log(1-V_h)\right)$$

Above steps are equivalent to Dunson and Xing (2009) except we set the maximum of the number of mixtures k such that $argmin\sum_{h=1}^{k}\nu_h > 0.99$ to avoid possible large number of mixtures, theoretically up to the data size N. This can lead to the bias which has little influence overall. In such a situation, we adjust V_h and ν_h to satisfy $\sum_h{}^k\nu_h = 1$. The practical gain from this method is that we can resort to known posterior sampling distributions and can anticipate facilitating the process.

Table A.8: Simulation results under two scenarios 1) $G_1 \perp E$, $G_1 \perp G_2$, and $G_2 \perp E$ association, 2) $G_1 \perp E$, $G_1 \sim G_2$, and $G_2 \sim E$. The results are based on 200 replicated datasets, each with 1,000 cases and 1,000 controls in phase I and 800 cases and 800 controls in phase II. The approaches listed, TPFB, TPFB$_{\mathrm{emp}}$, WL, PL, UML, CML, and EB where each represents Two-phase full Bayes (with empirically obtained prior variance), Weighted likelihood, Pseudolikeliohod, Unconstrained Maximum likelihood, Constrained Maximum Likelihood, and Empirical Bayes respectively. The CML imposes G_1-E and G_1-G_2 independence, however, no constraints on G_2-E association. We set $(\beta_E, \beta_{G_1G_2}, \beta_{G_1E}, \beta_{G_2E}) = (-1.5, 0, \log(2), \log(2))$ for all scenarios. In terms of MCMC chains from both TPFBs, the posterior results based on 1,000 samples after 'burn-in' of 40,000 among 50,000 iterations and a thinning of every 10 samples.

		$G_1 \perp E$, $G_1 \perp G_2$, $G_2 \perp E$				$G_1 \perp E$, $G_1 \sim G_2$, $G_2 \sim E$			
Stratified sampling (a)†		E	G_1 x G_2	G_1 x E	G_2 x E	E	G_1 x G_2	G_1 x E	G_2 x E
		$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (0,0,0)$				$(\lambda_{G_1G_2}, \lambda_{G_1E}, \lambda_{G_2E}) = (\log(2), 0, \log(1.5))$			
TPFB	Bias	0.019	0.011	-0.022	-0.070	-0.116	0.232	-0.018	0.190
	MSE	0.156	0.069	0.199	0.187	0.216	0.123	0.183	0.220
TPFB$_{\mathrm{emp}}$	Bias	0.019	0.038	-0.086	-0.109	-0.086	0.034	-0.048	0.126
	MSE	0.156	0.038	0.208	0.178	0.204	0.081	0.136	0.216
WL	Bias	-0.030	-0.009	0.049	0.030	-0.074	0.006	0.047	0.071
	MSE	0.192	0.099	0.290	0.264	0.260	0.094	0.243	0.270
PL	Bias	-0.030	-0.010	0.050	0.030	-0.074	0.004	0.047	0.071
	MSE	0.192	0.097	0.289	0.264	0.260	0.094	0.243	0.270
UML	Bias	-0.049	-0.010	0.050	0.030	-0.095	0.004	0.047	0.071
	MSE	0.196	0.097	0.289	0.264	0.267	0.094	0.243	0.270
CML	Bias	-0.042	-0.007	0.042	0.009	-0.080	0.705	0.023	0.061
	MSE	0.170	0.045	0.174	0.243	0.231	0.542	0.165	0.253
EB	Bias	-0.044	-0.010	0.042	0.014	-0.083	0.112	0.035	0.062
	MSE	0.175	0.060	0.205	0.249	0.235	0.116	0.179	0.254

†All subjects with $E = 1$ in case and control are sub-sampled for phase II.
TPFB uses the informative prior $N(0, 10^{-2})$ on G-G and G-E associations in the model (4.2)
TPFB$_{\mathrm{emp}}$ uses the prior $N(0, \hat{\theta}^2)$ on G-G and G-E associations in the model (4.2) where
$\hat{\theta}^2$ is empirically estimated G-G or G-E association parameter under controls.

A.4 Chapter V : Supplementary Material

A.4.1 The Log Gaussian Cox Process

In this section, we review the log Gaussian Cox process. Suppose that $\mathbf{\Lambda} = \{\Lambda(s) : s \in S\}$ in a bounded study window $S \subseteq \mathbb{R}^d$ and $\mathbf{\Lambda}$ is a nonnegative random field and is almost surely locally integrable, i.e., $\int_S \Lambda(s)ds < \infty$. When the conditional distribution of $\boldsymbol{X}$ given $\mathbf{\Lambda}$ is a Poisson process with the intensity $\mathbf{\Lambda}$, we call X a Cox process driven by $\mathbf{\Lambda}$.

When a Cox process $\boldsymbol{X}$ on $\mathbb{R}^d$ is driven by the intensity process $\mathbf{\Lambda} = \exp(\boldsymbol{Y})$ where $\boldsymbol{Y} = \{Y(s) : s \in S\}$ is a real valued Gaussian random field (GRF), we call $\boldsymbol{X}$ a log Gaussian Cox process (LGCP). It is well documented that the distribution of $\boldsymbol{Y}$, as well as $\boldsymbol{X}$, is determined by its mean $\mu(s) = \mathbb{E}Y(s),\ s \in S$ and covariance $\mathbb{C}ov(Y(s_1), Y(s_2)),\ s_1, s_2 \in S$ of $\boldsymbol{Y}$.

