NEW ADVANCES IN LOGISTIC REGRESSION FOR HANDLING MISSING AND MISMEASURED DATA WITH APPLICATIONS IN BIOSTATISTICS

A Dissertation

by

JINGANG MIAO

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	Suojin Wang
Co-Chair of Committee,	Samiran Sinha
Committee Members,	Michael Longnecker
	E. Lisako J. McKyer
Head of Department,	Simon Sheather

August 2014

Major Subject: Statistics

ABSTRACT

As a probabilistic statistical classification model, logistic regression (or logit regression) is widely used to model the outcome of a categorical dependent variable based on one or more predictor variables/features. We study two problems related to logistic regression with applications in biostatistics.

In the first problem, we study multivariate disease classification in the presence of partially missing disease traits. In modern cancer epidemiology, diseases are classified based on pathologic and molecular traits, and different combinations of these traits give rise to many disease subtypes. The effect of predictor variables can be measured by fitting a polytomous logistic model to such data. The differences (heterogeneity) among the relative risk parameters associated with subtypes are of great interest to better understand disease etiology. Due to the heterogeneity of the relative risk parameters, when a risk factor is changed, the prevalence of one subtype may change more than that of another subtype does. Estimation of the heterogeneity parameters is difficult when disease trait information is only partially observed and the number of disease subtypes is large. We consider a robust semiparametric approach based on the pseudo conditional likelihood for estimating these heterogeneity parameters. Through simulation studies, we compare the robustness and efficiency of our approach with the maximum likelihood approach. The method is then applied to analyze data from the American Cancer Society Cancer Prevention Study (CPS) II Nutrition Cohort. Weight gain was associated with the risk of breast cancer and the association varies by disease subtype.

In the second problem, we use a semiparametric Bayesian method to handle measurement errors. In nutritional epidemiological studies, nutrient intakes are often measured via food frequency questionnaires and 24-hour dietary recalls. Due to self reporting, recall error, and other reasons, the measured nutrient intakes can involve a substantial amount of noise. While independence assumption between the measurement error and the true predictor is likely to be a reasonable assumption for the main effect of the predictors, this assumption is not tenable for the interaction effect of two predictors measured with error. Although there are a number of flexible methods for handling additive, homogeneous measurement error in predictors in logistic regression models, relatively less attention has been paid to handling measurement error that depends on the unobserved predictor. Therefore, we propose a semiparametric Bayesian method for handling this unorthodox measurement error scenario in logistic regression models in the presence of the interaction term. The proposed method is also designed to handle partially missing values for the error-prone surrogate variables. Through simulation studies, we assess some operating characteristics of the proposed method and compare it with the simulation extrapolation and the regression calibration method. Our method has smaller biases than the other methods. In addition, we analyze the NHANES data and assess the association between some important nutrients and high cholesterol level. Total fat and protein reinforce each other's association with the risk of having high cholesterol level.

DEDICATION

To Wenhua, Maggie, and my parents

ACKNOWLEDGEMENTS

I'd like to thank a few people for their help.

I would like to express the deepest appreciation to my advisors, Dr. Suojin Wang and Dr. Samiran Sinha, whose guidance and persistent help have made this dissertation possible. Dr. Wang, with his rich research experience and strong expertise in and beyond statistics, has been a role model. Dr. Sinha, with his immense patience and hands-on coaching, has given me the needed tools, such as Latex and Fortran programming, and helped conquer one challenge after another in my research.

My gratitude also goes to other members of my committee. Dr. Michael Longnecker is one of my favorite professors, and his courses on methods of statistics were among the best courses I've ever taken. Those courses opened the door for me to the world of statistics and helped me decide to pursue a degree in statistics. Dr. E. Lisako McKyer has been a mentor as well as a friend. In addition to mentoring and directly helping me when we collaborate on some research projects, she has been helping me indirectly too — by being the best advisor to my wife.

I'd like to thank my parents for trying their very best to keep me in school when life was so hard that sometimes even feeding the family became a challenge. You are the best parents in the world. Now all three of your sons have finished college and are on their ways to achieve great things, so you have every reason to be proud.

Last but not least, my wife Wenhua and daughter Maggie. Wenhua, thank you for your encouragement by writing your own dissertation around the same time and doing a good job. More importantly, thank you for the delicious food you prepare everyday, although I was hoping that writing the dissertation could help me lose a few pounds. Maggie, writing a dissertation can be stressful sometimes, so thank

you for all the joy and happiness, especially during our fishing trips, which quickly recharged my batteries. Also, thank you for all the good advice you gave to your dad and mom — “check your data.”

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1. OVERVIEW OF LOGISTIC REGRESSION

A categorical variable, as its name suggests, has a measurement scale of a set of categories, and examples include gender (male vs. female) and disease status (case vs. control). A categorical variable is said to be ordinal if an ordering is present among the levels of the variable, and an example would be education level (high school or below vs. associate or bachelor degrees vs. advanced degrees). A categorical variable is classified as nominal when no ordering is present among the levels. Categorical scales are widely used in the social sciences and medical sciences.

As a probabilistic statistical classification model, logistic regression (or logit regression) is widely used to explain/predict the outcome of a categorical dependent variable based on one or more predictor variables/features. The predictor variables/features themselves can be categorical or continuous. Logistic regression seems to be very popular in different industries as well as in scientific research. For example, social and behavioral researchers interested in obesity prevention may use it to find out what factors influence whether or not a child actively commutes to school; a bank may use it to predict whether a credit card transaction is valid or fraudulent; and an insurance company may use it to predict whether a customer will stay with them or switch to a competitor.

Depending on the number of possible values of the outcome variable, different variants of logistic regression can be used, including (1) the binary/binomial logistic model for a binary (e.g., true vs. false) outcome and (2) the polytomous/multinomial logistic model for an outcome with more than two levels (e.g., no change vs. better vs. worse).

1.1 Why Logistic Regression

Linear regression is perhaps the most used method to explain/predict an outcome based on predictor variables. However, linear regression may not work well when the outcome is categorical. For example, when modeling the probability that a student will complete college within 4 years, a linear model may give predictions outside the range of 0 to 1; in addition, the error terms may violate linear regression's assumptions of equal variance and normal distribution. Logistic regression, on the other hand, is designed to handle categorical outcomes.

The following properties may have contributed to the popularity of logistic regression: (1) elegant interpretation: unlike methods such as neural networks and decision trees, logistic regression's parameters can be elegantly interpreted as the log odds ratio associated with every 1-unit increase in the corresponding predictor variable; (2) easy estimation and inference: as a member of the exponential family, logistic regression's estimation and statistical inference are straightforward and computationally fast; (3) a fairly complete toolbox: various methods have been developed for logistic regression's model selection, model diagnostics, etc.

1.2 Logistic Models Related to This Dissertation

In binary logistic regression, the outcome Y is usually coded as"0" for a negative result and "1" for a positive result for convenience in interpretation and mathematical derivations. For a length-P vector of predictor variables $\boldsymbol{X}$, which includes as its first element a 1 for the intercept term, the logistic model is $\pi(\boldsymbol{X}) = \Pr(Y = 1|X) = 1 - \Pr(Y = 0|X) = \{1 + \exp(-(X^T\boldsymbol{\beta}))\}^{-1}$. This way, $\pi(\boldsymbol{X})$ is guaranteed to be in the range of 0 to 1. An alternative way to express the model is $\text{logit}(\Pr(Y = 1|\boldsymbol{X})) = \boldsymbol{X}^T\boldsymbol{\beta}$.

The interpretation of $\boldsymbol{\beta}$ is straightforward: If we let X_p be the p-th component of

$\boldsymbol{X}$ and β_p be the corresponding element in $\boldsymbol{\beta}$, then it is easy to see that the logit, or log odds ratio, of $\pi(\boldsymbol{X})$ increases by β_p with every 1-unit increase in X_p. An equivalent but more commonly used interpretation is that the odds increase multiplicatively be $\exp(\beta_p)$ with every 1-unit increase in X_p.

Parameter estimation and statistical inference for binary logistic regression can be done with the usual maximum likelihood mechanics. With a sample size of n, the likelihood function is simply the product of n binomial probability mass functions, i.e., $L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i}(1-\pi(\boldsymbol{x}_i))^{1-y_i}$. The corresponding log likelihood is $l(\boldsymbol{\beta}) = log(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} \boldsymbol{x}_i^T\boldsymbol{\beta} - \sum_{i=1}^{n} \log(1-\pi(\boldsymbol{\beta};\boldsymbol{x}_i))$. Then, the score equations can be obtained by setting the derivatives of the log likelihood with respect to $\boldsymbol{\beta}$ to zero, i.e.,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \pi(\boldsymbol{\beta};\boldsymbol{x}_i))\boldsymbol{x}_i^T.$$

The solution to score equations is the maximum likelihood estimate (MLE), which can be found numerically since an analytical form does not exist. One of the numeric tools can be used is Newton's method (also known as the Newton-Raphson method), which in the context of logistic regression has a new name "Iterative Weighted Least Squares" (IWLS). The variance/covariance of the estimate is just the inverse of the Fisher information matrix.

For an outcome with J nominal levels, where $J > 2$, a polytomous logistic model with a reference category specified by the user can be fit. For example, suppose we are interested in the result of cancer screening in women and the possible outcomes include breast cancer ($J = 1$), lung cancer ($J = 2$), colorectal cancer ($J = 3$), and no cancer ($J = 4$), we can use no cancer as the reference category and model $J-1$ probabilities $\Pr(Y = 1), \ldots, \Pr(Y = J-1)$. Formally, $\log(\Pr(Y = j|\boldsymbol{X})/\Pr(Y =$

$J|\boldsymbol{X})) = \boldsymbol{X}^T\boldsymbol{\beta}_j$, $j = 1, \ldots, J-1$. The probabilities for the outcomes are $\Pr(Y = j|\boldsymbol{X}) = \{\sum_{k=1}^{J} \boldsymbol{X}^T(\boldsymbol{\beta}_k - \boldsymbol{\beta}_j)\}^{-1}$ with $\boldsymbol{\beta}_J = 0$. Its parameter estimation and statistical inference can be done in a similar fashion as in binary logistic regression.

1.3 Organization of Dissertation

In Section 2, a method will be proposed to handle multivariate disease classification data in the presence of partially missing disease traits, and the method will be used to analyze the associations between weight gain and the risk of breast cancer subtypes using data from the American Cancer Society Cancer Prevention Study (CPS) II Nutrition Cohort. In Section 3, we will address a non-standard measurement error problem with a semiparametric Bayesian method and analyze the NHANES data to assess the effects of some important nutrients on high cholesterol level. A summary and some discussions on future work will be contained in Section 4.

2. ANALYSIS OF MULTIVARIATE DISEASE CLASSIFICATION DATA IN THE PRESENCE OF PARTIALLY MISSING DISEASE TRAITS

2.1 Motivating Data and Problem to Solve

While disease trait information has been used in understanding survival of patients, relatively less research has been done on incorporating disease trait information into etiologic investigations. In this section, we analyze data from the American Cancer Society's Cancer Prevention Study (CPS) II Nutrition Cohort (Calle et al. 2002) to investigate whether the association between weight gain and risk of breast cancer varies among different disease trait subtypes in women not using postmenopausal hormones, adjusting for important risk factors. If the association of a predictor variable varies across the subtypes, we examined how much of this variation is due to each of the disease traits. Understanding "etiologic heterogeneity" of a risk factor sheds light on the pathogenesis of disease (Morton et al. 2008). In the CPS-II Nutrition Cohort, there are 5 tumor characteristics, including stage (2 levels), histology (3 levels), estrogen receptor (2 levels), progesterone receptor (2 levels), and grade (3 levels), leading to 72 (i.e., $2 \times 3 \times 2 \times 2 \times 3$) different disease subtypes.

To examine the effect of risk factors on different disease subtypes, we consider the polytomous logistic regression, which is commonly used for handling multinomial data (Fagerland et al. 2008; Engel 1998; Hosmer 2000). There are two variants of the model: one for nominal and one for ordinal scale outcomes (Goeman and le Cessie 2006), and we focus on modeling nominal outcomes. For each disease subtype, we have a set of disease-predictor association/regression parameters and a set of nuisance intercept parameters. The etiologic heterogeneity will be measured via differences among the regression parameters across subtypes. The number of regression parameters can be large when several disease characteristics (traits) are of interest and each characteristic has multiple levels. In this context, a second-stage model was proposed to reduce the dimension of the heterogeneity parameters when all disease traits are observed (Chatterjee 2004). The problem becomes even more challenging when the disease traits are partially missing, which was not address in Chatterjee (2004). In the CPS-II Nutrition Cohort data, the missingness percentages for the five traits are 23.2%, 21.2%, 0.0%, 30.0%, and 33.6%, respectively. In particular, among the cases, approximately 45.5% had at least one missing trait.

While estimation of the heterogeneity parameters was considered in the Cox regression model in the presence of partially missing disease traits (Chatterjee et al. 2010), the same issue has not been considered before in the context of polytomous logistic model, which will be considered in this section. We will propose a method to estimate the heterogeneity parameters using a pseudo conditional likelihood. We would like to point out that in the presence of missing data, the pseudo conditional likelihood is not free from the nuisance intercept parameters. For estimating these nuisance parameters, we use a different type of pseudo conditional likelihood. For handling the large dimension of the nuisance parameters, we adopt another second-stage model, and estimate them from another objective function. The idea of using two objective functions, one for the main parameters of interest and the other for the nuisance parameters, was inspired by Goetghebeur and Ryan (1995).

Alternative to the proposed approach, one could consider a maximum likelihood based inference for the heterogeneity parameters using the full likelihood of the data. However, misspecification of the model for the intercepts will have less bearing on our inference than on the full likelihood based approach. This robustness property of our approach will be demenstrated through simulation studies. Our inference is based on an artificially constructed pseudo conditional likelihood function. To show its validity, we derive the large sample properties of the resulting estimator.

2.2 Model and Notation

For each subject in a cohort of n subjects, when no missingness occurs we observe (D, Y, X), where D takes on one or zero according to whether the subject is diagnosed with the disease or not during the follow-up period. For the sake of simplicity and easy understanding, we first consider only two disease traits (i.e., $K = 2$) and assume that X is a scalar covariate (i.e., $P = 1$). The general case of $K > 2$ and $P > 1$ can be derived in the same fashion. Thus, $Y = (Y_1, Y_2)^T$ carries information on 2 disease traits. For a disease-free subject, we have $D = 0$ and $Y = (0, 0)^T$. If the k-th trait has M_k levels, then there are a total of $M = M_1 \times M_2$ disease subtypes. Our model is

$$
\begin{aligned}
p_{i,(y_1,y_2)} \equiv \mathrm{pr}(D_i = 1, Y_i = (y_1, y_2)|X_i) &= \frac{\exp(\alpha_{(y_1,y_2)} + \beta_{(y_1,y_2)} X_i)}{1 + \sum_{(y_1,y_2)} \exp(\alpha_{(y_1,y_2)} + \beta_{(y_1,y_2)} X_i)}\ , \\
\mathrm{pr}(D_i = 0|X_i) &= \frac{1}{1 + \sum_{(y_1,y_2)} \exp(\alpha_{(y_1,y_2)} + \beta_{(y_1,y_2)} X_i)}\ , \qquad (2.1)
\end{aligned}
$$

for $i = 1, \ldots, n$, where $\beta_{(y_1,y_2)}$ denotes the log-odds ratio parameter of the disease subtype (y_1, y_2) for the covariate, $\alpha_{(y_1,y_2)}$ denotes the nuisance intercept parameter, and $\sum_{(y_1,y_2)}$ means summing over all M subtypes of the disease.

For a scalar continuous covariate scenario, there are M main regression (log-odds ratio) parameters of interest along with M intercept parameters, which are not the main interest. Etiologic heterogeneity is measured via the differences among the regression parameters for a given covariate, and our focus is on the estimation of the heterogeneity parameters.