The covariance function $c(s_1, s_2) = \mathbb{C}ov(Y(s_1), Y(s_2)),\ s_1, s_2 \in S$ is frequently assumed to be translation invariant and isotropic. For example, the covariance is commonly assumed to be a function of the distance $d = \|s_1 - s_2\|$, i.e., $c(s_1, s_2) = \sigma^2 r(d)$. Here, $\sigma^2 = Var(Y(s))$, is the marginal variance of the process and $r(d)$ is the correlation function. A commonly employed correlation function is the power exponential correlation function:

$$r(d) = \exp(-\alpha \parallel d \parallel^k), \qquad \alpha > 0, \qquad 0 < k \leq 2,$$

where α and k play key roles in characterizing the smoothness of the intensity function.

For longitudinal data across time points, we can postulate a Cox process $\boldsymbol{X}_t,\ t = 1, \ldots, T$ at each time t, which is driven by a time specific intensity process $\boldsymbol{\Lambda}_t = \{\Lambda_t(s) : s \in S\}, t = 1, \ldots, T$. Then, $\Lambda_t(s) = \exp(Y_t(s)),\ t = 1, \ldots, T$, provided that the corresponding Gaussian random fields $\boldsymbol{Y}_t = \{Y_t(s) : s \in S\}, t = 1, \ldots, T$, are well-defined. The intensity can easily accommodate temporally and spatially referenced covariates $V(t),\ t = 1, \ldots, T$ and $W(s),\ s \in S$ by specifying

$$\Lambda_t(s) = \exp(Y_t(s) + \beta_1 V(t) + \beta_2 W(s)).$$

A.4.2 Full Conditionals

We describe an approximated posterior distribution in stage I,

$$\pi(\boldsymbol{\Omega}_1 | \boldsymbol{x}_S) \ \propto \ \prod_{t=1}^{7} \prod_{c=1}^{21} \Big[\exp \Big\{ \sum_{\xi \in S_c} \big(\log(\widetilde{\Lambda_{c,t}}(\xi)) n_{c,t}(\xi) - |S_c(\xi)| \widetilde{\Lambda_{c,t}}(\xi) \big) \Big\} \Big] p(\boldsymbol{\Omega}_1),$$

where we replace the integral in (5.1) by a sum using the approximation, $\int_{\xi \in S_c} \exp(\widetilde{\Lambda_{c,t}}(\xi)) d\xi \approx \sum_{\xi \in S_c} \exp(\widetilde{\Lambda_{c,t}}(\xi)) |S_c(\xi)|$. Here $|S_c(\xi)|$ represents the area of a single cell in S_c and $n_{c,t}(\xi)$ is the number of observed cases in a single cell ξ and $p(\boldsymbol{\Omega}_1)$ prior for $\boldsymbol{\Omega}_1$.

The full conditionals for each parameter are listed below,

$$
\begin{aligned}
\pi(\Gamma_{c,t}^{ext}|x,\cdot) &\propto \exp\Big(\Big[\sum_{\xi\in S_c}\big(\log(\widetilde{\Lambda_{c,t}}(\xi))n_{c,t}(\xi)-|S_c(\xi)|\widetilde{\Lambda_{c,t}}(\xi)\big)-\parallel\Gamma_{c,t}^{ext}\parallel^2/2\Big]\Big),\\
\pi(\sigma_t^2|x,\cdot) &\propto \frac{1}{\sigma_t^2}\exp\Big[\sum_{t=1}^{7}\sum_{c=1}^{21}\Big\{\sum_{\xi\in S_c}\big(\log(\widetilde{\Lambda_{c,t}}(\xi))n_{c,t}(\xi)-|S_c(\xi)|\widetilde{\Lambda_{c,t}}(\xi)\big)\Big\}-\frac{(\log\sigma_t^2)^2}{2\times10^4}\Big],\\
\pi(\eta|x,\cdot) &\propto \exp\Big[\sum_{t=1}^{7}\sum_{c=1}^{21}\Big\{\sum_{\xi\in S_c}\big(\log(\widetilde{\Lambda_{c,t}}(\xi))n_{c,t}(\xi)-|S_c(\xi)|\widetilde{\Lambda_{c,t}}(\xi)\big)\Big\}-\frac{\eta^2}{2\times10^4}\Big],\\
\pi(\beta_1|x,\cdot) &\propto \exp\Big[\sum_{t=1}^{7}\sum_{c=1}^{21}\Big\{\sum_{\xi\in S_c}\big(\log(\widetilde{\Lambda_{c,t}}(\xi))n_{c,t}(\xi)-|S_c(\xi)|\widetilde{\Lambda_{c,t}}(\xi)\big)\Big\}-\frac{{\beta_1}^2}{2\times10^4}\Big],\\
\pi(\beta_2|x,\cdot) &\propto \exp\Big[\sum_{t=1}^{7}\sum_{c=1}^{21}\Big\{\sum_{\xi\in S_c}\big(\log(\widetilde{\Lambda_{c,t}}(\xi))n_{c,t}(\xi)-|S_c(\xi)|\widetilde{\Lambda_{c,t}}(\xi)\big)\Big\}-\frac{{\beta_2}^2}{2\times10^4}\Big],\\
\pi(\mu_c^{com}|x,\cdot) &\propto \exp\Big[\sum_{t=1}^{7}\sum_{c=1}^{21}\Big\{\sum_{\xi\in S_c}\big(\log(\widetilde{\Lambda_{c,t}}(\xi))n_{c,t}(\xi)-|S_c(\xi)|\widetilde{\Lambda_{c,t}}(\xi)\big)\Big\}-\frac{(\mu_c^{com}-\mu_{com})^2}{2\sigma_{com}^2}\Big],\\
\pi(\mu_t^{time}|x,\cdot) &\propto \exp\Big[\sum_{t=1}^{7}\sum_{c=1}^{21}\Big\{\sum_{\xi\in S_c}\big(\log(\widetilde{\Lambda_{c,t}}(\xi))n_{c,t}(\xi)-|S_c(\xi)|\widetilde{\Lambda_{c,t}}(\xi)\big)\Big\}-\frac{(\mu_t^{time})^2}{2\times10^4}\Big],\\
\pi(\mu_{com}|x,\cdot) &\sim N(\frac{20\hat{\mu}}{21},\frac{2\times10^5}{21}),\\
\pi(\sigma_{com}^2|x,\cdot) &\sim IG(21/2+1/2+0.1,\frac{(\mu_{com})^2}{40}+\frac{\sum_{c=1}^{21}(\mu_c^{com})^2}{2}+0.1),
\end{aligned}
$$

where $\parallel\cdot\parallel$ is the Euclidean norm and $\hat{\mu}=\sum_{c=1}^{21}\mu_c^{com}/21$ and $1/21$ is a shrinkage factor.

The full conditionals for stage II parameters are,

$$
\begin{aligned}
\pi(\tau_{0t}|\cdot) &\propto \frac{1}{|\boldsymbol{\Phi}^t(\cdot|\rho_t^2,\phi)|}\exp\Big\{-\frac{(\boldsymbol{I}^t-\boldsymbol{\mu}^t(\boldsymbol{\tau}))^\top\boldsymbol{\Phi}^t(\cdot|\rho_t^2,\phi)^{-1}(\boldsymbol{I}^t-\boldsymbol{\mu}^t(\boldsymbol{\tau}))}{2}-\frac{(\tau_{0t})^2}{2\times10^4}\Big\},\\
\pi(\tau_1|\cdot) &\propto \prod_t\frac{1}{|\boldsymbol{\Phi}^t(\cdot|\rho_t^2,\phi)|}\exp\Big\{-\frac{(\boldsymbol{I}^t-\boldsymbol{\mu}^t(\boldsymbol{\tau}))^\top\boldsymbol{\Phi}^t(\cdot|\rho_t^2,\phi)^{-1}(\boldsymbol{I}^t-\boldsymbol{\mu}^t(\boldsymbol{\tau}))}{2}-\frac{(\tau_1)^2}{2\times10^4}\Big\},\\
\pi(\rho_t^2|\cdot) &\propto \frac{1}{\rho_t^2|\boldsymbol{\Phi}^t(\cdot|\rho_t^2,\phi)|}\exp\Big\{-\frac{(\boldsymbol{I}^t-\boldsymbol{\mu}^t(\boldsymbol{\tau}))^\top\boldsymbol{\Phi}^t(\cdot|\rho_t^2,\phi)^{-1}(\boldsymbol{I}^t-\boldsymbol{\mu}^t(\boldsymbol{\tau}))}{2}-\frac{(\log\rho_t^2)^2}{2\times10^4}\Big\},\\
\pi(\phi|\cdot) &\propto \prod_t\frac{1}{\phi|\boldsymbol{\Phi}^t(\cdot|\rho_t^2,\phi)|}\exp\Big\{-\frac{(\boldsymbol{I}^t-\boldsymbol{\mu}^t(\boldsymbol{\tau}))^\top\boldsymbol{\Phi}^t(\cdot|\rho_t^2,\phi)^{-1}(\boldsymbol{I}^t-\boldsymbol{\mu}^t(\boldsymbol{\tau}))}{2}-\frac{(\log\phi)^2}{2\times10^4}\Big\},
\end{aligned}
$$

where $t=1,\ldots,7$. $\boldsymbol{I}^t=(\boldsymbol{I}_S^t,\boldsymbol{I}_U^t)^\top$ and $\boldsymbol{\mu}^t(\boldsymbol{\tau})=(\boldsymbol{\mu}_S^t(\boldsymbol{\tau}),\boldsymbol{\mu}_U^t(\boldsymbol{\tau}))^\top$.

A.4.3 Posterior sampling strategy

We describe the posterior sampling steps for parameters in stage I. Since full conditionals do not follow closed form standard distributions, we consider the Metropolis-within-Gibbs algorithm. We employ a random walk proposal for each parameter except $\mathbf{\Gamma}^{ext} = \{\Gamma^{ext}_{c,t}; c = 1, \ldots, 21, t = 1, \ldots, 7\}$. In order to accelerate the sampling process and to have well-mixed samples, we adaptively tune the variances of the random walk proposals respectively until desired acceptance rates are achieved.

We repeat the Metropolis-within-Gibbs sampling steps regarding the first stage parameters, $(\boldsymbol{\mu}^{com}, \boldsymbol{\mu}^{time}, \mu_{com}, \sigma^2_{com}, \boldsymbol{\sigma}^2, \eta, \beta_1, \beta_2)^\top_{39} \equiv \mathbf{\Omega}_1 = (\omega_1, \ldots, \omega_{39})^\top$ and $\mathbf{\Gamma}^{ext}$ given below.

(i) Start the algorithm by setting initial values for each parameter in $\mathbf{\Omega}_1$ and $\mathbf{\Gamma}^{ext}$. The initial vector of $\Gamma^{ext}_{c,t}$ for community c and cycle t is drawn from $N_{2K_{1c} \times 2K_{2c}}(0, I)$ for $c = 1, \ldots, 21,\ t = 1, \ldots, 7$. We also set proposal variance h_{ω_j} corresponding to $\omega_j, j = 1, \ldots, 39$.

(ii) At the n-th iteration, we use the Metropolis-adjusted Langevin algorithm (MALA, Besag 1994) for sampling $\mathbf{\Gamma}^{ext n+1}$. Details are in Møller *et al.* (1998).

(iii) At the n-th iteration, we sample a candidate value from $\omega^*_j \sim N(\omega^n_j, h_{\omega_j})$ where ω^n_j is the value from the previous iteration.