To measure heterogeneity and reduce the dimension of subtype-specific regression parameters, following Chatterjee (2004) we use the following second-stage model for the log-odds ratio parameters in model (2.1):

$$\beta_{(y_1,y_2)} = \theta^{(0)} + \theta^{(1)}_{1(y_1)} + \theta^{(1)}_{2(y_2)} + \theta^{(2)}_{12(y_1,y_2)}, \tag{2.2}$$

where $\theta^{(0)}$ is the regression coefficient corresponding to the reference subtype of the disease, and the first-order and second-order parameter contrasts are respectively represented by $\theta^{(1)}_{k(y_k)}, k = 1, 2,$ and $\theta^{(2)}_{12(y_1,y_2)}$. By assuming certain contrasts to be zero, we can reduce the number of parameters. In addition, these assumptions can be tested. Assuming the second- and higher-order contrasts are equal to zero, which we call a second-stage additive model, $\theta^{(1)}_{1(y)} - \theta^{(1)}_{1(y^*)}$ tells us the degree of etiologic heterogeneity with respect to the first trait, regardless of the levels of other traits. For identifiability, we set $\theta^{(1)}_{1(1)} = \theta^{(1)}_{2(1)} = 0$, and $\theta^{(2)}_{12(1,y_2)} = \theta^{(2)}_{12(y_1,1)} = 0$. More elaborately, the heterogeneity of the log-odds ratio parameters due to the first trait can be measured via the contrasts $\theta^{(1)}_{1(2)}, \ldots, \theta^{(1)}_{1(M_1)}$. By assuming the second-order contrast parameters to be zero, we reduce the dimension of regression parameters from $M_1 \times M_2$ to $1 + M_1 - 1 + M_2 - 1 = M_1 + M_2 - 1$.

To simplify the notation in the second-stage model, we use a design matrix $\mathcal{B}$ to relate the coefficient β that contains all the $\beta_{(y_1,y_2)}$ parameters of the unstructured polytomous model to the parameters θ of the second-stage model (2.2) as $\beta = \mathcal{B}\theta$. In particular, $\beta^T_{(y_1,y_2)} = \mathcal{B}^T_{(y_1,y_2)}\theta$, where $\mathcal{B}_{(y_1,y_2)}$ denotes the row of $\mathcal{B}$ corresponding to disease subtype (y_1, y_2). Also, using a second-stage model we can write $\alpha = \mathcal{A}\xi$, where α is a length-M vector of all $\alpha_{(y_1,y_2)}$ parameters. We use ξ to denote the second-stage parameters for the nuisance intercept parameters. For clarity, we write $\alpha_{(y_1,y_2)} = \mathcal{A}^T_{(y_1,y_2)}\xi$, where $\mathcal{A}^T_{(y_1,y_2)}$ denotes the row of $\mathcal{A}$ that corresponds to disease subtype (y_1, y_2).

Note that the use of the second-stage model for the regression parameters is not just for dimension reduction. More importantly, these second-stage model parameters are our main interest. As mentioned previously, these parameters directly measure the heterogeneity in the log-odds ratio parameters due to each of the disease trait. For the purpose of dimension reduction we set second- and higher-order contrasts to be zero.

We introduce non-missing value indicator variables, $R_i = (R_{i1}, R_{i2})^T$, where $R_{ik} = 1$ $(k = 1, 2)$ if the k-th trait is observed for *diseased* subject i and 0 otherwise. Since for a non-diseased subject there is no relevance of disease traits, for all *non-diseased* subjects we set $R = (1, 1)^T$ for convenience. Note that there are at most 2^2 types of missing data patterns: $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$, with $(1, 0)$, for example, meaning the first trait was observed but the second trait was not observed. We assume that the probability of observing missingness pattern r, $\text{pr}(R = r|Y, X) = \pi(r, X)$, does not depend on the disease traits. However, we not only allow the missingness probabilities to depend on X (a case of missing at random, MAR, Little 1998; Rubin 1976) but also allow the missingness indicators of different traits, R_1 and R_2, to be dependent on each other.

We introduce some additional notations to be used in the following subsections. For the i-th subject, whose missing data pattern is r, we partition its vector of disease traits into the observed traits $y_i^{o_r}$ and the missing traits $y_i^{m_r}$. Similarly, we will use $\sum_{y_i^{m_r}}$ to sum over all the possible values of $y_i^{m_r}$. For example, if $Y_1 = y_1$ but Y_2 is missing, then $r = (1, 0)$, $y^{o_r} = y_1$, $y^{m_r} = Y_2$, whose value is missing, and $\sum_{y^{m_r}}$ means summing over all the terms corresponding to $(Y_1 = y_1, Y_2 = 1), (Y_1 = y_1, Y_2 = 2) \ldots, (Y_1 = y_1, Y_2 = M_2)$. When both traits are observed, $\sum_{y^{m_r}}$ just uses the term corresponding to $(Y_1 = y_1, Y_2 = y_2)$.

2.3 Estimation Methodology

2.3.1 Maximum likelihood method in the context of missing data

To estimate θ, one can use the maximum likelihood estimator (MLE), which is obtained by maximizing the full likelihood

$$
\begin{aligned}
L \;=\; \prod_{i=1}^{n} \Bigg[& \Bigg\{ \frac{1}{1+\sum_{(y_1,y_2)} \exp(\mathcal{A}^T_{(y_1,y_2)}\xi + X_i\mathcal{B}^T_{(y_1,y_2)}\theta)} \Bigg\}^{1-D_i} \\
& \times \prod_{r} \Bigg\{ \frac{\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + X_i\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta)}{1+\sum_{(y_1,y_2)} \exp(\mathcal{A}^T_{(y_1,y_2)}\xi + X_i\mathcal{B}^T_{(y_1,y_2)}\theta)} \Bigg\}^{I(R_i=r)D_i} \Bigg].
\end{aligned}
$$

The resulting score functions for θ and ξ can be compactly written as the following:

$$
S_\theta \equiv \frac{\partial \log(L)}{\partial\theta} = \sum_{i=1}^{n} \Bigg\{ D_i X_i \sum_r I(R_i = r) \sum_{y_i^{m_r}} \mathcal{B}_{(y_i^{o_r},y_i^{m_r})}\omega_{(y_i^{o_r},y_i^{m_r},X_i)} - X_i \sum_{(y_1,y_2)} \mathcal{B}_{(y_1,y_2)} p_{i,(y_1,y_2)} \Bigg\},
$$

$$
S_\xi \equiv \frac{\partial \log(L)}{\partial\xi} = \sum_{i=1}^{n} \Bigg\{ D_i \sum_r I(R_{\overset{}{i}} = r) \sum_{y_i^{m_r}} \mathcal{A}_{(y_i^{o_r},y_i^{m_r})}\omega_{(y_i^{o_r},y_i^{m_r},X_i)} - \sum_{(y_1,y_2)} \mathcal{A}_{(y_1,y_2)} p_{i,(y_1,y_2)} \Bigg\},
$$

where

$$
\begin{aligned}
\omega_{(y_i^{o_r},y_i^{m_r},X_i)} &= \frac{\exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + X_i\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta)}{\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + X_i\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta)}, \\
p_{i,(y_1,y_2)} &= \frac{\exp(\mathcal{A}^T_{(y_1,y_2)}\xi + X_i\mathcal{B}^T_{(y_1,y_2)}\theta)}{1+\sum_{(y_1,y_2)} \exp(\mathcal{A}^T_{(y_1,y_2)}\xi + X_i\mathcal{B}^T_{(y_1,y_2)}\theta)}.
\end{aligned}
$$

We have some model assumptions:

Let $S_n(\eta) = n^{-1}(S^T_{\text{EE},\theta^{(1)}}, \ldots, S^T_{\text{EE},\theta^{(P)}}, S^T_{\text{EE},\xi})^T$.

C1. The parameter space for η is a compact subset of an Euclidean space.

C2. $0 < \exp(\sum_{p=1}^{p} X_p \mathcal{B}_y^{(p)}\theta^{(p)}) < \infty$ for all $\theta^{(p)}$ and y.

C3. $0 < \exp(\mathcal{A}_y\xi) < \infty$ for all ξ and y.

C4. The elements of the second-stage design matrices $\mathcal{B}$ and $\mathcal{A}$ remain uniformly bounded in absolute value by constants, say $C_{\mathcal{B}}$ and $C_{\mathcal{A}}$, respectively.

C5. The information matrix H_n is positive definite.

C6. The deterministic equation $E\{S_n(\eta)\} = 0$ has only one root in the neighborhood of the true parameters.

Conditions C1-C4 are required for uniform convergence, i.e., $\sup_\eta ||S_n(\eta) - E\{S_n(\eta)\}|| \xrightarrow{P} 0$. Condition C5, C6 (identifiability) and the asymptotic unbiasedness of $S_n(\eta)$ for zero (to be proved) together imply convergence of the estimator in probability towards the true value (Theorem 5.9 of van der Vaart 1998).

If the model assumptions above hold, then under standard regularity conditions given in Theorem 5.41 of van der Vaart (1998), the MLE $\tilde{\eta} = (\tilde{\theta}^T, \tilde{\xi}^T)^T$ asymptotically follows a normal distribution with mean $\eta = (\theta^T, \xi^T)^T$, and the asymptotic variance can be obtained by inverted the Fisher information matrix.

As evident from the above discussion, the inference of the heterogeneity parameters, θ, depends on the intercept parameters α and their model $\alpha = \mathcal{A}\xi$. Next we discuss an alternative inference for the heterogeneity parameters, which is more robust against the misspecification of the second-stage model for α.

2.3.2 Pseudo conditional likelihood in the context of missing data

In order to form a pseudo conditional likelihood (PCL), for every subject with disease, we define a matched set $\mathcal{S}$ consisting of the subject itself and all subjects without the disease. Thus, if $D_i = 1$, then $\mathcal{S}_i = \{i\} \cup \{j : D_j = 0\}$. If there are n_0 controls, then the cardinality of $\mathcal{S}_i$ is $(n_0 + 1)$. We form the pseudo conditional likelihood $\mathcal{L}_{\mathrm{PCL},i}$ such that the i-th subject has a disease of subtype $(y_i^{o_r}, y_i^{m_r})$ given that there is only one subject with disease $(y_i^{o_r}, y_i^{m_r})$ in the set $\mathcal{S}_i$:

$$\mathcal{L}_{\mathrm{PCL},i}=\prod_r \left\{ \sum_{y_i^{m_r}} \frac{\mathrm{pr}(D_i=1, Y_i=(y_i^{o_r}, y_i^{m_r})|X_i)\prod_{j\in\mathcal{S}_i\backslash\{i\}}\mathrm{pr}(D_j=0|X_j)}{\sum_{k\in\mathcal{S}_i}\sum_{y_i^{m_r}}\mathrm{pr}(D_k=1,Y_k=(y_i^{o_r},y_i^{m_r})|X_k)\prod_{j\in\mathcal{S}_i\backslash\{k\}}\mathrm{pr}(D_j=0|X_j)}\right\}^{I(R_i=r)}$$

$$=\prod_r \left\{ \frac{\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + X_i \mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta)}{\sum_{j\in\mathcal{S}_i}\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + X_j \mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta)} \right\}^{I(R_i=r)}.$$

Then the pseudo conditional likelihood is defined as the product of $\mathcal{L}_{\mathrm{PCL},i}$ over i, i.e., $\mathcal{L}_{\mathrm{PCL}} = \prod_{i=1}^n L_{\mathrm{PCL},i}$, and the estimating functions are defined as the derivatives of $\log(\mathcal{L}_{\mathrm{PCL}})$ with respect to θ:

$$S_{\mathrm{EE},\theta} \equiv \frac{\partial \log(\mathcal{L}_{\mathrm{PCL}})}{\partial\theta} = \sum_{i=1}^n D_i \sum_r I(R_i=r)\Bigg\{ X_i \sum_{y_i^{m_r}} \mathcal{B}_{(y_i^{o_r},y_i^{m_r})}\omega_{(y_i^{o_r},y_i^{m_r},X_i)} \\ - \frac{\sum_{j\in\mathcal{S}_i} X_j \sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + X_j \mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta)\mathcal{B}_{(y_i^{o_r},y_i^{m_r})}}{\sum_{j\in\mathcal{S}_i}\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + X_j \mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta)} \Bigg\} = 0.$$

Note that $\mathcal{L}_{\mathrm{PCL}}$ is free from ξ (or α_y) if there are no missing disease traits for any of the diseased subjects. Therefore, $\mathcal{L}_{\mathrm{PCL}}$ contains somewhat limited information regarding ξ. Hence, we shall estimate ξ from another set of estimating equations. Goetghebeur and Ryan (1995) first introduced two different sets of estimating equations in the context of missing causes of failure in the competing risk model. Here, to estimate ξ we consider another pseudo conditional likelihood $L^*_{\mathrm{PCL},i}$ such that the i-th subject has a disease of subtype $(y_i^{o_r}, y_i^{m_r})$ given that there is only one diseased subject in $\mathcal{S}_i$ without specifying the observed disease subtype. It is given as

$$\mathcal{L}^*_{\mathrm{PCL},i} \equiv \prod_r \left\{ \frac{\sum_{y_i^{m_r}} \mathrm{pr}(D_i=1, Y_i=(y_i^{o_r},y_i^{m_r})|X_i)\prod_{j\in\mathcal{S}_i\setminus\{i\}}\mathrm{pr}(D_j=0|X_j)}{\sum_{k\in\mathcal{S}_i}\sum_{(y_1,y_2)}\mathrm{pr}(D_k=1, Y=y|X_k)\prod_{j\in\mathcal{S}_i\setminus\{k\}}\mathrm{pr}(D_j=0|X_j)} \right\}^{I(R_i=r)} \\ = \prod_r \left[\frac{\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{y_i^{o_r},y_i^{m_r}}\xi + X_i \mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta)}{\sum_{j\in\mathcal{S}_i}\sum_{(y_1,y_2)} \exp(\mathcal{A}^T_{(y_1,y_2)}\xi + X_j \mathcal{B}^T_{(y_1,y_2)}\theta)} \right]^{I(R_i=r)}.$$

Hence, by defining $L^*_{\mathrm{PCL}} = \prod_{i=1}^n \mathcal{L}^*_{\mathrm{PCL},i}$, the estimating equations for ξ is

$$S_{\mathrm{EE},\xi} \equiv \frac{\partial \log(L^*_{\mathrm{PCL}})}{\partial\xi} = \sum_{i=1}^n D_i \Bigg\{ \sum_r I(R_i=r) \sum_{y_i^{m_r}} \mathcal{A}_{y_i^{o_r},y_i^{m_r}}\omega_{y_i^{o_r},y_i^{m_r}}$$

$$-\frac{\sum_{j\in\mathcal{S}_i}\sum_{(y_1,y_2)}\exp(\mathcal{A}^T_{(y_1,y_2)}\xi+X_j\mathcal{B}^T_{(y_1,y_2)}\theta)\mathcal{A}^T_{(y_1,y_2)}}{\sum_{j\in\mathcal{S}_i}\sum_{(y_1,y_2)}\exp(\mathcal{A}^T_{(y_1,y_2)}\xi+X_j\mathcal{B}^T_{(y_1,y_2)}\theta)}\Bigg\}=0.$$

We estimate θ and ξ by solving $S_{\mathrm{EE},\theta}=0$ and $S_{\mathrm{EE},\xi}=0$ simultaneously. Denote the resulting estimates as $\hat{\eta}=(\hat{\theta}^T,\hat{\xi}^T)^T$.

The estimating equations are asymptotically unbiased, which we will show in the next subsection. The asymptotic distribution of the estimators is a multivariate normal with the asymptotic covariance of $\hat{\eta}$ given by a sandwich estimator. The middle component of the sandwich estimator is obtained via a linearization technique applied to the estimating equations. The left and right multipliers of the sandwich estimator are the derivative of the estimating equations with respect to the parameters.