(iv) When $\mathbf{\Omega}^n_{(-j)} = (\omega^{n+1}_1, \ldots, \omega^{n+1}_{j-1}, \omega^n_{j+1}, \ldots, \omega^n_{39})$, the log posterior ratio or log acceptance probability is computed:

$$\log(p_a) = l(\omega^*_j | \mathbf{\Omega}^n_{(-j)}, \cdot) - l(\omega^n_j | \mathbf{\Omega}^n_{(-j)}, \cdot),$$

where $l({\omega^n}_j | \mathbf{\Omega}^n_{(-j)}, \cdot)$ is the full conditional for ω^n_j on the log scale.

(v) Generate $U \sim Unif[0, 1]$. $\omega^{n+1}_j = \omega^*_j$ if $\log U < \log(p_a)$, otherwise $\omega^{n+1}_j = \omega^n_j$.

(vi) Iterate steps (iii) through (v) for $j = 1, \ldots, 39$.

After simulating the stage I posterior, we calculate $\boldsymbol{I}_S^t$. Then we proceed to the stage II sampling.

(i) Set initial values for parameters for $\boldsymbol{\Omega}_2 = (\omega_1, \ldots, \omega_{15})^\top$, $\boldsymbol{I}_U^t$, and $N(\boldsymbol{x}_{U,t})$ for $t = 1, \ldots, 7$. We also set h_{ω_j}, the proposal variance for ω_j.

(ii) At the n-th iteration, sampling of ${\boldsymbol{\Omega}_2}^{n+1}$:

Since $\omega_j, j = 1, \ldots, 15$, does not follow any standard distribution, we use a random walk Metropolis-Hasting algorithm. Sample ω_j^* from $N(\omega_j^n, h_{\omega_j})$.

Given ω_j^*, the acceptance ratio p_a is calculated as

$$p_a = \min\Big\{1, \frac{\pi(\omega_j^*|\cdot)}{\pi(\omega_j^n|\cdot)}\Big\},$$

where $\pi(\omega_j^*|\cdot)$ corresponds to the full conditional of ω_j^*. With probability p_a, we update $\omega_j^{n+1} = \omega_j^*$. We repeat this for $j = 1, \ldots, 15$.

(iii) At the n-th iteration, sampling of ${\boldsymbol{I}_U^t}^{n+1}$:

We sample based on the Metropolis-Hasting algorithm. For the proposal density, we use $[{\boldsymbol{I}_U^t}^*|\boldsymbol{I}_S^t, \boldsymbol{\Omega}_2]$ which follows a conditional multivariate normal distribution with mean $\boldsymbol{\mu}_U^t(\boldsymbol{\tau}^{n+1}) + {\boldsymbol{\Phi}_{S,U}^t}^\top((\rho_t^2, \phi)^{n+1}){\boldsymbol{\Phi}_S^t}^{-1}((\rho_t^2, \phi)^{n+1})\ \{\boldsymbol{I}_S^t - \boldsymbol{\mu}_S^t(\boldsymbol{\tau}^{n+1})\}$ and variance $\boldsymbol{\Phi}_U^t((\rho_t^2, \phi)^{n+1})\ -{\boldsymbol{\Phi}_{S,U}^t}^\top((\rho_t^2, \phi)^{n+1}){\boldsymbol{\Phi}_S^t}^{-1}((\rho_t^2, \phi)^{n+1})\ \boldsymbol{\Phi}_{S,U}^t((\rho_t^2, \phi)^{n+1})$. Then, the acceptance ratio p_a is calculated as

$$p_a = \min\left\{1, \frac{\prod_{l=1}^{106}\left[\exp\left\{-pop_{l,t}^n \exp({I_{U_l}^t}^*)\right\}\{pop_{l,t}^n \exp({I_{U_l}^t}^*)\}^{n(x_{U_l,t}^n)}\right]}{\prod_{l=1}^{106}\left[\exp\left\{-pop_{l,t}^n \exp({I_{U_l}^t}^n)\right\}\{pop_{l,t}^n \exp({I_{U_l}^t}^n)\}^{n(x_{U_l,t}^n)}\right]}\right\},$$

We update ${\boldsymbol{I}_U^t}^{n+1} = {\boldsymbol{I}_U^t}^*$ if $U < p_a$ when $U \sim Unif[0,1]$. We repeat this step for $t = 1, \ldots, 7$.

(iv) We draw $N(x_{U_l,t}^{n+1})$ for unsampled communities using a Poisson distribution with mean $pop_{l,t}^n \exp({I_{U_l}^t}^{n+1})$.

(v) Iterate (ii)-(iv) until convergence.

REFERENCES

REFERENCES

Agresti, A. (2002). Categorical data analysis (2nd ed.). New York: John Wiley and Sons.

Ahn, J., Mukherjee, B., Banerjee, M., and Cooney, K. A. (2009). Bayesian Inference for the Stereotype Regression Model: Application to a Case-control Study of Prostate Cancer. *Statistics in medicine* **28**, 3139-3157.

American Joint Committee on Cancer (2002).: AJCC Cancer Staging Manual. 6th ed. New York, NY: Springer, 113–124.

Amundadottir, L. *et al.* (2009). Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet.* **41** 986-990.

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

Anderson, J. A. (1984). Regression and ordered categorical variable. *J. R. Stat. Soc. B.* **46**, 1-30.

Ashby, D., Hutton, J. L. and McGee, M. A. (1993). Simple Bayesian analyses for casecontrolled studies in cancer epidemiology. *Statistician* **42**, 385-389.

Baddeley. A. J., Møller, J. and Waagepetersen, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica* **54**, 329-350.

Baddeley, A. and Turner, R. (2005). spatstat: An R Package for Analyzing Spatial Point Patterns *Journal of Statistical software* **12**, *Issue 6.*

Barreto, M. L., Genser, B., Strina, A., Teixeira, M. L., Assis, A. M. O., Rego, R. F., Teles, C. A., Prado, M. S., Matos, S. M. A., Santos, D. N., Santos, L, A., Cairncross, S. (2007). Effect of city-wide sanitation programme on reduction in rate of childhood diarrhoea in northeast Brazil: assessment by two cohort studies. *Lancet* **370**, 1622-1628.