2.4 General Case and Asymptotics

2.4.1 General case

Suppose that $X=(X_1,\ldots,X_P)$ is a vector of P covariates, and $Y=(Y_1,\ldots,Y_K)$ carries information on K disease traits, and $M=M_1\times M_2\times\ldots\times M_K$ is the total number of disease subtypes, based on all possible combinations of the various traits. We will use y for $(y_1,\ldots,y_K)$. Our model is $p_{i,y}\equiv\mathrm{pr}(D_i=1,Y_i=y|X_i)=\exp(\alpha_y+\sum_{p=1}^P\beta_y^{(p)}X_{i,p})/\{1+\sum_y\exp(\alpha_y+\sum_{p=1}^P\beta_y^{(p)}X_{i,p})\}$, and $\mathrm{pr}(D_i=0|X_i)=1/\{1+\sum_y\exp(\alpha_y+\sum_{p=1}^P\beta_y^{(p)}X_{i,p})\}$, for $i=1,\ldots,n$. For M disease subtypes, we have $M\times P$ main regression parameters of interest along with M intercept parameters. The second-stage model for the log-odds ratio parameter is

$$\beta^{(p)}_{(y_1,\ldots,y_K)}=\theta^{(0)}{}^{(p)}+\sum_{k=1}^K\theta^{(1)}{}^{(p)}_{k(y_k)}+\sum_{k=1}^K\sum_{k'\geq k}^K\theta^{(2)}{}^{(p)}_{kk'(y_k,y_{k'})}+\cdots+\theta^{(K)}{}^{(p)}_{12\ldots K(y_1,\ldots,y_K)}. \quad (2.3)$$

Suppose that $\beta^{(p)}$ is the set of log-odds ratio parameter corresponding to X_p, then the second-stage model can be written as $\beta^{(p)}=\mathcal{B}^{(p)}\theta^{(p)}$. From here on, we denote

$(\theta^{(1)T}, \ldots, \theta^{(P)T})^T$ by θ. For each subject we introduce a vector of binary variables $R = (R_1, \ldots, R_K)^T$, where $R_k = 1$ if the k^{th} trait is observed and 0 otherwise. For our convenience, we set $R = (1, \ldots, 1)^T$ for a non-diseased subject. Using our methodology the estimating functions for θ are

$$
\begin{aligned}
S_{\mathrm{EE},\theta^{(p)}} \equiv \frac{\partial \mathrm{log}(\mathcal{L}_{\mathrm{PCL}})}{\partial \theta^{(p)}} = \sum_{i=1}^{n} D_i \sum_{r} I(R_i = r) \Big\{ X_{i,p} \sum_{y_i^{m_r}} \mathcal{B}^{(p)}_{(y_i^{o_r}, y_i^{m_r})} \omega_{(y_i^{o_r}, y_i^{m_r})} \\
- \frac{n_0^{-1} \sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r}, y_i^{m_r})} \xi + \sum_{p=1}^{P} X_{i,p} \mathcal{B}^{(p)T}_{(y_i^{o_r}, y_i^{m_r})} \theta^{(p)}) X_{i,p} \mathcal{B}^{(p)}_{y_i^{o_r}, y_i^{m_r}} + \mathcal{M}^{(1)}_{y_i^{o_r}, p}(Q_{0n})}{n_0^{-1} \sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r}, y_i^{m_r})} \xi + \sum_{p=1}^{P} X_{i,p} \mathcal{B}^{(p)T}_{(y_i^{o_r}, y_i^{m_r})} \theta^{(p)}) + \mathcal{M}^{(0)}_{y_i^{o_r}, p}(Q_{0n})} \Big\},
\end{aligned}
$$

where for $k = 0, 1$,

$$
\begin{aligned}
\mathcal{M}^{(k)}_{y_i^{o_r}, p}(Q_{0n}) &= n_0^{-1} \sum_{j \in \mathcal{S}_i / \{i\}} \exp(\mathcal{A}^T_{(y_i^{o_r}, y_i^{m_r})} \xi + \sum_{p=1}^{P} X_{j,p} \mathcal{B}^{(p)T}_{(y_i^{o_r}, y_i^{m_r})} \theta^{(p)}) (X_{j,p} \mathcal{B}^{(p)}_{y_i^{o_r}, y_i^{m_r}})^{\otimes k} \\
&= \int \sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r}, y_i^{m_r})} \xi + \sum_{p=1}^{P} X_p \mathcal{B}^{(p)T}_{(y_i^{o_r}, y_i^{m_r})} \theta^{(p)}) (X_p \mathcal{B}^{(p)}_{(y_i^{o_r}, y_i^{m_r})})^{\otimes k} dQ_{0n}(X),
\end{aligned}
$$

and $Q_{0n}(x) = n_0^{-1} \sum_{i=1}^{n} I(D_i = 0, X_i = x)$ denotes the empirical distribution function of X among the controls which converges in probability to the true distribution of X among the controls denoted by $Q_0(x)$. Here $a^{\otimes k} = 1, a, aa^T$ for $k = 0, 1, 2$, respectively. We want to clarify that $\sum_r$ in $S_{\mathrm{EE},\theta^{(p)}}$ signifies a summation over all possible values of the indicator vector r. If there are three traits, then the possible values of r are $(0,0,0), (1,0,0), (0,1,0), (1,1,0), (0,0,1), (1,0,1), (0,1,1)$ and $(1,1,1)$.

The estimating functions for ξ are

$$
\begin{aligned}
S_{\mathrm{EE},\xi} \equiv \quad \frac{\partial \mathrm{log}(L^*_{\mathrm{PCL}})}{\partial \xi} = \sum_{i=1}^{n} D_i \Big\{ \sum_{r} I(R_i = r) \sum_{y_i^{m_r}} \mathcal{A}_{y_i^{o_r}, y_i^{m_r}} \omega_{y_i^{o_r}, y_i^{m_r}} \\
- \frac{n_0^{-1} \sum_{y} \exp(\mathcal{A}_y^T \xi + \sum_{p=1}^{P} X_{i,p} \mathcal{B}_y^{(p)T} \theta^{(p)}) \mathcal{A}_y^T + \mathcal{N}^{(1)}(Q_{0n})}{n_0^{-1} \sum_{y} \exp(\mathcal{A}_y^T \xi + \sum_{p=1}^{P} X_{i,p} \mathcal{B}_y^{(p)T} \theta^{(p)}) + \mathcal{N}^{(0)}(Q_{0n})} \Big\},
\end{aligned}
$$

where for $k = 0, 1$,

$$\begin{aligned}\mathcal{N}^{(k)}(Q_{0n}) &= n_0^{-1} \sum_{j \in \mathcal{S}_i / \{i\}} \sum_y \exp(\mathcal{A}_y^T \xi + \sum_{p=1}^P X_{j,p} \mathcal{B}_y^{(p)T} \theta^{(p)}) \mathcal{A}_y^{\otimes k} \\ &= \int \sum_y \exp(\mathcal{A}_y^T \xi + \sum_{p=1}^P X_p \mathcal{B}_y^{(p)T} \theta^{(p)}) \mathcal{A}_y^{\otimes k} dQ_{0n}(X).\end{aligned}$$

We estimate $\theta^{(p)}$, $p = 1, \ldots, P$, and ξ by solving $S_{\mathrm{EE},\theta^{(p)}} = 0$, $p = 1, \ldots, P$, $S_{\mathrm{EE},\xi} = 0$ simultaneously. Denote the resulting estimator as $\hat{\eta} = (\hat{\theta}^T, \hat{\xi}^T)^T$.

2.4.2 Asymptotic properties

In this section, we discuss the large sample properties of $\hat{\eta}$. We show that $n^{-1} S_{\mathrm{EE},\theta^{(p)}} \to 0$ ($p = 1, \ldots, P$) and $n^{-1} S_{\mathrm{EE},\xi} \to 0$ in probability, i.e., the estimating equations are asymptotically unbiased.

Here we first show asymptotic unbiasedness, i.e., $n^{-1} S_{\mathrm{EE},\theta^{(p)}} \xrightarrow{P} 0$ as $n \to \infty$ at the true parameter value. Due to the law of large numbers, $n^{-1} S_{\mathrm{EE},\theta^{(p)}}$ converges to its expectation. In order to calculate this expectation, we shall use the conditional probability that the i-th subject has disease of type $y = (y_i^{o_r}, y_i^{m_r})$ given that there is one diseased subject in the matched set $\mathcal{S}_i$ with this disease type. Hence,

$$\begin{aligned}E\left(\frac{S_{\mathrm{EE},\theta^{(p)}}}{n}\right) = E\Bigg[\sum_r \int_{y_i^{o_r}} \sum_{k \in \mathcal{S}_i} &\frac{\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r}, y_i^{m_r})} \xi + \sum_{p=1}^P X_{k,p} \mathcal{B}^T_{(y_i^{o_r}, y_i^{m_r})} \theta^{(p)}) \pi(r, X_k)}{\sum_{j \in \mathcal{S}_i} \sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r}, y_i^{m_r})} \xi + \sum_{p=1}^P X_{j,p} \mathcal{B}^T_{(y_i^{o_r}, y_i^{m_r})} \theta^{(p)})} \\ \times \Bigg\{ &\frac{\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r}, y_i^{m_r})} \xi + \sum_{p=1}^P X_{k,p} \mathcal{B}^T_{(y_i^{o_r}, y_i^{m_r})} \theta^{(p)}) X_{k,p} \mathcal{B}_{(y_i^{o_r}, y_i^{m_r})}}{\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r}, y_i^{m_r})} \xi + \sum_{p=1}^P X_{k,p} \mathcal{B}^T_{(y_i^{o_r}, y_i^{m_r})} \theta^{(p)})} - \frac{\mathcal{M}^{(1)}_{y_i^{o_r}, p}(Q_0)}{\mathcal{M}^{(0)}_{y_i^{o_r}, p}(Q_0)} \Bigg\} d\mu(y_i^{o_r}) \Bigg] + o(1).\end{aligned}$$

Now, the first term on the right hand side above is

$$E\Bigg[\sum_r \int_{y_i^{o_r}} \sum_{k \in \mathcal{S}_i} \frac{\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r}, y_i^{m_r})} \xi + \sum_{p=1}^P X_{k,p} \mathcal{B}^T_{(y_i^{o_r}, y_i^{m_r})} \theta^{(p)}) \pi(r, X_k)}{\sum_{j \in \mathcal{S}_i} \sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r}, y_i^{m_r})} \xi + \sum_{p=1}^P X_{j,p} \mathcal{B}^T_{(y_i^{o_r}, y_i^{m_r})} \theta^{(p)})}$$

$$\times \frac{\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + \sum_{p=1}^P X_{k,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})X_{k,p}\mathcal{B}_{(y_i^{o_r},y_i^{m_r})}}{\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + \sum_{p=1}^P X_{k,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})} d\mu(y_i^{o_r})\Big]$$

$$= E\Big[\sum_r \int_{y_i^{o_r}} \frac{\sum_{k\in\mathcal{S}_i}\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + \sum_{p=1}^P X_{k,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})X_{k,p}\mathcal{B}_{(y_i^{o_r},y_i^{m_r})}\pi(r,X_k)}{\sum_{j\in\mathcal{S}_i}\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + \sum_{p=1}^P X_{j,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})} d\mu(y_i^{o_r})\Big],$$

and the second term is

$$E\Big[\sum_r \int_{y_i^{o_r}} \frac{\sum_{k\in\mathcal{S}_i}\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + \sum_{p=1}^P X_{k,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})\pi(r,X_k)}{\sum_{j\in\mathcal{S}_i}\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + \sum_{p=1}^P X_{j,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})} \times \frac{\mathcal{M}^{(1)}_{y_i^{o_r},p}(Q_0)}{\mathcal{M}^{(0)}_{y_i^{o_r},p}(Q_0)} d\mu(y_i^{o_r})\Big].$$

The difference between the two terms is easily seen to be asymptotically the expected weighted conditional covariance between $\pi(r, X)$ and $X_{\cdot,p}\mathcal{B}_{(y_i^{o_r},y_i^{m_r})}$ with weight

$$\frac{\exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + \sum_{p=1}^P X_{j,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})}{\sum_{j\in\mathcal{S}_i}\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + \sum_{p=1}^P X_{j,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})}.$$

Let

$$\begin{aligned}
&\mathrm{cov}^w\{\pi(r,X), X_{\cdot,p}\mathcal{B}_{(y_i^{o_r},y_i^{m_r})}\} \\
&= \sum_{j\in\mathcal{S}_i}\sum_{y_i^{m_r}} \frac{\exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + \sum_{p=1}^P X_{j,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})}{\sum_{j\in\mathcal{S}_i}\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + \sum_{p=1}^P X_{j,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})}\{\pi(r,X_j) - \bar{\pi}(r,X)\} \\
&\quad \times \Big(X_{j,p}\mathcal{B}_{(y_i^{o_r},y_i^{m_r})} - \frac{\mathcal{M}^{(1)}_{y_i^{o_r},p}(Q_0)}{\mathcal{M}^{(0)}_{y_i^{o_r},p}(Q_0)}\Big),
\end{aligned}$$

where

$$\bar{\pi}(r,X) \quad = \quad \sum_{j\in\mathcal{S}_i}\sum_{y_i^{m_r}} \frac{\exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + \sum_{p=1}^P X_{j,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})\pi(r,X_j)}{\sum_{j\in\mathcal{S}_i}\sum_{y_i^{m_r}} \exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + \sum_{p=1}^P X_{j,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})}.$$

Then, we can write

$$E\left(\frac{S_{\mathrm{EE},\theta^{(p)}}}{n}\right) \;=\; E\Big[\sum_r \int_{y_i^{o_r}} \mathrm{cov}^w\{\pi(r,X), X_{\cdot,p}\mathcal{B}_{(y_i^{o_r},y_i^{m_r})}\}d\mu(y_i^{o_r})\Big]+o(1)=o(1),$$

where the last equality follows due to the fact that $\sum_r \pi(r,X)=1$.

Similarly, due to the law of large numbers, $n^{-1}S_{\mathrm{EE},\xi}$ converges to its expectation. In order to calculate this expectation, we shall use the conditional probability that the i-th subject has disease of type $y=(y_i^{o_r},y_i^{m_r})$ given that there is one diseased subject in the matched set $\mathcal{S}_i$ but without specifying any disease subtype information. Hence,

$$\begin{aligned}
E\left(\frac{S_{\mathrm{EE},\xi}}{n}\right) &= E\Big[\int_{y_i^{o_r}}\sum_r\sum_{k\in\mathcal{S}_i}\pi(r,X_k)\frac{\sum_{y_i^{m_r}}\exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi+\sum_{p=1}^P X_{k,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})}{\sum_{j\in\mathcal{S}_i}\sum_y \exp(\mathcal{A}_y^T\xi+\sum_{p=1}^P X_{j,p}\mathcal{B}_y^T\theta^{(p)})}\\
&\quad\times\left\{\frac{\sum_{y_i^{m_r}}\exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi+\sum_{p=1}^P X_{k,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})\mathcal{A}_{(y_i^{o_r},y_i^{m_r})}}{\sum_{y_i^{m_r}}\exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi+\sum_{p=1}^P X_{k,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})}-\frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)}\right\}d\mu(y_i^{o_r})\Big]+o(1)\\
&=E\Big[\frac{\sum_r\sum_{k\in\mathcal{S}_i}\pi(r,X_k)\int_{y_i^{o_r}}\sum_{y_i^{m_r}}\exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi+\sum_{p=1}^P X_{k,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})\mathcal{A}_{(y_i^{o_r},y_i^{m_r})}d\mu(y_i^{o_r})}{\sum_{j\in\mathcal{S}_i}\sum_y\exp(\mathcal{A}_y^T\xi+\sum_{p=1}^P X_{j,p}\mathcal{B}_y^T\theta^{(p)})}\\
&\quad-\frac{\sum_r\sum_{k\in\mathcal{S}_i}\pi(r,X_k)\int_{y_i^{o_r}}\sum_{y_i^{m_r}}\exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi+\sum_{p=1}^P X_{k,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)})d\mu(y_i^{o_r})}{\sum_{j\in\mathcal{S}_i}\sum_y\exp(\mathcal{A}_y^T\xi+\sum_{p=1}^P X_{j,p}\mathcal{B}_y^T\theta^{(p)})}\times\frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)}\Big]+o(1)\\
&=E\Big[\frac{\sum_r\sum_{k\in\mathcal{S}_i}\pi(r,X_k)\sum_y\exp(\mathcal{A}_y^T\xi+\sum_{p=1}^P X_{k,p}\mathcal{B}_y^T\theta^{(p)})\mathcal{A}_y}{\sum_{j\in\mathcal{S}_i}\sum_y\exp(\mathcal{A}_y^T\xi+\sum_{p=1}^P X_{j,p}\mathcal{B}_y^T\theta^{(p)})}\\
&\quad-\frac{\sum_r\sum_{k\in\mathcal{S}_i}\pi(r,X_k)\sum_y\exp(\mathcal{A}_y^T\xi+\sum_{p=1}^P X_{k,p}\mathcal{B}_y^T\theta^{(p)})}{\sum_{j\in\mathcal{S}_i}\sum_y\exp(\mathcal{A}_y^T\xi+\sum_{p=1}^P X_{j,p}\mathcal{B}_y^T\theta^{(p)})}\times\frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)}\Big]+o(1).
\end{aligned}$$

Now using the facts that $\sum_r \pi(r,X_k)=1$ and

$$\frac{\sum_{k\in\mathcal{S}_i}\sum_y\exp(\mathcal{A}_y^T\xi+\sum_{p=1}^P X_{k,p}\mathcal{B}_y^T\theta^{(p)})\mathcal{A}_y}{\sum_{j\in\mathcal{S}_i}\sum_y\exp(\mathcal{A}_y^T\xi+\sum_{p=1}^P X_{j,p}\mathcal{B}_y^T\theta^{(p)})}\xrightarrow{P}\frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)},$$

we obtain that $n^{-1}S_{\mathrm{EE},\xi} \xrightarrow{P} 0$.