Bates, S. J., Trostle, J., Cevallos, W. T., Hubbard, A., Eisenberg, J. N. O. (2007). Relating Diarrheal Disease to Social Networks and the Geographic Configuration of Communities in Rural Ecuador. *American Journal of Epidemiology* **166**, 1088-1095.

Beneš, V., Bodlák, K., Møller, J., and Waagepetersen, R. (2005). A case study on point process modelling in disease mapping. *Image Anal. Stereol.* **24**. 159-168.

Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis, (2nd ed.) New York: Springer-Verlag.

Besag, J. E. (1994). Discussion on 'Representations of knowledge in complex systems' by Grenander and Miller. *Journal of the Royal Statistical Society, B* **56**. 591-592.

Bhattacharya, A. and Dunson, D.B. (2011). Simplex factor models for multivariate unordered categorical data. Journal Amer. Stat. Assoc. Revision invited.

Bhattacharjee, S., Chatterjee, N. and Wheeler, W. (2011). Package CGEN, Version 1.0.0, An R package for analysis of case-control studies in genetic epidemiology. `http://dceg.cancer.gov/bb/tools/genetanalcasecontdata`.

Breslow, N. E. and Day, N. E. (1980). Statistical Methods in Cancer Research: Vol. 1 - The Analysis of Case-Control Studies. Lyon, France, IARC Scientific Publications.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L., and Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology* **108**, 299-307.

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two-stage case control data. *Biometrika* **75**, 11-20.

Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association* **91**, 14-48.

Breslow, N. E. and Holubkov, R. (1997a). Maximum likelihood estimation of logistic regression parameters under two-phase outcome-dependent sampling. *J.R. Stat. Soc. B.* **59**, 447-461.

Breslow, N. E. and Holubkov, R. (1997b). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine* **16**, 103-116.

Breslow, N. E. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Appl. Statist.* **48**, 457-468.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. Appl. Probab.* **31**, 929-953.

Brix, A. and Diggle, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society, B* **63**, 823-841.

Brix, A. and Møller, J. (2001). Space-time multitpe log Gaussian Cox processes with a view to modelling weed data. *Scandinavian Journal of Statistics* **28**, 471-488.

Brix, A. and Kendall, W. S. (2002). Simulation of cluster point processes without edge effects. *Adv. Appl. Prob.* **34**, 267-280.

Burr I. (1942). Cumulative frequency functions. *Annals of Mathematical Statistics* **13**, 215-232.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis, Chapman & Hall/CRC*

Chatterjee, N., Chen, Y. H., and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal Amer. Stat. Assoc.* **98**, 158-168.

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

Chatterjee, N. and Chen, Y. H. (2007). Maximum likelihood inference on a mixed conditionally and marginally specified regression model for genetic epidemiologic studies with two-phase sampling. *J.R. Stat. Soc. B.* **69**, 123-142.

Checkley, W., Robert, G. H. Robert, B. E., Leonardo, E. D., Lilia, C., Charles, S. R., and Lawrence, M. H. (2004). Effect of water and sanitation on childhood health in a poor Peruvian peri-urban community. *Lancet* **363**, 112-118.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *J. R. Stat. Soc. Ser. B* **63**, 19-29.

Christensen, O.F. and Ribeiro, P.J.J. (2011) geoRglm: geoRglm - a package for generalised linear spatial models. `http://cran.r-project.org/web/packages/geoRglm/`.

Choi, J., Hui, S.K., Bell, D.R. (2007). Bayesian Spatio-Temporal Analysis of Imitation Behavior Across New Buyer at Online Grocery Retailer ıJournal of Marketing Research **XLV**.

Cochran, W. G. (1963). Sampling Techniques. New York: Wiley.

Cooney, K. A., Strawderman, M. S., Wojno, K. J., Doerr, K. M., Taylor, A., Alcser, K.H., Heeringa, S. G., Taylor, J. M., Wei, J. T., Montie, J. E., and Schottenfield, D. (2001). Age-specific distribution of serum prostate-specific antigen in a community-based study of African-American men. *Urology* **57**, 91-96.

Congdon, P. (2005). Bayesian Models for Categorical Data. New York: Wiley.

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

Cornfield, J., Gordon, T. and Smith, W. W. (1961). Quantal response curves for experimentally uncontrolled variables. *Bulletin of the International Statistical Institute* **38**, 97-115.

Cox, D. R. (1966). A simple example of a comparison involving quantal data. *Biometrika* **53**, 215-220.

Cowles, M. K. and Zimmerman, D. L. (2003). A Bayesian Space-Time Analysis of Acid

Deposition Data Combined From Two Monitoring Networks. *Journal of Geophysical Research* **DOI:10.1029/2003JD004001.**

Curtis, V. and Cairncross, S. (2003). Effect of washing hands with soap on diarrhea risk in the community, a systematic review. *Lancet Infectious Disease* **3**, 275-281.

Davis, P. J. (1979). Circulant Matrices, *Wiley*, Chichester.

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

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

Diggle, P. J., and Elliott, P. (1995). Disease risks near point sources: Statistical issues for analysis using individual or spatially aggregated data. *Journal of Epidemiology and Community Health* **49**, S20-S27.

Diggle, P. J., Rowlingson, B., and Su, Ting-li. (2005). Point process methodology for on-line spatio-temporal disease surveillance *Environmetrics* **16**, 424-434.

Diggle, P.J, Gomez-Rubio, V., Brown,P. E., Chetwynd, A. G. and Gooding, S. (2007). Second-order analysis of inhomogeneous spatial point process data. *Biometrics* **63**, 550-557.