Now, we prove symptotic normality. Note that for large n

$$\begin{aligned}
\frac{1}{\sqrt{n}}S_{\mathrm{EE},\theta^{(p)}} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i \sum_{r} I(R_i{=}r)\Bigg\{X_{i,p}\sum_{y_i^{m_r}}\mathcal{B}_{(y_i^{o_r},y_i^{m_r})}\omega_{(y_i^{o_r},y_i^{m_r},X_i)} - \frac{\mathcal{M}^{(1)}_{y_i^{o_r},p}(Q_{0n})}{\mathcal{M}^{(0)}_{y_i^{o_r},p}(Q_{0n})}\Bigg\} + o(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i \sum_{r} I(R_i{=}r)\Bigg\{X_{i,p}\sum_{y_i^{m_r}}\mathcal{B}_{(y_i^{o_r},y_i^{m_r})}\omega_{(y_i^{o_r},y_i^{m_r},X_i)} - \frac{\mathcal{M}^{(1)}_{y_i^{o_r},p}(Q_0)}{\mathcal{M}^{(0)}_{y_i^{o_r},p}(Q_0)}\Bigg\} \\
&\quad - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i \sum_{r} I(R_i{=}r)\Bigg\{\frac{\mathcal{M}^{(1)}_{y_i^{o_r},p}(Q_{0n})}{\mathcal{M}^{(0)}_{y_i^{o_r},p}(Q_{0n})} - \frac{\mathcal{M}^{(1)}_{y_i^{o_r},p}(Q_0)}{\mathcal{M}^{(0)}_{y_i^{o_r},p}(Q_0)}\Bigg\} + o_p(1). \qquad (2.4)
\end{aligned}$$

Let $\hat{a}_i = \mathcal{M}^{(1)}_{y_i^{o_r},p}(Q_{0n})$ and $\hat{b}_i = \mathcal{M}^{(0)}_{y_i^{o_r},p}(Q_{0n})$. Then using the fact that

$$\frac{\hat{a}_i}{\hat{b}_i} - \frac{a_i}{b_i} = \frac{\hat{a}_i - a_i}{b_i} - \frac{a_i}{b_i^2}(\hat{b}_i - b_i) + o_p(n^{-1/2}),$$

the summand of the second term of (2.4) is

$$\begin{aligned}
&\frac{\mathcal{M}^{(1)}_{y_i^{o_r},p}(Q_{0n})}{\mathcal{M}^{(0)}_{y_i^{o_r},p}(Q_{0n})} - \frac{\mathcal{M}^{(1)}_{y_i^{o_r},p}(Q_0)}{\mathcal{M}^{(0)}_{y_i^{o_r},p}(Q_0)} \\
&\quad = \frac{1}{n\mathcal{M}^{(0)}_{y_i^{o_r},p}(Q_0)}\sum_{j=1}^{n}(1-D_j)\sum_{y_i^{m_r}}\exp(\mathcal{A}^T_{(y_i^{o_r},y_i^{m_r})}\xi + \sum_{p=1}^{P} X_{j,p}\mathcal{B}^T_{(y_i^{o_r},y_i^{m_r})}\theta^{(p)}) \\
&\quad \times\Bigg\{X_{j,p}\mathcal{B}_{(y_i^{o_r},y_i^{m_r})} - \frac{\mathcal{M}^{(1)}_{y_i^{o_r},p}(Q_0)}{\mathcal{M}^{(0)}_{y_i^{o_r},p}(Q_0)}\Bigg\} + o_p(n^{-1/2}). \qquad (2.5)
\end{aligned}$$

Plugging (2.5) into (2.4) and changing the order of the two summations in the second term, we have

$$\frac{1}{\sqrt{n}}S_{\mathrm{EE},\theta^{(p)}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{r} I(R_i = r)D_i\Bigg\{X_{i,p}\sum_{y_i^{m_r}}\mathcal{B}_{(y_i^{o_r},y_i^{m_r})}\omega_{(y_i^{o_r},y_i^{m_r},X_i)} - \frac{\mathcal{M}^{(1)}_{y_i^{o_r},p}(Q_0)}{\mathcal{M}^{(0)}_{y_i^{o_r},p}(Q_0)}\Bigg\}$$

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-D_i)\frac{1}{n}\sum_{j=1}^{n}\sum_{r} I(R_j=r)D_j\frac{1}{\mathcal{M}^{(0)}_{y_j^{or},p}(Q_0)}\sum_{y_j^{mr}}\exp(\mathcal{A}^T_{(y_j^{or},y_j^{mr})}\xi$$

$$+\sum_{p=1}^{P}X_{i,p}\mathcal{B}^T_{(y_j^{or},y_j^{mr})}\theta^{(p)})\{X_{i,p}\mathcal{B}_{(y_j^{or},y_j^{mr})}-\frac{\mathcal{M}^{(1)}_{y_j^{or},p}(Q_0)}{\mathcal{M}^{(0)}_{y_j^{or},p}(Q_0)}\}+o_p(1).$$

Finally, applying the strong law of large numbers and the Slutsky's Theorem, we obtain $n^{-1/2}S_{\mathrm{EE},\theta^{(p)}}\overset{d}{=}n^{-1/2}\sum_{i=1}^{n}\Phi_{i,\theta^{(p)}}(\theta,\xi)$ asymptotically. Similarly, for large n,

$$\begin{aligned}\frac{1}{\sqrt{n}}S_{\mathrm{EE},\xi} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{r}I(R_i=r)D_i\Big\{\sum_{y_i^{mr}}\omega_{(y_i^{or},y_i^{mr},X_i)}\mathcal{A}_{(y_i^{or},y_i^{mr})}-\frac{\mathcal{N}^{(1)}(Q_{0n})}{\mathcal{N}^{(0)}(Q_{0n})}\Big\}+o(1)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{r}I(R_i=r)D_i\Big\{\sum_{y_i^{mr}}\omega_{(y_i^{or},y_i^{mr},X_i)}\mathcal{A}_{(y_i^{or},y_i^{mr})}-\frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)}\Big\}\\ &\quad -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{r}I(R_i=r)D_i\Big\{\frac{\mathcal{N}^{(1)}(Q_{0n})}{\mathcal{N}^{(0)}(Q_{0n})}-\frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)}\Big\}+o_p(1). \qquad (2.6)\end{aligned}$$

Employing the same technique as that used in (2.5), we can write

$$\begin{aligned}\frac{\mathcal{N}^{(1)}(Q_{0n})}{\mathcal{N}^{(0)}(Q_{0n})}-\frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)} &= \frac{1}{n\mathcal{N}^{(0)}(Q_0)}\sum_{j=1}^{n}(1-D_j)\sum_{y}\exp(\mathcal{A}_y^T\xi+\sum_{p=1}^{P}X_{j,p}\mathcal{B}_y^T\theta^{(p)})\\ &\quad\times\Big\{\mathcal{A}_y-\frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)}\Big\}+o_p(n^{-1/2}). \qquad (2.7)\end{aligned}$$

Plugging (2.7) into (2.6) and changing the order of the two summations in the second term, we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}S_{\mathrm{EE},\xi}\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{r}I(R_i=r)D_i\Big\{\sum_{y_i^{mr}}\omega_{(y_i^{or},y_i^{mr},X_i)}\mathcal{A}_{(y_i^{or},y_i^{mr})}-\frac{\mathcal{N}^{(1)}(Q_{0n})}{\mathcal{N}^{(1)}(Q_{0n})}\Big\}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-D_i)\\ &\times\sum_{j=1}^{n}\sum_{r}I(R_j=r)D_j\frac{1}{n\mathcal{N}^{(0)}(Q_0)}\sum_{y}\exp(\mathcal{A}_y^T\xi+\sum_{p=1}^{P}X_{i,p}\mathcal{B}_y^T\theta^{(p)})\Big\{\mathcal{A}_y-\frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)}\Big\}+o_p(1)\end{aligned}$$

$$\overset{d}{=}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{r} I(R_i = r)D_i\Big\{\sum_{y_i^{m_r}}\omega_{(y_i^{o_r},y_i^{m_r},X_i)}\mathcal{A}_{(y_i^{o_r},y_i^{m_r})} - \frac{\mathcal{N}^{(1)}(Q_{0n})}{\mathcal{N}^{(1)}(Q_{0n})}\Big\}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-D_i)$$

$$\times E\Big[\sum_{r} I(R_j=r)D_1\frac{1}{\mathcal{N}^{(0)}(Q_0)}\sum_{y}\exp(\mathcal{A}_y^T\xi+\sum_{p=1}^{P}X_{i,p}\mathcal{B}_y^T\theta^{(p)})\Big\{\mathcal{A}_y-\frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)}\Big\}\Big]+o_p(1).$$

The last equality follows due to the application of the strong law of large numbers and Slutsky's Theorem. Thus we have shown that $S_{\mathrm{EE},\theta}$ and $S_{\mathrm{EE},\xi}$ are approximately a sum of asymptotically independent random variables whose means are zero. Now,

$$\sqrt{n}\begin{bmatrix}\hat{\theta}-\theta\\ \hat{\xi}-\xi\end{bmatrix} = H_n^{-1}\frac{1}{\sqrt{n}}S_n(\eta)+o_p(1) = H^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix}\Phi_{i,\theta}(\theta,\xi)\\ \Phi_{i,\xi}(\theta,\xi)\end{bmatrix}+o_p(1),$$

where $\theta = ({\theta^{(1)}}^T, \cdots, {\theta^{(p)}}^T)^T$,

$$H = \begin{bmatrix} H_{11} & H_{12} & \cdots & H_{1P} & H_{1\xi} \\ \vdots & & & & \\ H_{P1} & H_{P2} & \cdots & H_{PP} & H_{P\xi} \\ H_{\xi 1} & H_{\xi 2} & \cdots & H_{\xi P} & H_{\xi\xi} \end{bmatrix} = \lim_{n\to\infty} H_n,\ H_n \equiv -\frac{1}{n}\begin{bmatrix}\frac{\partial S_{\mathrm{EE},\theta}}{\partial\theta} & \frac{\partial S_{\mathrm{EE},\theta}}{\partial\xi}\\ \frac{\partial S_{\mathrm{EE},\xi}}{\partial\theta} & \frac{\partial S_{\mathrm{EE},\xi}}{\partial\xi}\end{bmatrix},$$

and the asymptotically independent terms are $\Phi_{i,\theta}(\theta,\xi) = (\Phi_{i,\theta^{(1)}}^T(\theta,\xi), \ldots, \Phi_{i,\theta^{(P)}}^T(\theta,\xi), \Phi_{i,\xi}^T(\theta,\xi))^T$. Also,

$$\begin{aligned}\Phi_{i,\theta^{(p)}}(\theta,\xi) \ =\ & \sum_{r} I(R_i = r)D_i\Big\{\sum_{y_i^{m_r}}\omega_{(y_i^{o_r},y_i^{m_r},X_i)}X_{i,p}\mathcal{B}_{(y_i^{o_r},y_i^{m_r})} - \frac{\mathcal{M}^{(1)}_{y^{o_r},p}(Q_0)}{\mathcal{M}^{(0)}_{y^{o_r},p}(Q_0)}\Big\}\\ & -(1-D_i)E\Big[\sum_{r} I(R=r)\times\frac{D}{\mathcal{M}^{(0)}_{y^{o_r}}(Q_0)}\sum_{y^{m_r}}\exp(\mathcal{A}^T_{(y^{o_r},y^{m_r})}\xi\\ & +\sum_{p=1}^{P}X_{i,p}\mathcal{B}^T_{(y^{o_r},y^{m_r})}\theta^{(p)})\Big\{X_{i,p}\mathcal{B}_{(y^{o_r},y^{m_r})} - \frac{\mathcal{M}^{(1)}_{y^{o_r},p}(Q_0)}{\mathcal{M}^{(0)}_{y^{o_r},p}(Q_0)}\Big\}|X_i\Big],\end{aligned}$$

$$\begin{aligned}\Phi_{i,\xi}(\theta,\xi) \;=\; & \sum_{r} I(R_i=r) D_i \Big\{ \sum_{y_i^{m_r}} \omega_{(y_i^{o_r}, y_i^{m_r}, X_i)} \mathcal{A}_{(y_i^{o_r}, y_i^{m_r})} - \frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)} \Big\} \\ & -(1-D_i) E\Big[\sum_{r} I(R=r) \frac{D}{\mathcal{N}^{(0)}(Q_0)} \sum_{y} \exp(\mathcal{A}_y^T \xi + \sum_{p=1}^{P} X_{i,p} \mathcal{B}_y^T \theta^{(p)}) \Big\{ \mathcal{A}_y - \frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)} \Big\} |X_i \Big],\end{aligned}$$

where Q_0 represents the true distribution of X among the controls.

So the asymptotic covariance of $\hat{\eta}$ can be consistently estimated by

$$H_n^{-1} \sum_{i=1}^{n} \begin{bmatrix} \hat{\Phi}_{i,\theta}(\hat{\theta}, \hat{\xi}) \\ \hat{\Phi}_{i,\xi}(\hat{\theta}, \hat{\xi}) \end{bmatrix}^{\otimes 2} H_n^{-T},$$

where $\hat{\Phi}^T_{i,\theta}(\hat{\theta}, \hat{\xi})$ and $\hat{\Phi}^T_{i,\xi}(\hat{\theta}, \hat{\xi})$ are obtained by replacing the expectations by the empirical averages, Q_0 by Q_{0n}, and the true parameters by their consistent estimators.

2.5 Simulation Studies

2.5.1 Simulation design and methods of analysis

One of main goals of this numerical investigation was to show how robust our method is towards a misspecification of the intercept model in the presence of partially missing disease traits. We simulated cohort data of size $n = 5,000$ by simulating (X, Y, D). The scalar covariate X was simulated from a Normal$(0, 1)$ distribution. We considered two scenarios each with 3 traits. First with 8 $(= 2 \times 2 \times 2)$ disease subtypes, and second with 30 $(= 2 \times 3 \times 5)$ disease subtypes. For each scenario we considered a correctly specified (denoted by a) second-stage model and a misspecified one (denoted by b) for the intercepts. We created missing values in each trait where missingness probabilities depended on X. Two mechanisms were used: M1) the missingness probabilities were dependent on X but the missingness of different traits was independent; and M2) the missingness probabilities were dependent on X

and the missingness of different traits was dependent. Overall disease probability lies between 6% and 9%. The details of the simulation designs are given in Appendix C.

For scenario 1, we considered three disease characteristics each with two levels, resulting in $2 \times 2 \times 2 = 8$ disease subtypes. Assuming that the second- and higher-order contrasts for the relative risk parameters are negligible, we write

$$\beta = \mathcal{B}\theta,\, \beta = \begin{bmatrix} \beta_{(1,1,1)} \\ \beta_{(1,1,2)} \\ \beta_{(1,2,1)} \\ \beta_{(1,2,2)} \\ \beta_{(2,1,1)} \\ \beta_{(2,1,2)} \\ \beta_{(2,2,1)} \\ \beta_{(2,2,2)} \end{bmatrix},\, \mathcal{B} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix},\, \theta = (\theta^{(0)}, \theta^{(1)}_{1(2)}, \theta^{(1)}_{2(2)}, \theta^{(1)}_{3(2)})^T,$$

and we chose $\theta = (0.35, 0.15, 0, 0.5)^T$. Thus the disease subtypes were generated using the model $\mathrm{pr}(Y{=}(y_1, y_2, y_3)|X){=}\exp(\alpha_{(y_1,y_2,y_3)}{+}\beta_{(y_1,y_2,y_3)}X)\{1{+}\sum_{(y_1,y_2,y_3)} \exp(\alpha_{(y_1,y_2,y_3)}+\beta_{(y_1,y_2,y_3)}X)\}^{-1}$. We chose $\alpha_{(y_1,y_2,y_3)}$ to follow the same model as $\beta_{(y_1,y_2,y_3)}$ with $\mathcal{A} = \mathcal{B}$ and $\xi = (-5, 0, 0, 0)^T$ (scenario 1a). In addition, to study the robustness of the approach against the misspecification of the model for the intercepts (scenario 1b), we used $\alpha = (-5.193, -4.477, -5.297, -5.033, -5.170, -5.160, -4.340,\ -5.330)^T$ by adding vector $(-5, -5, -5, -5, -5, -5, -5, -5)^T$ in the column space of $\mathcal{B}$, which is the correctly specified part, to vector $(-0.193, 0.523, -0.297, -0.033, -0.170, -0.160, 0.660, -0.330)^T$ perpendicular to the column space, which is the misspecified part.

Finally, we created missing values in the diseases traits using two mechanisms. For M1, the missing probabilities for each of the traits were allowed to depend on X

through the logistic function $\exp(-1.5+0.5X)\{1+\exp(-1.5+0.5X)\}^{-1}$, resulting in missingness probabilities of around 0.2 for each disease trait. For M2, 3 traits had $2^3 = 8$ possible missingness patterns. For each case subject these patterns were generated from a multinomial distribution with the following probabilities $\text{pr}\{R = (1,0,0)|X\} = d^{-1}\exp(\gamma_1 + 0.5X), \text{pr}\{R = (0,1,0)|X\} = d^{-1}\exp(\gamma_2 + 0.5X), \text{pr}\{R = (1,1,0)|X\} = d^{-1}\exp(\gamma_3 + 0.5X), \text{pr}\{R = (0,0,1)|X\} = d^{-1}\exp(\gamma_4 + 0.5X), \text{pr}\{R = (1,0,1)|X\} = d^{-1}\exp(\gamma_5 + 0.5X), \text{pr}\{R = (0,1,1)|X\} = d^{-1}\exp(\gamma_6 + 0.5X), \text{pr}\{R = (1,1,1)|X\} = d^{-1}\exp(\gamma_7 + 0.5X)$, where $d = 1 + \sum_{i=1}^{7}\exp(\gamma_i + 0.5X)$, and $\gamma_1, \ldots, \gamma_7$ were chosen so that marginally each trait had about 20% missing values.

For scenario 2, we considered three disease traits with numbers of levels 2, 3, and 5, resulting in $2 \times 3 \times 5 = 30$ disease subtypes. With the corresponding $\mathcal{A} = \mathcal{B}$ defined by the second-stage additive model, we took $\theta = (0.35, 0.15, 0, 0.5, 0.35, 0.15, 0, 0.5)^T$ and $\xi = (-5, 0, 0, 0, 0, 0, 0, 0)^T$ (scenario 2a). For scenario 2b, we chose α the same way as in scenario 1b.

Finally, we created missing values in the disease traits. For mechanism one, the missingness probabilities were allowed to depend on X through the logistic function $\exp(\gamma_k{+}0.5X)\{1{+}\exp(\gamma_k{+}0.5X)\}^{-1}$, where γ_k was chosen to be $(-1.5, -1.5, -0.85)^T$, resulting in missing probabilities of around 0.2, 0.2 and 0.3 for the three disease traits, respectively. For mechanism two, we allowed the missingness probalilities to depend on each other in a similar pattern as in scenario 1.

Each of the simulated datasets was analyzed by the maximum likelihood approach (MLE) and by the pseudo conditional likelihood method (PCL). Furthermore, we analyzed the data considering only the subjects without any missing disease traits using the maximum likelihood approach, and we refer to it as the complete-case maximum likelihood estimator (CMLE). In all these analyses, we adopted the second-stage additive models for the regression and intercept parameters, $\beta = \mathcal{B}\theta$ and

$\alpha = \mathcal{A}\xi$.

We present mean, median, median absolute deviation (MAD), empirical standard errors (Emp. SE), estimated standard errors (Est. SE), 95% coverage probabilities, and root mean square errors (RMSE) of all the methods based on 2,000 replications. To assess asymptotic bias, we present B.score = $\sqrt{2000}$(mean estimate − truth)/Emp.SE.

2.5.2 Results of the simulation studies

To save space, in both scenarios we omit the results for missingness mechanism two, which are very similar to those for mechanism one. Also, we leave out results for the correctly specified intercept model case in scenario two. The conclusions that could be drawn from the results not presented were not different from those presented here. We would be happy to provide these omitted results upon request.

Table 2.1: Simulation results for the MLE, complete-case MLE, and the pseudo conditional likelihood method. Here MAD, Emp. SE, Est. SE, Bias, B. Score, RMSE, and CP denote median absolute deviation, empirical standard error, estimated standard error, bias, bias score, root mean squared error, and 95% coverage probability based on the Wald-type confidence intervals, respectively. The results were based on 2,000 replications. There were $2 \times 2 \times 2 = 8$ disease subtypes. The missingness probabilities depended on the covariate.

Scenario 1a: Correctly Specified Model for Intercepts

	MLE				Complete-case MLE				Pseudo Conditional Likelihood Method			
	$\theta^{(0)}=0.35$	$\theta^{(1)}_{1(2)}=0.15$	$\theta^{(1)}_{2(2)}=0$	$\theta^{(1)}_{3(2)}=0.5$	$\theta^{(0)}=0.35$	$\theta^{(1)}_{1(2)}=0.15$	$\theta^{(1)}_{2(2)}=0$	$\theta^{(1)}_{3(2)}=0.5$	$\theta^{(0)}=0.35$	$\theta^{(1)}_{1(2)}=0.15$	$\theta^{(1)}_{2(2)}=0$	$\theta^{(1)}_{3(2)}=0.5$
Mean	0.354	0.147	0.001	0.498	0.357	0.149	0.000	0.497	0.351	0.148	0.002	0.500
Median	0.349	0.146	0.004	0.496	0.353	0.149	0.003	0.498	0.348	0.146	0.005	0.498
MAD	0.127	0.130	0.123	0.128	0.170	0.157	0.161	0.154	0.129	0.133	0.129	0.127
Emp. SE	0.126	0.128	0.125	0.126	0.161	0.152	0.153	0.158	0.129	0.132	0.128	0.128
Est. SE	0.129	0.125	0.124	0.129	0.163	0.153	0.152	0.159	0.131	0.128	0.128	0.133
Bias	0.004	−0.003	0.001	−0.002	0.007	−0.001	0.000	−0.003	0.001	−0.002	0.002	−0.000
B. Score	1.406	−0.904	0.455	−0.791	1.985	−0.189	0.046	−0.720	0.448	−0.550	0.693	−0.064
RMSE	0.127	0.128	0.124	0.126	0.161	0.152	0.153	0.158	0.129	0.132	0.128	0.127
CP	0.958	0.949	0.948	0.956	0.948	0.955	0.952	0.949	0.950	0.948	0.951	0.963

Scenario 1b: Misspecified Model for Intercepts

	MLE				Complete-case MLE				Pseudo Conditional Likelihood Method			
	$\theta^{(0)}=0.35$	$\theta^{(1)}_{1(2)}=0.15$	$\theta^{(1)}_{2(2)}=0$	$\theta^{(1)}_{3(2)}=0.5$	$\theta^{(0)}=0.35$	$\theta^{(1)}_{1(2)}=0.15$	$\theta^{(1)}_{2(2)}=0$	$\theta^{(1)}_{3(2)}=0.5$	$\theta^{(0)}=0.35$	$\theta^{(1)}_{1(2)}=0.15$	$\theta^{(1)}_{2(2)}=0$	$\theta^{(1)}_{3(2)}=0.5$
Mean	0.476	0.023	−0.076	0.459	0.475	0.027	−0.077	0.456	0.383	0.122	−0.017	0.482
Median	0.476	0.022	−0.075	0.457	0.478	0.031	−0.077	0.451	0.383	0.120	−0.016	0.479
MAD	0.117	0.125	0.124	0.122	0.147	0.146	0.155	0.147	0.133	0.130	0.134	0.124
Emp. SE	0.117	0.122	0.124	0.122	0.148	0.150	0.152	0.152	0.133	0.128	0.130	0.127
Est. SE	0.126	0.122	0.122	0.125	0.159	0.149	0.149	0.155	0.136	0.128	0.127	0.132
Bias	0.126	−0.127	−0.076	−0.041	0.125	−0.123	−0.077	−0.044	0.033	−0.028	−0.017	−0.018
B. Score	48.513	−46.309	−27.410	−15.099	37.776	−36.571	−22.603	−12.880	11.160	−9.918	−5.717	−6.248
RMSE	0.172	0.176	0.145	0.129	0.194	0.194	0.171	0.158	0.137	0.131	0.131	0.129
CP	0.858	0.812	0.898	0.942	0.898	0.860	0.916	0.942	0.954	0.946	0.946	0.954

Table 2.2: Simulation results for the MLE, complete-case MLE, and the pseudo conditional likelihood method. Here MAD, Emp. SE, Est. SE, Bias, B. Score, RMSE, and CP denote median absolute deviation, empirical standard error, estimated standard error, root mean squared error, bias, bias score, root mean squared error, and 95% coverage probability based on the Wald-type confidence intervals, respectively. The results were based on 2,000 runs. There were $2\times3\times5=30$ disease subtypes. The model for the intercepts was misspecified. The missingness probabilities depended on the covariate. This is Scenario 2b.

	$\theta^{(0)}=0.35$	$\theta^{(1)}_{1(2)}=0.15$	$\theta^{(1)}_{2(2)}=0$	$\theta^{(1)}_{2(3)}=0.5$	$\theta^{(1)}_{3(2)}=0.35$	$\theta^{(1)}_{3(3)}=0.15$	$\theta^{(1)}_{3(4)}=0$	$\theta^{(1)}_{3(5)}=0.5$
MLE								
Mean	0.474	0.153	0.006	0.428	0.252	0.012	−0.148	0.402
Median	0.474	0.150	0.005	0.428	0.252	0.008	−0.148	0.402
MAD	0.139	0.102	0.133	0.120	0.171	0.172	0.185	0.170
Emp. SE	0.144	0.101	0.136	0.126	0.171	0.176	0.182	0.168
Est. SE	0.147	0.102	0.139	0.126	0.171	0.178	0.181	0.166
Bias	0.124	0.003	0.006	−0.072	−0.098	−0.138	−0.148	−0.098
B. Score	38.546	1.148	1.919	−25.553	−25.665	−35.047	−36.248	−26.070
RMSE	0.190	0.101	0.136	0.146	0.197	0.223	0.235	0.194
CP	0.881	0.959	0.954	0.910	0.909	0.886	0.864	0.902
Complete-case MLE								
Mean	0.475	0.154	0.004	0.425	0.251	0.017	−0.145	0.407
Median	0.473	0.154	0.004	0.426	0.252	0.014	−0.144	0.411
MAD	0.184	0.134	0.180	0.166	0.209	0.204	0.218	0.212
Emp. SE	0.185	0.131	0.178	0.164	0.214	0.217	0.221	0.207
Est. SE	0.190	0.134	0.182	0.166	0.209	0.217	0.221	0.204
Bias	0.125	0.004	0.004	−0.075	−0.099	−0.133	−0.145	−0.093
B. Score	30.220	1.249	1.066	−20.372	−20.787	−27.353	−29.367	−20.149
RMSE	0.223	0.131	0.178	0.180	0.236	0.254	0.265	0.227
CP	0.910	0.954	0.960	0.930	0.917	0.911	0.900	0.920
Pseudo Conditional Likelihood Method								
Mean	0.381	0.160	0.001	0.476	0.327	0.119	−0.039	0.478
Median	0.379	0.156	−0.000	0.473	0.330	0.115	−0.039	0.479
MAD	0.156	0.108	0.138	0.131	0.177	0.173	0.190	0.178
Emp. SE	0.157	0.108	0.139	0.132	0.179	0.180	0.191	0.177
Est. SE	0.164	0.111	0.141	0.165	0.192	0.188	0.188	0.208
Bias	0.031	0.010	0.001	−0.024	−0.023	−0.031	−0.039	−0.022
B. Score	8.836	4.061	0.426	−8.187	−5.687	−7.594	−9.079	−5.497
RMSE	0.160	0.108	0.139	0.135	0.181	0.183	0.195	0.178
CP	0.952	0.964	0.957	0.962	0.952	0.955	0.947	0.963

The results for scenarios 1a (top panel of Table 2.1) indicate that when the intercept model is correctly specified: (1) all three methods are asymptotically unbiased; (2) the standard errors of the PCL method were slightly larger than that of the MLE method, but smaller than that of the CMLE method, which suggests that the PCL's efficiency is close to that of the MLE method; (3) similar to the standard errors, the RMSEs of the PCL method were slightly larger than that of the MLE method, but smaller than that of the CMLE method; (4) the estimated standard errors of the PCL method were close to that of the empirical standard errors; and (5) all methods' coverage probabilities were close to the nominal level (95%). The trend of the results remain the same for scenario 2a.

The results for scenarios 1b (bottom panel of Table 2.1) and 2b (Table 2.2) indicate that when the intercept model is misspecified: (1) the biases of both the MLE and the CMLE methods were prominent, but the biases of the PCL method were far less serious; (2) the comparisons of the three methods in terms of standard errors, RMSEs and estimated and empirical standard errors agreement were similar to those in the model with correctly specified model for the intercepts; and (3) the coverage probabilities of the MLE and the CMLE methods deviated from the nominal level, but the coverage probabilities of the PCL stayed close to the nominal level. Finally the PCL method was almost as efficient as the MLE method in all scenarios. The bias of the CMLE method can be attributed to model misspecification of the model for the intercepts and ignoring the subjects with missing traits. However, the main source of bias in the MLE method is due to model misspecification.

2.6 Analysis of the CPS-II Data

We now report the results of analysis by our method of the data from the CPS-II Nutrition Cohort that motivated our research. For comparison purposes, it is useful to examine the same dataset as that used in (Chatterjee et al. 2010).

Table 2.3: Results of the CPS-II Nutrition Cohort data analysis with five disease traits. In this analysis, we used weight gain as the only covariate. EST, estimate; SE, standard error; PRED, predictor; METH, method; ER, estrogen receptor; PR, progesterone receptor; WG, weight gain.

P	M			Grade		Stage	Histogoly		ER	PR
R	E		Ref.	(Well)		(Localized)	(Ductal)		(ER+)	(PR+)
E	T			Moderate	Poor	Distant	Lobular	Other	ER−	PR−
D	H		%missing	23.2		21.2	0.0		30.0	33.6
			$\theta^{(0)}$	$\theta^{(1)}_{1(2)}$	$\theta^{(1)}_{1(3)}$	$\theta^{(1)}_{2(2)}$	$\theta^{(1)}_{3(2)}$	$\theta^{(1)}_{3(3)}$	$\theta^{(1)}_{4(2)}$	$\theta^{(1)}_{5(2)}$
Analysis One: without controlling for effects of other covariates										
	CMLE	EST	1.312	−0.160	−0.033	0.703	−0.652	0.379	−0.126	−0.879
		SE	0.357	0.386	0.415	0.354	0.443	0.523	0.420	0.339
		p-value	< 0.001	0.679	0.937	0.047	0.142	0.469	0.764	0.009
WG	MLE	EST	0.961	0.040	0.268	0.795	−0.666	0.404	0.233	−0.693
		SE	0.305	0.332	0.357	0.263	0.307	0.349	0.376	0.308
		p-value	0.002	0.904	0.452	0.003	0.030	0.246	0.535	0.025
	PCL	EST	1.066	0.025	0.128	0.810	−0.685	0.368	0.883	−1.222
		SE	0.273	0.317	0.346	0.261	0.303	0.351	0.439	0.359
		p-value	< 0.001	0.937	0.711	0.002	0.024	0.294	0.045	0.001

The CPS-II Nutrition Cohort is a prospective study of cancer incidence and mortality in 86,402 men and 97,786 women and has been described in detail elsewhere (Calle et al. 2002). Briefly, the Nutrition Cohort is a subgroup of the approximately 1.2 million participants of the CPS-II Cohort, a prospective study of cancer mortality established by the American Cancer Society in 1982 (Garfinkel 1985). Nutrition Cohort participants resided in 21 states with population-based cancer registries, were aged 50-74 years, and completed a 10-page confidential, self-administered mailed questionnaire at enrollment in 1992 or 1993.