Dunson, D. B. and Xing, C. (2009). Nonparametric Bayes Modeling of Multivariate Categorical Data. *Journal Amer. Stat. Assoc.* **104**, 1042-1051.

Eisenberg, J. N. S., Cevallos, W., Ponce, K., Levy, K., Bates, S., Scott, J., Hubbard, A., Viera, N., Segovia, R., Espinel, M., Trueba, G., Riley, L., and Trostle, J. (2006). Environmental change and infectious disease: How new roads affect the transmission of diarrheal pathogens in rural Ecuador. *Proceedings of the National Academies of Science* **103**, 19460-19465.

Flanders, W. D. and Greenland, S. (1991). Analytic methods for 2-stage case-control studies and other stratified designs. *Statist. Med.* **10**, 739-747.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling Based Approaches to Calculating Marginal Densities. *Journal Amer. Stat. Assoc.* **85**, 398-409.

Gelfand, A. E., Zhu, L., and Carlin, B. P. (2001). On the Change of Support Problem for Spatio-Temporal Data. *Biostatistics* **2**, 31-45.

Gelfand, A. E., Kim, H. J., Sirmans, C. F., and Banerjee, S. (2003). Spatial Modeling With Spatially Varying Coefficient Processes. *Journal Amer. Stat. Assoc.* **98**, 387-396.

Gelfand, A.E. and Barber, J. (2007). Hierarchical Spatial Modeling for Estimation of Population Size. *Environmental and Ecological Statistics* **14**, 193-205

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

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Anaysis*, 2nd Edition. Chapman & Hall/CRC, Boca Raton, Florida.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721-741.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881-889.

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

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337-348.

Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling. *Applied Statistics* **44**, 455-472.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *J. Am. Statist. Assoc.* **97**, 590-600.

Gómez-Rubio, V., Ferrándiz-Ferragud, J., López-Quilez, A., and Bivand, R. (2011) DCluster - Functions for the detection of spatial clusters of diseases. `http://http://cran.r-project.org/web/packages/DCluster/index.html/`.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternative(with discussion). *J.R. Stat. Soc. B.* **46**, 149-192.

Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in medicine* **13**, 1665-1677.

Gushulak,B. D. and MacPherson, D. W. (2004). Globalization of Infectious Diseases: The Impact of Migration. *Clinical. Infec. Dis.* **38**, 1742-1748.

Haberman, S. J. (1981). Tests for independence in two-way contingency tables based on canonical correlation and linear-by-linear interaction. *The Annals of Statistics* **9**, 1178-1186.

Hachem, C., Morgan, R., Johnson, M., Kuebeler, M., and El-Serag, H. (2009). Statins and the risk of colorectal carcinoma: a nested case-control study in veterans with diabetes. *Am. J. Gastroenterol.* **104**, 1241-1248.

Hanfelt, J. J. and Liang, K. Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika* **82**, 461-477.

Haneuse, S. and Wakefield, J. (2007). Hierarchical models for combining ecological and case-control data. *Biometrics* **63**, 128-136.

Haneuse, S. and Chen, J. (2011). A Multiphase Design Strategy for Dealing with Participation Bias. *Biometrics* **67**, 309-318.

Heikkinen, J. and Arjas, E. (1998). Non-parametric Bayesian Estimation of a Spatial Poisson Intensity. *Scand. Jour. of Stat.* **25**, 435-450.

Herikstad, H., Yang, S., Van, G. T. J., Vugia, D., Hadler, J., Blake, P., Deneen, V., Shiferaw, B., and Angulo, F. J. (2002). A population-based estimate of the burden of diarrhoeal illness in the United States: Food Net, 1996-7. *Epidemiol Infect* **129**, 9-17.

Holtbrügge, W. and Schumacher, M. (1991). A comparison of regression models for the analysis of ordered categorical data. *Appl. Statist.* **40**, 249-259.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **55**, 663-685.

Hossain, M. and Lawson A. B. (2009). Approximate methods in Bayesian point process spatial models. *Computational Statistics and Data Analysis* **53**, 2831-2842.

Hunter, D. J. *et al.* (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet.* **39**, 870-874.

Ishwaran, H. and Rao, J. S. (2003). Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection. *Journal Amer. Stat. Assoc.* **98**, 438-455

Jewett, H. J. (1975). The present status of radical prostatectomy for stages A and B prostatic cancer. *Urol. Clin. North Am.* **2**, 105-124.

Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*, Springer.

Kosek, M., Bern, C., and Guerrant, R. (2003). The Global Burden of Diarrheal Disease, As Estimated from Studies Published Between 1992 and 2000. *Bulletin of the World Health Organization* **81**, 197-204.

Kottas, A., Müller, P., and Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *J. Computational and Graphical Statistics.* **14**, 610-625.

Krige, D. G. (1951). A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand, *Journal of Chemical, Metallurgical, and Mining Society of South Africa*

52, 119-139.

Kuss, O. (2004). Modelling physicians' recommendations for optimal medical care by random effects stereotype regression, in Proceedings of the 18th Workshop on Statistical Modelling, eds G. Verbeke, G. Molenberghs, M.Aerts, S. Fieuws, 245-249.

Kuss, O. (2006). On the estimation of the stereotype regression model. *Computational Statistics* & *Data Analysis* **50**, 1877–1890.

Lall R., Campbell, M. J., Walters, S. J., Morgan, K. and MRC CFAS Co-operative. (2002). A review of ordinal regression models applied on health-related quality of life assessments, *Stat. Meth. Med. Res.* **11**, 49-67.

Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J. (1999). Semiparametric Methods for Response-Selective and Missing Data Problems in Regression. *J.R. Stat. Soc. B.* **61**, 413-438.

Lee, A. J., Scott, A. J., and Wild, C. J. (2010). Efficient estimation in multi-phase case-control studies. *Biometrika* **97**, 361–374.