Excluded from this analysis were Nutrition Cohort participants who were men

($n = 86,402$); women who were using hormone replacement therapy ($n = 33,407$), not post-menopausal ($n = 3,514$), lost to follow-up (i.e., alive at the first follow-up questionnaire in 1997 but did not return the 1997 or any subsequent follow-up questionnaires) ($n = 2,178$), reported a personal history of cancer other than nonmelanoma skin cancer in 1992 ($n = 9,520$), reported a diagnosis of breast cancer on the first survey that could not be verified through medical or cancer registry records or an in situ breast cancer ($n = 174$), or the subjects with missing values in any of the predictor variables or whose weight gain was more than 100 lbs ($n = 7,979$). Included in the analysis were 41,014 women. There were 1,555 incident cases of breast cancer (International Classification of Disease for Oncology, Second and Third Editions site code C50) that occurred between the date of the baseline questionnaire and June 30, 2007.

We considered four predictors (covariates) in the analysis: (1) total weight change since age 18 to 1992 (WG), (2) number of live births (LB) with three categories: no live birth, 1-2 live births, 3 or more live births; (3) age at menarche (MC) with three categories: ≤ 11, $12-13$, and > 13; (4) age at menopause (MP) with four categories: < 44, $44 - 49$, $50 - 54$, and > 54.

Using $(y_1, \ldots, y_5)$ to represent levels of the five traits, stage (2 levels), histology (3 levels), estrogen receptor (2 levels), progesterone receptor (2 levels), and grade (3 levels), we can write the polytomous logistic model and the corresponding second-stage additive model as

$$\begin{aligned}
\text{pr}(D_i = 1, Y_i = (y_1, \ldots, y_5)|X_i) &= \frac{\exp(\alpha_{(y_1,\ldots,y_5)} + \sum_{p=1}^{P} \beta_{(y_1,\ldots,y_5)}^{(p)} X_{i,p})}{1 + \sum_y \exp(\alpha_{(y_1,\ldots,y_5)} + \sum_{p=1}^{P} \beta_{(y_1,\ldots,y_5)}^{(p)} X_{i,p})}, \\
\text{pr}(D_i = 0|X_i) &= \frac{1}{1 + \sum_{(y_1,\ldots,y_5)} \exp(\alpha_{(y_1,\ldots,y_5)} + \sum_{p=1}^{P} \beta_{(y_1,\ldots,y_5)}^{(p)} X_{i,p})},
\end{aligned}$$

and

$$\beta^{(p)}_{(y_1,\ldots,y_5)} = \theta^{(0)}{}^{(p)} + \theta^{(1)}{}^{(p)}_{1(y_1)} + \theta^{(1)}{}^{(p)}_{2(y_2)} + \theta^{(1)}{}^{(p)}_{3(y_3)} + \theta^{(1)}{}^{(p)}_{4(y_4)} + \theta^{(1)}{}^{(p)}_{5(y_5)},$$

for $i = 1, \ldots, n$.

First, we used weight gain as the only covariate ($P = 1$) since it has been shown to be related to risk of breast cancer in previous studies (e.g., Chatterjee et al. 2010; Zaman et al. 2010; McCullough et al. 2012), and the results are presented in Table 2.3.

Predictor WG was re-scaled to be between 0 and 1. We used the second-stage additive models for both the intercepts and regression parameters for all three methods. For the MLE and PCL methods, we used all 1,555 cases. For the CMLE approach, we used 848 cases, whose disease traits information was complete.

The results due to the PCL method indicate that (1) the estimate of $\theta^{(0)}$ due to weight gain is positive and statistically significant at the 5% level. The odds ratio for the incidence of breast cancer with well differentiated grade, localized stage, histology ductal, ER status positive and PR status positive for the 3rd quartile (45 lbs, re-scaled to be 0.476) of weight gain versus 1st quartile of weight gain (15 lbs, re-scaled to be 0.190) was 1.356 ($\exp\{(0.476 - 0.190) \times 1.066\}$, 95% confidence interval (CI): 1.164–1.580); (2) the PCL method produced statistically significant estimates of $\theta^{(1)}_{2(2)}$, $\theta^{(1)}_{3(2)}$, $\theta^{(1)}_{4(2)}$, and $\theta^{(1)}_{5(2)}$ for the covariate weight gain, which can be interpreted as follows. For a women who gained 45 pounds versus one who gained 15 pounds, the odds ratio of the disease with *distant* tumor is 1.260 (95% CI: 1.089–1.459) times the odds ratio of the disease with *localized* tumor, keeping all other traits fixed; the odds ratio of the disease with *lobular* histology is 0.822 (95% CI: 0.694–0.974) times the odds ratio of the disease with *ductal* histology, keeping all other traits fixed; the odds ratio of the disease with $ER-$ status is 1.287 (95% CI: 1.006–1.646) times the odds ratio of the disease with $ER+$ status, keeping all other traits fixed; the odds ratio of the disease with $PR-$ status is 0.705 (95% CI: 0.577–0.862) times the odds ratio of the disease with $PR+$ status, keeping all other traits fixed.

In the second analysis, we again examined weight gain but included 3 additional covariates (age at menopause, age at menarche, and number of live births), all of which have been shown to be related to risk of breast cancer (e.g., Goldman and Hatch 2000; Orgéas et al. 2008). Thus here $P = 4$. While additional factors are known to be associated with breast cancer, we chose only 3 for the purposes of this example to illustrate the ability of this method to simultaneously adjust for multiple predictors.

The associations with age at menopause, age at menarche, and number of live births in this analysis may be of etiologic relevance; however, caution should be exercised when interpreting these results since the sub-population of CPS-II was selected mainly to examine the effect of weight gain.

In analysis two (Table 2.4), the estimates of $\theta^{(0)}$ due to weight gain are positive and statistically significant at the 5% level in all three methods. Based on the PCL method, the odds ratio for the incidence of breast cancer with well differentiated grade, localized stage, histology ductal, ER status positive and PR status positive for the 3rd quartile (45 lbs, re-scaled to be 0.476) of weight gain versus 1st quartile of weight gain (15 lbs, re-scaled to be 0.190) was 1.374 ($\exp\{(0.476 - 0.190) \times 1.110\}$, 95% confidence interval (CI): 1.179–1.600). Although weight gain was associated with the overall risk of breast cancer, there was not enough evidence to show that this association varied across the subtypes, controlling for the effects of age at menopause, age at menarche, and number of live births.

Table 2.4: Results of the CPS-II Nutrition Cohort data analysis with five disease traits and four covariates. ER, estrogen receptor; PR, progesterone receptor; WG, weight gain; LB, number of live births; MC, menarche; MP, menopause.

P	M			Grade		Stage	Histogoly		ER	PR
R	E		Ref.	(Well)		(Localized)	(Ductal)		(ER+)	(PR+)
E	T			Moderate	Poor	Distant	Lobular	Other	ER-	ER+
D	H		%missing	23.2		21.2	0.0		30.0	33.6
			$\theta^{(0)}$	$\theta^{(1)}_{1(2)}$	$\theta^{(1)}_{1(3)}$	$\theta^{(1)}_{2(2)}$	$\theta^{(1)}_{3(2)}$	$\theta^{(1)}_{3(3)}$	$\theta^{(1)}_{4(2)}$	$\theta^{(1)}_{5(2)}$
	CMLE	EST	1.313	−0.074	0.107	−0.038	−0.173	0.003	0.274	0.076
		SE	0.358	0.319	0.302	0.208	0.242	0.255	0.230	0.286
		p-value	< 0.001	0.816	0.722	0.853	0.475	0.990	0.233	0.791
WG	MLE	EST	0.988	0.095	0.041	−0.019	0.010	0.078	0.297	0.252
		SE	0.306	0.270	0.258	0.179	0.203	0.218	0.196	0.238
		p-value	0.001	0.725	0.873	0.914	0.963	0.719	0.129	0.290
	PCL	EST	1.110	0.027	0.012	−0.019	0.083	0.129	0.308	0.228
		SE	0.273	0.260	0.248	0.168	0.192	0.214	0.182	0.226
		p-value	< 0.001	0.918	0.962	0.910	0.665	0.546	0.091	0.312
	CMLE	EST	−0.141	−0.024	−0.166	0.054	0.024	−0.114	0.050	0.436
		SE	0.387	0.335	0.316	0.224	0.260	0.279	0.252	0.309
		p-value	0.715	0.943	0.600	0.810	0.926	0.682	0.844	0.158
LB	MLE	EST	0.031	−0.013	−0.087	−0.013	−0.091	−0.161	−0.111	0.169
(2)		SE	0.332	0.288	0.275	0.194	0.220	0.237	0.213	0.257
		p-value	0.925	0.964	0.751	0.945	0.679	0.498	0.600	0.511
	PCL	EST	−0.002	0.051	−0.046	0.025	−0.110	−0.212	−0.151	0.110
		SE	0.316	0.308	0.297	0.199	0.225	0.252	0.211	0.259
		p-value	0.996	0.869	0.877	0.900	0.624	0.399	0.475	0.671
	CMLE	EST	0.008	0.237	−0.082	0.239	0.234	−0.307	−0.060	0.175
		SE	0.416	0.366	0.349	0.247	0.284	0.298	0.266	0.330
		p-value	0.986	0.517	0.815	0.333	0.411	0.304	0.821	0.595
LB	MLE	EST	0.279	0.118	−0.108	0.172	−0.037	−0.184	−0.137	−0.124
(3)		SE	0.357	0.311	0.299	0.212	0.243	0.254	0.227	0.281
		p-value	0.436	0.706	0.718	0.418	0.877	0.467	0.545	0.658
	PCL	EST	0.106	0.265	−0.043	0.232	−0.128	−0.314	−0.195	−0.168
		SE	0.345	0.350	0.342	0.229	0.265	0.286	0.238	0.296
		p-value	0.758	0.450	0.899	0.311	0.629	0.273	0.411	0.570
	CMLE	EST	0.740	−0.331	−0.547	−0.234	−0.134	0.328	0.318	0.552
		SE	0.354	0.288	0.275	0.207	0.239	0.274	0.247	0.282
		p-value	0.037	0.250	0.047	0.257	0.575	0.231	0.199	0.050
MC	MLE	EST	0.805	−0.507	−0.575	−0.290	−0.098	0.213	0.154	0.335
(2)		SE	0.264	0.216	0.206	0.153	0.174	0.191	0.175	0.205
		p-value	0.002	0.019	0.005	0.059	0.574	0.266	0.380	0.101
	PCL	EST	0.819	−0.534	−0.571	−0.325	−0.090	0.256	0.184	0.378
		SE	0.260	0.222	0.211	0.158	0.179	0.196	0.179	0.208
		p-value	0.002	0.016	0.007	0.039	0.615	0.192	0.305	0.070

Table 2.4 Continued

P	M			Grade		Stage	Histogoly		ER	PR
R	E		Ref.	(Well)		(Localized)	(Ductal)		(ER+)	(PR+)
E	T			Moderate	Poor	Distant	Lobular	Other	ER-	ER+
D	H		%missing	23.2		21.2	0.0		30.0	33.6
			$\theta^{(0)}$	$\theta^{(1)}_{1(2)}$	$\theta^{(1)}_{1(3)}$	$\theta^{(1)}_{2(2)}$	$\theta^{(1)}_{3(2)}$	$\theta^{(1)}_{3(3)}$	$\theta^{(1)}_{4(2)}$	$\theta^{(1)}_{5(2)}$
	CMLE	EST	−0.610	−0.311	−0.471	−0.046	0.051	−0.042	0.350	0.527
		SE	0.444	0.337	0.321	0.255	0.292	0.347	0.297	0.335
		p-value	0.169	0.355	0.142	0.856	0.861	0.903	0.239	0.115
MC	MLE	EST	−0.655	−0.039	−0.060	−0.142	−0.134	0.095	0.226	0.245
(3)		SE	0.308	0.257	0.247	0.172	0.197	0.217	0.194	0.228
		p-value	0.033	0.879	0.809	0.407	0.498	0.661	0.244	0.283
	PCL	EST	−0.669	−0.026	−0.018	−0.123	−0.090	0.114	0.232	0.201
		SE	0.303	0.264	0.255	0.175	0.201	0.220	0.198	0.231
		p-value	0.028	0.920	0.944	0.483	0.655	0.604	0.242	0.384
	CMLE	EST	0.384	0.425	0.212	−0.224	0.087	−0.044	−0.092	−0.424
		SE	0.525	0.508	0.495	0.309	0.346	0.368	0.331	0.431
		p-value	0.465	0.403	0.668	0.469	0.801	0.906	0.780	0.325
MP	MLE	EST	0.424	0.071	−0.035	−0.021	−0.121	0.221	0.190	−0.049
(2)		SE	0.350	0.310	0.299	0.206	0.241	0.249	0.229	0.285
		p-value	0.225	0.818	0.906	0.918	0.615	0.374	0.406	0.863
	PCL	EST	0.386	0.051	−0.056	−0.033	−0.144	0.214	0.178	−0.038
		SE	0.348	0.308	0.296	0.206	0.242	0.251	0.228	0.284
		p-value	0.268	0.867	0.849	0.873	0.551	0.394	0.436	0.895
	CMLE	EST	−0.113	0.283	0.379	0.158	0.313	0.365	0.057	−0.344
		SE	0.422	0.393	0.376	0.258	0.289	0.299	0.277	0.350
		p-value	0.789	0.472	0.314	0.539	0.280	0.222	0.838	0.326
MP	MLE	EST	0.243	−0.017	0.213	0.154	0.377	0.083	−0.062	−0.392
(3)		SE	0.378	0.337	0.317	0.237	0.261	0.264	0.242	0.308
		p-value	0.521	0.959	0.502	0.515	0.149	0.755	0.798	0.203
	PCL	EST	0.925	0.094	0.314	0.110	0.510	0.171	0.178	−0.110
		SE	0.465	0.505	0.463	0.278	0.328	0.350	0.308	0.374
		p-value	0.047	0.852	0.497	0.694	0.120	0.624	0.563	0.769
	CMLE	EST	−0.884	−0.238	−0.179	0.002	0.047	0.268	−0.038	−0.206
		SE	0.340	0.277	0.263	0.195	0.225	0.238	0.216	0.261
		p-value	0.009	0.390	0.497	0.991	0.836	0.260	0.858	0.429
MP	MLE	EST	−0.707	−0.175	0.037	0.033	0.072	0.043	−0.152	−0.330
(4)		SE	0.310	0.263	0.249	0.181	0.207	0.214	0.194	0.238
		p-value	0.022	0.505	0.882	0.857	0.728	0.840	0.435	0.165
	PCL	EST	−1.248	−0.222	−0.056	−0.051	−0.148	0.004	−0.205	−0.243
		SE	0.372	0.368	0.335	0.204	0.250	0.268	0.229	0.279
		p-value	0.001	0.547	0.867	0.804	0.554	0.988	0.369	0.384

2.7 Discussion

In this section, we have addressed an multivariate classification problem complicated with missing data. The two-stage model is an efficient and flexible way to measure heterogeneity of the odds ratios, and it allows a sensible way to dimension reduction. For parameter estimation of the second-stage model, one can use the MLE, PCL, or the CMLE methods. Compared with the MLE method, our method reduces the effects of the intercepts on the estimation of the regression parameters, and thus it is more robust against the misspecification of the model for the intercepts.

When the model is correct, the PCL method is asymptotically unbiased. In addition, our simulations suggest (1) when the second-stage model for the intercepts is misspecified, our bias is usually smaller than that of either the MLE method or the CMLE method, and (2) with either correctly specified or misspecified model for the intercepts, our method can usually achieve efficiency that is very close to the MLE method.

3. SEMIPARAMETRIC BAYESIAN ANALYSIS OF LOGISTIC MODELS WITH NON-STANDARD MEASUREMENT ERRORS

3.1 Brief Overview of Measurement Error Problems

Statistics is often defined as the study of the collection, organization, analysis, interpretation and presentation of data (Dodge 2006). Statistics without "good" data is just like cooking without good ingredients. Two of the commonly seen data quality issues are missing data and measurement error in data. While both issues can lead to bad consequences, the latter can be more devastating since missing data can be easily seen but measurement error problems may easily go unnoticed and neglected.