Levy, K., Hubbard, A., Nelson, K. L., and Eisenberg, J. N. S. (2009). Drivers of Water Quality Variability in Northern Coastal Ecuador. *Environ. Sci. Technol.* **43**, 1788-1797.

Li, D. and Conti, D.V. (2009). Detecting Gene-Environment Interactions Using a Combined Case-Only and Case-Control Approach. *American Journal of Epidemiology* **169**, 497-504.

Liang, S., Carlin, B. P., and Gelfand, A. E. (2009). Analysis of Minnesota colon and rectum cancer point patterns with spatial and non-spatial covariate information. *The Annals of applied statistcs* **3**, 943-962.

Lipkin, S. M., Chao, L., Moreno, V. M., Rozek, L. S., Rennert, H., Pinchev, M., Dizon, D., Rennert, G., Koelovich, L., Gruber, S. B. (2010). Genetic Variation in 3-Hydroxy-3-Methylglutaryl CoA Reductase Modifies the Chemopreventive Activity of Statins for Colorectal Cancer. *Cancer Prevention Research* **5**, 597-603.

Lipsitz, S. R., Parzen, M. and Ewell, M. (1998). Inference using Conditional Logistic Regression with Missing Covariates. *Biometrics* **54**, 295-303.

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

Liu, J., Gustafson, P., Cherry, N., and Bursty, I. (2009). Bayesian analysis of a matched case-control study with expert prior information on both the misclassification of exposure and the exposure-disease association. *Statistics in Medicine* **28**, 3411-3423.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J.R. Stat. Soc. B.* **44**, 226-233.

Lumley, T. (2011). Package Survey, Version 3.2.4. R for analyzing data from complex surveys. `http://cran.r-project.org/package/survey`.

Lunt, M. (2004). Prediction of ordinal outcomes when the association between predictors and outcome differs between outcome levels. *Statistics in Medicine* **24**, 1357-1369.

Manski, C.F. and Lerman, S. (1977). The Estimation of Choice Probabilities from Choice-Based Samples. *Econometrica* **45**, 1977-1988.

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.

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

McCullagh, P. and Nelder, J. A. (1983). Generalized linear models. London: Chapman and Hall.

McCullagh, P. (1984). Discussion of professor Anderson's paper. *J. R. Stat. Soc. B.* **46**, 1-30.

Meng, X. L. and Rubin, D. B. (1993). Maximum Likelihood Estimation via the ECM Algorithm : A General Framework. *Biometrika* **80**, 267-278.

Mitchell, T. J., and Beauchamp, J. J. (1988), Bayesian Variable Selection in Linear Regression (with discussion). *Journal of the American Statistical Association* **83**, 1023-1036.

Montie, J. E. (1995). Staging of prostate cancer: current TNM classification and future prospects for prognostic factors. *Cancer* **75**, 1814-1818.

Mukherjee, B., Liu, I. and Sinha, S. (2007). Analysis of Matched case-control data with ordinal disease states: possible choices and comparisons, *Statistics in Medicine* **26**, No 17, 3240-3257.

Mukherjee B, Zhang L, Ghosh M, Sinha S. (2007). Semiparametric Bayesian analysis of case-control data under conditional gene-environment independence. *Biometrics* **63**, 834-844.

Mukerjee, B., Ahn, J., Liu, I., Rathouz, P.J. and Sanchez, BN (2008). On elimination of nuisance parameters in a stratified proportional odds model by amalgamating conditional likelihoods. *Statistics in Medicine* **27**, 4950-4971.

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

Mukherjee, B., Ahn, J., Rennert, G., Gruber, S.B., Moreno, V. and Chatterjee, N. (2008). Testing gene-environment interaction from case-control data: A novel study of Type-1 error, power and designs. *Genetic Epidemiology* **32**, 615-626.

Mukherjee, B. and Liu, I. (2009) A characterization of bias for fitting multivariate generalized linear models under choice-based sampling. *Journal of Multivariate Analysis* **100**, 459-472.

Mukherjee, B., Ahn, J., Gruber, S.B., Ghosh, M. and Chatterjee, N. (2010). Case-Control Studies of Gene Environment Interaction: Bayesian Design and Analysis. *Biometrics* **66**, 934–948.

Murcray, C.E., Lewinger, J. P., and Gauderman, W. J. (2009). Gene-environment interaction in genome-wide association studies. *Am. J. Epidemiol.* **169**, 219-226.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Statist.* **25**, 451-482.

Møller, J. and Waagepetersen, R. P. (2002). Statistical inference for Cox processes, in A. B. Lawson and D. Denison (eds), *Spatial cluster Modelling* , *Chapman & Hall/CRC*, Boca Raton, FL. 37-60.

Møller, J. and Waagepetersen, R. P. (2003). Statistical Inference and Simulation for Spatial Point Processes. *Chapman & Hall/CRC*, Boca Raton, FL.

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

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

National Cancer Institute. (2009). Physical Activity and Cancer.
`http://www.cancer.gov/cancertopics/factsheet/prevention/physicalactivity`.

Neyman, J. (1938). Contribution to the theory of sampling from human populations. *Journal Amer. Stat. Assoc.* **33**, 101-116.

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

Paik, M. C. and Sacco, R. L. (2000). matched case-control data analyses with missing covariate. *J.R. Stat. Soc. C.* **49**, 145-156.

Paik, M. C. (2004). Nonignorable missingness in matched case-control data analyses. *Biometrics* **60**, 306-314.

Park, J.H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J. and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Gen.* **42** , 570-575.

Park, E. and Kim, Y. (2004). Analysis of longitudinal data in case control studies. *Biometrika* **91**, 321-330.

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers and Geosciences* **30**, 683-691

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

Plummer M., Best, N., Cowles, K., Vines, K. (2009). Package CODA, Version 0.13-4, Output analysis and diagnostics for MCMC. `http://cran.r-project.org/web/packages/coda`.