Measurement error is simply the difference between the measured value and the true value. Measurement error can be introduced into the data in many ways, including self-reporting, recall error, processing error, and instrument error. Both categorical and continuous variable can be measured with error, and in regression type of analysis, measurement errors can potentially be seen in both the independent/predictor variables as well as in the dependent/outcome variables.

Measurement errors, if neglected, can jeopardize the validity of any statistical analysis. Unfortunately, the problem created by measurement errors are ignored in the majority of statistical analysis (Schwartz 1985). The theory for measurement error in the continuous variables are fairly well developed, but the problem of measurement error in the categorical variables, which is typically more difficult to handle, seems to have drawn less attention although its consequences might be as bad as, if not worse than, the continuous case.

We will address the error in the continuous covariates case here since the error in the outcome problem can be easy to handle; for example, in linear regression the error in the outcome variable will be absorbed in the error term.

We now use an simple linear regression example to demonstrate that ignoring measurement error could bias the slope estimate. Data were generated from the model $Y = \beta_0 + \beta_1 X + \epsilon$, where $\beta_0 = beta_1 = 1$, X came from the standard normal distribution, and ϵ came from a Normal(mean $=0$, sd $=0.5$) that was independent of X. Instead of observing X, we observe $W = X+U$, and U came from a Normal$(0, 0.5)$ that was independent of X, which is the classical measurement error model. The sample size used was 20, and the simulation was done 1,000 times to evaluate the bias in the slope estimate.

The mean bias of the model regressing Y on X was 0.008, but the mean bias of the modeling regressing Y on W, which is called the naive model, was -0.194. The phenomenon where ignoring the measurement error biases the slope estimate in the direction of zero is commonly referred to as attenuation or attenuation to the null. Please note that ignoring the measurement error has other consequences, such as increased/decreased standard errors and loss of statistical power, although only the bias in the slope estimate is discussed in this example. Result of one replication in the simulation is displayed in Figure 3.1.

Various methods have been proposed to eliminate or reduce bias caused by measurement error. See Fuller (1987) for a summary of methods for linear regression and Carroll et al. (2006) for a summary of more recent methods with an emphasis on nonlinear models. Those methods fall into two broad categories — structural methods and functional methods. Structural methods rely on distributional assumptions on the true predictors $\boldsymbol{X}$, which are not observed; functional methods, on the other hand, make no such assumptions, meaning that $\boldsymbol{X}$ could be fixed (the usual definition) or random (Carroll 1998). There is no consensus on which methods are

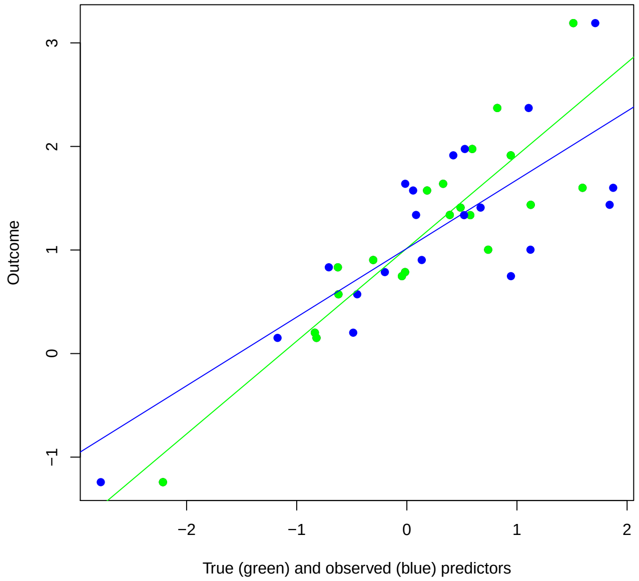


Figure 3.1: Attenuation caused by measurement error. When X is observed with additive error, the slope of the regression line is less steep in a phenomenon called attenuation.

better. In functional methods, since no assumptions are made about $\boldsymbol{X}$, there is no need to worry about model misspecification for $\boldsymbol{X}$, and those methods are generally applicable. Many such methods are based on rather intuitive ideas and often easy to implement.

In structural methods, specification of a parametric distribution for $\boldsymbol{X}$ is required, and issue of model misspecification naturally arise. However, they allow for maximum likelihood estimation-based inference and can yield gains in efficiency if the model for $\boldsymbol{X}$ is correct. In addition, structural approaches can be easily tailored to handle complicated epidemiological designs,

If a wrong model is assumed for $\boldsymbol{X}$, then the parameter estimation could become inconsistent, which is one reason why structural approaches are criticized. Methods have been proposed to evaluate the effects of model misspecification in structural measurement error models (Huang et al. 2006). One way to avoid model misspecification is to flexible models (e.g., Roeder et al. 1996; Richardson et al. 2002; Bolfarine and Lachos 2007; Hossain and Gustafson 2009). What we are proposing in this dissertation is one such method. We assume distributions for $\boldsymbol{X}$ as well as $\boldsymbol{U}$ and use very flexible models to avoid the consequences of model misspecification. The problem we are addressing is unique in that we allow for an interaction effect between variables measured with errors.

3.2 Motivating Data and Problem to Solve

Regression models with main effects and interaction effects of potential predictors are common in epidemiological studies that help to understand how the association between the response and one predictor changes with the values of the other predictors. The motivating example comes from the National Health and Nutrition Examination Survey (NHANES, CDC 2013) where we wish to study how high cholesterol level is associated with total fat and protein.

Although fat's and protein's roles have been investigated in many studies (e.g., Mensink and Katan 1989; Appel et al. 2005), we have noticed a lack of investigation on their interaction effect. Here we aim to evaluate the interaction effect along with the main effects. The main difficulty is that these nutrient intakes are measured via recalls that involve substantial amount of measurement error. Nonetheless, estimation of the main effect and interaction effect association parameters while covariates are measured with errors has not drawn much attention. In this paper, we aim to fill the gap and propose a semiparametric Bayesian method of estimation logistic regression models.

To be more specific, we use Y, $\boldsymbol{Z}$ and $\boldsymbol{X}$ to denote the binary response variable (high cholesterol level in our data example), the control covariates measured without any error, and the nutrient intakes that are not recorded in the data, respectively. Instead of $\boldsymbol{X}$, an error-prone surrogate for $\boldsymbol{X}$ is observed, and we denote the surrogate by $\boldsymbol{W}$. Suppose that there are two nutrient intakes, i.e., $\boldsymbol{X} = (X_1, X_2)^T$, and the regression model is

$$\text{logit}\{\text{pr}(Y = 1|\boldsymbol{X}, \boldsymbol{Z})\} = \beta_0 + \beta_{11}X_1 + \beta_{12}X_2 + \beta_2 X_1 X_2 + \boldsymbol{\beta}_3^T \boldsymbol{Z},$$

While the nutrient intakes are generally non-negative and the product of two non-negative variables is correlated with each of the two, the intakes are usually centered in numerical analysis to reduce multicollinearity. Our interest is in estimating the regression parameters. In our data example, X_1 and X_2 will denote the true fat and protein intakes that were not observed. The term X_1X_2 is included in the model to capture the potential interaction effect, and for a non-zero β_2, $(\beta_{11} + \beta_2 X_2)$ measures the degree of association between the response and X_1 while X_2 is fixed. If the interaction parameter is non-zero, any analysis ignoring the interaction term may lead to erroneous conclusions.

Here $\boldsymbol{W} = (W_1, W_2)^T$. We assume an additive non-differential measurement error model, that is, for $j = 1, 2$, $W_j = X_j + U_j$, where the measurement error U_j is assumed to be independent of $\boldsymbol{X}$, $\boldsymbol{Z}$ and Y and have mean zero and finite (homogeneous) variance. Then $W_1W_2 = X_1X_2 + U_{12}$, where $U_{12} = X_1U_2 + X_2U_1 + U_1U_2$. Hence, the measurement error U_{12} is not independent of X_1X_2, resulting in heteroscedastic measurement errors that depend on the unobserved truth, X_1 and X_2.

Two widely used measurement error methods are regression calibration (RC), which replaces $\boldsymbol{X}$ with an estimate based on other covariates but not on Y (Spiegelman et al. 1997), and simulation extrapolation (SIMEX), which adds artificial errors to the observed values and then extrapolates back to the error-free case (Stefanski and Cook 1995). Both of the methods are consistent in special cases such as linear regression and loglinear mean models, but they are not consistent in general (Carroll et al. 2006; Gustafson 2004). Murad and Freedman (2007) considered the estimation of interaction parameter in linear regression models, but their method does not apply to the logistic regression case.

Two examples of the functional approach are the conditional-score method (Stefanski and Carroll 1987) and the corrected-score method (Nakamura 1990). The corrected score and conditional score methods deal with the linear logistic model (i.e., only the main effects are present), and they make no assumptions on the distribution of the unobserved $\boldsymbol{X}$. These methods critically depend on the linear logistic structure of the model and the additive structure of the measurement error, so they methods are not applicable in our context. Even more, the conditional score approach critically depends on the normal distribution assumption of the measurement error.

To develop a unified method for handling measurement error and partially missing surrogate variables, we consider a flexible structural model which, we believe, is lacking in the literature. The model is flexible in the sense that both the distribution of the measurement error and the unobserved true nutrients are modeled nonparametrically using a mixture of Dirichlet process prior (DPP). By modeling the distribution of $\boldsymbol{X}$ as a Dirichlet process mixture of a multivariate normal kernel, we in principle can capture any continuous distribution. We model the distribution of the measurement error as a Dirichlet process mixture of a symmetric kernel, which is again a flexible model to capture any distribution symmetric about zero. For efficient computation, we adopt Ishwaran and James' (2001) approximation of the stick-breaking presentation of the Dirichlet process prior. All model parameters are estimated in a fully Bayesian framework using the Markov Chain Monte Carlo (MCMC) method. Although the DPP has been used previously in the measurement error context (Johnson et al. 2007; Sinha et al. 2010), use of DPP in modeling the distributions of $\boldsymbol{X}$ as well as $\boldsymbol{U}$, the symmetric measurement errors, is completely new. Above all, this is the first attempt to investigate the effect of measurement errors on the estimation of the interaction parameter in non-linear models. In addition, the proposed flexible method can handle missing values in the surrogate variable.

3.3 Model and Notation

In general, the observed data can be presented as $(Y_i, \boldsymbol{W}_i^{(1)}, \ldots, \boldsymbol{W}_i^{(R)}, \boldsymbol{Z}_i, \boldsymbol{\Delta}_i^{(1)}, \ldots, \boldsymbol{\Delta}_i^{(R)})$ for $i = 1, \ldots, n$, where Y_i is the binary response variable and $\boldsymbol{Z}_i$ is a length-Q vector of control variables for the i^{th} subject. $\boldsymbol{X} = (X_1, \ldots, X_P)^T$ represents the true intakes of P nutrients of interest. Since $\boldsymbol{X}$ is not actually measured, some proxy of $\boldsymbol{X}$ is measured via up to R 24-hour food recalls, which are denoted by $\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(R)}$. In the NHANES data, for example, each subject had 2 recalls, so $R = 2$.

The dimensions of $\boldsymbol{X}$ and $\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(R)}$ are all equal to P. The missingness indicators $\boldsymbol{\Delta}_{ip}^{(r)}$, for $r = 1, \ldots, R$ and $p = 1, \ldots, P$, are defined as 1 if the p^{th} component of the r^{th} recall is available for the i^{th} subject and 0 otherwise, with $\boldsymbol{\Delta}_i^{(r)} = (\boldsymbol{\Delta}_{i1}^{(r)}, \ldots, \boldsymbol{\Delta}_{iP}^{(r)})$. We assume that missingness mechanism does not depend on the value of the missing variable $\boldsymbol{X}$ or the regression parameters β_0, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_3$. Note that $\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(R)}$ are considered to be unbiased surrogates for the actual long-term intake $\boldsymbol{X}$. We assume that conditional on $\boldsymbol{X}$, $\boldsymbol{W}^{(r)}$ is independent of $\boldsymbol{Z}$, for $r = 1, \ldots, R$, and write $\boldsymbol{W}_i^{(r)} = \boldsymbol{X}_i + \boldsymbol{U}_i^{(r)}$, where $\boldsymbol{U}_i^{(1)}, \ldots, \boldsymbol{U}_i^{(R)}$ are independent and identically distributed (iid) copies of $\boldsymbol{U}$, which follows a distribution with mean 0 and finite variance. We assume that the error is non-differential, i.e., its distribution does not depend on Y if $\boldsymbol{X}$ were observed. Moreover, we denote the observed parts of $\boldsymbol{W}_i^{(r)}$ as $\boldsymbol{W}_{i\text{obs}}^{(r)}$ and its un-observed parts as $\boldsymbol{W}_{i\text{miss}}^{(r)}$ for $i = 1, \ldots, n$, and $r = 1, \ldots, R$. The assumed model for Y given $\boldsymbol{X}$ and $\boldsymbol{Z}$ is

$$\text{pr}(Y = 1|\boldsymbol{X}, \boldsymbol{Z}) = H(\beta_0 + \sum_{p=1}^{P} \beta_{1p} X_p + \sum_{p_1=1}^{P} \sum_{p_2=p_1+1}^{P} \beta_{2p_1p_2} X_{p_1} X_{p_2} + \sum_{q=1}^{Q} \beta_{3q} Z_q), \quad (3.1)$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$. Here β_0 is the intercept term, $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1P})^T$ represents the main effects of $\boldsymbol{X}$, $\boldsymbol{\beta}_2$ is the vector of all $\beta_{2p_1p_2}$'s where $\beta_{2p_1p_2}$ represents the two-factor interaction of X_{p_1} and X_{p_2} for $p_2 > p_1$, and β_{3q} represents the effect of the control variable Z_q for $q = 1, \ldots, Q$. Under this model, the odds ratio of the disease for changing X_p from X_p^* to X_p^{**} by holding X_k fixed to X_k^* for $k \neq p$ is

$$\exp\{\beta_{1p}(X_p^{**} - X_p^*) + (X_p^{**} - X_p^*) \sum_{k=1}^{p-1} \beta_{2kp} X_k^* + (X_p^{**} - X_p^*) \sum_{k=p+1}^{P} \beta_{2pk} X_k^*\}.$$

Note that this odds ratio depends on the values of X_k^* $(k \neq p)$ as well as the change in X_p.

Our analysis setting has some nonstandard features. First, this is a problem of errors-in-covariates. We have internal calibration data as we have repeated measurements of the erroneous surrogate for X. In addition, we face the problem of missing data as some of the subjects do not have information on every recall. In particular, here we face a monotone missing data pattern. Following the terminology of measurement error literature, our approach is "structural" by assuming a distribution for $\boldsymbol{X}$. However, our modeling techniques can potentially capture any continuous distribution for $\boldsymbol{X}$, resulting in a robust procedure.

3.4 Likelihood and Priors

The key component of the Bayesian inference is the likelihood. For our design the observed data likelihood is

$$L_o = \prod_{i=1}^{n} \int \mathrm{pr}(Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i) \left\{ \prod_{r=1}^{R} \int f(\boldsymbol{W}_{i\mathrm{obs}}^{(r)}, \boldsymbol{W}_{i\mathrm{miss}}^{(r)}|\boldsymbol{X}_i, \boldsymbol{Z}_i) d\boldsymbol{W}_{i\mathrm{miss}}^{(r)} \right\} f(\boldsymbol{X}_i|\boldsymbol{Z}_i) d\boldsymbol{X}_i.$$

Modeling the distribution of $\boldsymbol{X}$ given $\boldsymbol{Z}$ is challenging as any mispecification of this model may result in biased parameter estimates. To circumvent the issue of robustness, we model this distribution as a Dirichlet Process (DP) mixture of multivariate normal distributions. Using the stick-breaking presentation, DP can be seen as a discrete probability measure with infinitely many random mass points with random probability masses. Thus, the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Z}$ is a mixture of infinitely many multivariate normals. However, Ishwaran and James (2001) showed that a DP with infinitely many random mass points can be reasonably approximated by a discrete measure with finitely many random mass points. That means if a probability measure $\mathcal{P}$ follows a DP with base measure H and precision parameter α, denoted by $\mathcal{P} \sim DP(\alpha H)$, then following the Ishwaran and James approximation we can write $\mathcal{P}(\cdot) \approx \sum_{i=1}^{K} p_k \delta_{V_k}(\cdot)$ for some sufficiently large K, where $\delta_{V_k}(\cdot)$ denotes a measure concentrated at V_k, V_k are iid from a distribution H, and p_k's are random probabilities such that $0 \leq p_k \leq 1$ and $\sum_{k=1}^{K} p_k = 1$. When $(p_1, \ldots, p_K) \sim \mathrm{Dirichlet}(\alpha/K, \ldots, \alpha/K)$, $\mathcal{P}$ is called a finite-dimensional Dirichlet process prior (Ishwaran and Zarepour 2002), and the approximated DP is denoted by $DP_N(\alpha H)$. Theorem 2 of Iswaran and Zarepour (2002) states that for any real valued measurable integrable function g, $DP_K(\alpha H)(g) \rightarrow DP(\alpha H)(g)$ in distribution as $K \rightarrow \infty$. They also described a convenient mechanism of selecting K. For notational convenience, we add a scalar 1 to the beginning of $\boldsymbol{Z}$ and denote it $\boldsymbol{Z}^*$.