Poynter, J. N., Gruber, S. B., Higgins, P. D. R., Almog, R., Bonner, J. D., Rennert, H. S., Low, M., Greenson, J. K., and Rennert, G. (2005). Statins and the Risk of Colorectal Cancer. *The New England Journal of Medicine* **352**, 2184-2192.

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

Prescott, G. J. and Garthwaite, P. H. (2005). Bayesian analysis of misclassified binary data from a matched case-control study with a validation sub-study. *Statistics in Medicine* **24**, 379-401.

Prostate In. (2002). American Joint Committee on Cancer.: AJCC Cancer Staging Manual. 6th ed. New York, NY: Springer, 309-316.

R Development Core Team. (2011). R: A Language and Environment for Statistical Computing.

`http://www.R-project.org/`.

Rathouz, P. J., Satten, G. A., and Carroll, R. J. (2002). Semiparametric inference in matched case-control studies with missing covariate data. *Biometrika* **89**, 905-916.

Rathouz, P. J. (2003). Likelihood methods for missing covariate data in highly stratified studies. *J.R. Stat. Soc. B.* **65**, 711-723.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299-314.

Rice, M. 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, M. K. (2006). On Bayesian analysis of misclassified binary data from a matched case-control study with a validation sub-study, by Gordon Prescott and Paul Garthwaite, with accompanying R code for 1:1 matched studies and 1:3 matched studies. *Statistics in Medicine* **25**, 537-539.

Rice, M. K. (2008). Equivalence between conditional and random-effects likelihoods for pair-matched case-control studies. *Journal of the American Statistical Association* **103**, 385-396.

Ripley, B. D. (1976). On stationarity and superposition of point processes. *Ann. Prob.* **4**, 999-1005.

Ripley, B. D. (1977). Modelling spatial patterns (with Discussion). *J. R. Statist. Soc. B.* **39**, 172-212.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal Amer. Stat. Assoc.* **89**, 846-866.

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

Sarma, A. V. and Schottenfield, D. (2002). Prostate cancer incidence, mortality, and survival trends in the United States: 1981-2001. *Semin. Urol. Oncol.* **20**, 3-9.

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

Satten, G. and Carroll, R. J. (2000). Conditional and unconditional categorical regression models with missing covariates. *Biometrics* **56**, 384-388.

Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57-71.

Schill, W., Jockel, K.H., Drescher, K. and Timm, J. (1993). Logistic analysis in Case-control studies under validation sampling. *Biometrika* **80**, 339-352.

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

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica* **4**, 639-650.

Sinha S., Mukherjee, B., and Ghosh, M. (2004). Bayesian semiparametric modeling for 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. (2005). Semiparametric Bayesian analysis of matched case-control studies with missing exposure. *Journal of the American Statistical Association* **100**, 591-601.

Sinha, S., Mukherjee, B., and Ghosh, M. (2007). Modeling association among multivariate exposures in matched case-control study. *Sankhya* **64**, 379-404.

Sinha S. and Maiti, T. (2008). Analysis of matched case-control data in presence of nonignorable missing exposure. *Biometrics* **64**, 106-114.

Sinha, S., Gruber, S. B, Mukherjee, B., and Rennert, G. (2008). Inference on haplotype effects in matched case-control studies using unphased genotype data. *International Journal of Biostatistics* **4**, No. 1, Article 6.

5estimation of gene-gene and gene-environment interactions for numerous loci using

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528-550.

Thomas, D.C. (2010). The calculation of Gene-environment-wide association studies: emerging approaches. *Nature Reviews, Genetics* **11** 259–272.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics* **22**, 1701-1728.

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**, 1731-1743.

Vansteelandt, S., DeMeo, D.L., Lasky-Su, J., Smoller, J.W. et al. (2008). Testing and Estimating Gene Environment Interactions in Family-Based Association Studies. *Biometrics* **64**, 458-467.

Van Lieshout, M. N. M. (2000). Markov Point Process and Their Applications. *Imperial College Press*, London.

Waagepetersen, R. P. (2004). Convergence of posteriors for discretized log Gaussian Cox processes. *Statistics & Probability Letters* **66**, 22-235.

Wacholder, S., Hartge, P., Prentice, R., Garcia-Closas, M. et al. (2010). Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.* **362**, 986-993.

Walker, S. G. (2007). Sampling the Dirichlet Mixture Model With Slices. *Simulation and Computation* **36**, 45-54.

Waller, L. A. and Gotway, C. A. (2004). Applied Spatial Statistics for Public HealthData. *New York: Wiley.*

Whittemore, A. S. and Halpern, J. (1998). Multi-stage sampling in Genetic Epidemiology. *Stat in Med.* **16**, 153-167.

Wood, A. T. A. and Chan, G. (1994). Simulation of sationary Guassian processes in $[0, 1]^d$. *Journal of Computational and Graphical Statistics* **3**, 409-432.

World Health Organization. (2006). "Cancer". Retrieved on 2007-05-24.

Yeager, M. *et al.* (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet.* **39**,645-49.

Yee, T. W. and Hastie, T. J. (2003). Reduced-rank vector generalized linear models. *Statistical Modeling* **3**, 15-41.

Yee, T. W. (2010). The VGAM Package for Categorical Data Analysis. *Journal of Statistical Software* **32**, 1-34.

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

Zhang, L., Mukherjee, B., Ghosh, M., and Wu, R. (2006). A Bayesian Framework for Genetic Association in Case-Control Studies: Accounting for Unknown Population Substructure. *Statistical Modeling* **6**, 352-372.

Zhang, L., Mukherjee, B., Ghosh, M., Gruber, S. B., and Moreno, V. (2008). Misclassification of exposures in case-control studies of gene-environment interaction *Statistics in Medicine*, **27**, 2756-2783.