We assume that $[\boldsymbol{X}|\boldsymbol{Z}^*, \boldsymbol{\mu}_x, \Sigma_x] \sim N_p(\Gamma \boldsymbol{Z}^* + \boldsymbol{\mu}_x, \Sigma_x)$, and then

$$\begin{aligned}
\boldsymbol{X}|\boldsymbol{\mu}_x, \Sigma_x &\overset{\text{indep}}{\sim} \frac{1}{(2\pi)^{P/2}|\Sigma_x|^{0.5}}\left[\exp\{-0.5(\boldsymbol{X}-\boldsymbol{\mu}_x-\Gamma\boldsymbol{Z}^*)^T\Sigma_x^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_x-\Gamma\boldsymbol{Z}^*)\}\right],\\
\boldsymbol{\mu}_x, \Sigma_x|\mathcal{P}_x &\overset{\text{iid}}{\sim} \mathcal{P}_x,\\
\mathcal{P}_x &\sim DP_L(\alpha_x H_x) \ \text{ for large integer } L.
\end{aligned}$$

Hence we can write

$$\begin{aligned}
f_x(\boldsymbol{X}|p_{1x},\ldots,p_{Lx}) =& \sum_{k=1}^{L} p_{kx}\frac{1}{(2\pi)^{P/2}|\Sigma_{kx}|^{0.5}}\\
&\times\left[\exp\{-0.5(\boldsymbol{X}-\boldsymbol{\mu}_{kx}-\Gamma\boldsymbol{Z}^*)^T\Sigma_{kx}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_{kx}-\Gamma\boldsymbol{Z}^*)\}\right],\\
(p_{1x},\ldots,p_{Lx}) \sim& \,\text{Dirichlet}(\alpha_x/L,\ldots,\alpha_x/L),\\
\boldsymbol{\mu}_{kx}, \Sigma_{kx} \overset{iid}{\sim}& \, H_x.
\end{aligned}$$

Under the base probability measure H_x, we assume that $\boldsymbol{\mu}_x$ follows a multivariate normal with mean $\boldsymbol{m}_x$ and variance $\text{Diag}(\boldsymbol{\tau}_x^{1/2})\Sigma_x\text{Diag}(\boldsymbol{\tau}_x^{1/2})$, and Σ_x follows an Inverse-Wishart$_P(\nu_x, D_x)$, i.e., $f(\Sigma_x|\nu_x, D_x) = |D_x|^{-\nu_x/2}2^{-\nu_x P/2}\{\Gamma_P(\nu_x/2)\}^{-1}$ $|\Sigma_x|^{-(\nu_x+P+1)/2)}\exp\{-0.5\text{tr}(D_x^{-1}\Sigma_x^{-1})\}$, where $\Gamma_P(\ )$ is the multivariate gamma function. The DP mixture model results in the marginal density of $\boldsymbol{X}$ as a finite dimensional Dirichlet process mixture of the kernel $f(\boldsymbol{x}|\boldsymbol{\mu}_x, \Sigma_x)$, where the parameters of the component density are coming from H_x.

Now we model the distribution of U as a finite dimensional Dirichlet process mixture of a symmetric kernel. That means

$$\begin{aligned}
\boldsymbol{U}|\boldsymbol{\mu}_u, \Sigma_u \overset{\text{indep}}{\sim} \frac{1}{2(2\pi)^{P/2}|\Sigma_u|^{0.5}}\Big[&\exp\{-0.5(\boldsymbol{U}+\boldsymbol{\mu}_u)^T\Sigma_u^{-1}(\boldsymbol{U}+\boldsymbol{\mu}_u)\} +\\
&\exp\{-0.5(\boldsymbol{U}-\boldsymbol{\mu}_u)^T\Sigma_u^{-1}(\boldsymbol{U}-\boldsymbol{\mu}_u)\}\Big],
\end{aligned}$$

$$
\begin{aligned}
\boldsymbol{\mu}_u, \Sigma_u|\mathcal{P}_u &\overset{\text{iid}}{\sim} \mathcal{P}_u,\\
\mathcal{P}_u &\sim DP_M(\alpha_u H_u) \text{ for large integer } M.
\end{aligned}
$$

Hence we can write

$$
\begin{aligned}
f_u(\boldsymbol{U}|p_{1u},\ldots,p_{Mu}) &= \sum_{k=1}^{M} \frac{p_{ku}}{2(2\pi)^{P/2}|\Sigma_{ku}|^{1/2}} \Big[\exp\{-0.5(U+\boldsymbol{\mu}_{ku})^T\Sigma_{ku}^{-1}(\boldsymbol{U}+\boldsymbol{\mu}_{ku})\} + \\
&\qquad \exp\{-0.5(\boldsymbol{U}-\boldsymbol{\mu}_{ku})^T\Sigma_{ku}^{-1}(\boldsymbol{U}-\mu_{ku})\}\Big],\\
(p_{1u},\ldots,p_{Mu}) &\sim \text{Dirichlet}(\alpha_u/M,\ldots,\alpha_u/M),\\
\mu_{ku}, \Sigma_{ku} &\overset{iid}{\sim} H_u.
\end{aligned}
$$

Under the base probability measure H_u, we assume that $\boldsymbol{\mu}_u$ follows a multivariate normal with mean $\boldsymbol{m}_u$ and variance $\text{Diag}(\boldsymbol{\tau}_u^{1/2})\Sigma_u\text{Diag}(\boldsymbol{\tau}_u^{1/2})$, and Σ_u follows an Inverse-Wishart$_P(\nu_u, D_u)$, i.e., $f(\Sigma_u|\nu_u, D_u) = |D_u|^{-\nu_u/2}2^{-\nu_u P/2}\{\Gamma_P(\nu_u/2)\}^{-1}$ $|\Sigma_u|^{-(\nu_u+P+1)/2)}\exp\{-0.5\text{tr}(D_u^{-1}\Sigma_u^{-1})\}$. The DP mixture model results in the marginal density of $\boldsymbol{U}$ as a finite dimensional Dirichlet process mixture of the symmetric kernel $f(\boldsymbol{U}|\boldsymbol{\mu}_u, \Sigma_u)$, where the parameters of the component density are coming from H_u.

We put normal$(0, \sigma^2_{\beta_0})$, normal$(0, \sigma^2_{\beta_1} I_P)$, normal$(0, \sigma^2_{\beta_2} I_{P(P-1)/2})$, normal$(0, \sigma^2_{\beta_3} I_Q)$ priors on β_0, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$, respectively. On each row of Γ, say $\boldsymbol{\Gamma}_j$, we use normal($\boldsymbol{\mu}_\gamma$, $\sigma^2_\gamma I_Q$) distribution, for $j = 1,\ldots,P$. On α_x and α_u, we use Gamma(a_γ, b_γ) prior. We further assume that a priori each component of $\boldsymbol{\tau}_x \sim$ IG(g_x, h_x), each component of $\boldsymbol{\tau}_u \sim$ IG(g_u, h_u), where IG means inverse gamma and $\pi(\tau_{px}|g_x, h_x) = g_x^{-h_x}\tau_{px}^{-h_x-1}\exp\{-(g_x\tau_{px})^{-1}\}$ $/\Gamma(h_x)$ for $p = 1,\ldots,P$. For the scale matrices D_u and D_x, we put IW(P, I_P) priors. The values of the prior parameters can be specified by practitioners or one may put hyper-prior on these prior parameters. We shall fix the values of $\sigma^2_{\beta_0}$, $\sigma^2_{\beta_1}$, $\sigma^2_{\beta_2}$, $\sigma^2_{\beta_3}$

g_u, h_u, g_x, h_x, μ_γ, $\boldsymbol{m}_x$, and $\boldsymbol{m}_u$.

Remark 1. Now we model the distribution of $\boldsymbol{U}$ as a finite dimensional Dirichlet process mixture of a symmetric kernel. The symmetry assumption on the distribution of $\boldsymbol{U}$ is the sufficient condition for identifiability of the distribution of $\boldsymbol{X}$ if the number of replicated measurements of the error-prone surrogate is at least 2. Let χ_u, χ_w, χ_x be the characteristic functions of $\boldsymbol{U}$, $\boldsymbol{W}$, and $\boldsymbol{X}$, respectively. Then $\chi_x = \chi_w/\chi_u$. In principle, χ_w is directly estimable from the data, and $E[\exp\{it(\boldsymbol{W}^{(1)} - \boldsymbol{W}^{(2)})\}] = E[\exp\{it(\boldsymbol{U}^{(1)} - \boldsymbol{U}^{(2)})\}] = \chi_u^2$ under the symmetry of $\boldsymbol{U}$ where $i = \sqrt{-1}$. Hence χ_x is estimable, so is the density of $\boldsymbol{X}$.

3.5 Posterior Computation

Inference of the parameters is based on the posterior distribution, and the summaries of the posterior distribution are made by drawing random numbers from the posterior distribution using the Markov chain Monte Carlo method. The symmetric kernel used in modeling the distribution of $\boldsymbol{U}_i$ is written as

$$f(\boldsymbol{U}|\Sigma_u, \boldsymbol{\mu}_u) = \frac{1}{2} f(U|\Sigma_u, \boldsymbol{\mu}_u, \psi = 1) + \frac{1}{2} f(\boldsymbol{U}|\Sigma_u, \boldsymbol{\mu}_u, \psi = -1),$$

where

$$f(\boldsymbol{U}|\Sigma_u, \boldsymbol{\mu}_u, \psi) = \frac{1}{(2\pi)^{P/2}|\Sigma_u|^{0.5}} \exp\{-0.5(\boldsymbol{U} - \psi\boldsymbol{\mu}_u)^T \Sigma_u^{-1} (\boldsymbol{U} - \psi\boldsymbol{\mu}_u)\}.$$

For consistency, the same i, p, r notations introduced for Δ will be used for W, U and ψ. The posterior computation is done by drawing each of the parameters and the latent unobserved variables $\boldsymbol{X}_i$, $i = 1, \ldots, n$ from their full conditional distributions. In addition, if $\Delta_{iq}^{(r)} = 0$, we sample $W_{iq}^{(r)}$ from its full conditional distribution.

For $\boldsymbol{X}$, define $\Theta = (\theta_1, \ldots, \theta_n)$, where $\theta = (\boldsymbol{\mu}_x, \Sigma_x)$. Let ϕ contain the distinct elements of Θ, which makes ϕ a list of L distinct elements. Let $\boldsymbol{s}$ be an n-vector of configuration indicators, i.e., $s_i = j$ if $\theta_i = \phi_j$, $j = 1, \dots, L$. We also define m_j as the number of s_i's equal to j. Therefore, $\sum_{j=1}^{L} m_j = n$. Since knowing $\boldsymbol{s}$ and ϕ is equivalent to knowing Θ, Θ is updated via resampling $\boldsymbol{s}$ and ϕ.

Similarly, for handling $(\boldsymbol{\mu}_u, \Sigma_u)$, we define $N = R \times n$, $\Upsilon = (\vartheta_1, \dots, \vartheta_N)$, where $\vartheta = (\boldsymbol{\mu}_u, \Sigma_u)$, and suppose that φ contains distinct elements of Υ. The N-vector of configuration indicators $\boldsymbol{t}$ is defined for $\boldsymbol{U}$. Also we define n_j as the number of t_i's equal to j. Therefore, $\sum_{j=1}^{M} n_j = N$.

We propose the following **17**-step algorithm to sample from the posterior distribution.

Step 0. We initialize all the parameters and the unobserved $\boldsymbol{X}_i$, $i = 1, \dots, n$.

Step 1. Update the β parameters using the Metropolis Hastings (MH) algorithm. Draw proposed values (β^*) from a (multivariate) normal distribution with the current value as the mean and the variance/covariance of the naive estimate as the variance/covariance. For each β, accept the proposed value with probability $\min\{1, f_\beta(\beta^*|\text{rest}) / f_\beta(\beta|\text{rest})\}$, where f_β is proportional to the product of the prior density of β and $\prod_{i=1}^{n} \exp(Y_i\eta_i)/\{1+\exp(\eta_i)\}$ with $\eta_i = \beta_0 + \sum_{p=1}^{P} \beta_{1p} X_{ip} + \sum_{p1=1}^{P} \sum_{p2=p1+1}^{P} \beta_{2p_1p_2} X_{ip_1} X_{ip_2} + \sum_{q=1}^{Q} \beta_{3q} Z_{iq}$.

Step 2. Update $\boldsymbol{X}_i$ using the Metropolis algorithm. Draw proposed value $\boldsymbol{X}_i^*$ from a multivariate normal distribution with mean $\boldsymbol{X}_i$, which is the current value, and covariance $\text{cov}(\boldsymbol{W}^{(1)}) - \text{cov}(\boldsymbol{W}^{(1)} - \boldsymbol{W}^{(2)})/2$. Accept the proposed value with probability $\min\{1, f_x(\boldsymbol{X}_i^*|\text{rest})/f_x(\boldsymbol{X}_i|\text{rest})\}$, where

$$\begin{aligned} f_x(\boldsymbol{X}_i|\text{rest}) \propto & \frac{\exp(Y_i\eta_i)}{1+\exp(\eta_i)} \\ & \times \frac{1}{(2\pi)^{P/2}|\Sigma_{x_i}|^{0.5}} \left[\exp\{-0.5(\boldsymbol{X}_i - \boldsymbol{\mu}_{x_i} - \Gamma\boldsymbol{Z}_i^*)^T \Sigma_{x_i}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}_{x_i} - \Gamma\boldsymbol{Z}_i^*)\} \right] \end{aligned}$$

$$\times \prod_{j=1}^{R} \frac{1}{(2\pi)^{P/2}|\Sigma_{u_i^{(r)}}|^{0.5}}$$

$$\times \exp\{-0.5(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i - \psi_{ij}\boldsymbol{\mu}_{u_i^{(r)}})^T \Sigma_{u_i^{(r)}}^{-1}(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i - \psi_{ij}\boldsymbol{\mu}_{u_i^{(r)}})\}.$$

Step 3. Sample $\psi_i^{(r)}$ from the following distribution

$$\begin{aligned}\mathrm{pr}(\psi_i^{(r)} = 1|\mathrm{rest}) &= 1 - \mathrm{pr}(\psi_i^{(r)} = -1|\mathrm{rest}) \\ &= \frac{\exp\{-0.5(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i - \boldsymbol{\mu}_{u_i^{(r)}})^T \Sigma_{u_i^{(r)}}^{-1}(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i - \boldsymbol{\mu}_{u_i^{(r)}})\}}{\sum_{k=-1,1}\exp\{-0.5(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i - k\boldsymbol{\mu}_{u_i^{(r)}})^T \Sigma_{u_i^{(r)}}^{-1}(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i - k\boldsymbol{\mu}_{u_i^{(r)}})\}}\end{aligned}$$

for $i = 1, \ldots, n$ and $r = 1, \ldots, R$.

Step 4. If $\Delta_{ij}^{(r)} = 0$, sample $W_{ij}^{(r)}$ from the Normal$\{(X_{ij} + \psi_i^{(r)}\mu_{u_{ij}^{(r)}}), \Sigma_{u_{jj}^{(r)}}\}$ distribution for $r = 1, \ldots, R$, $i = 1, \ldots, n$, and $j = 1, \ldots, P$.

Step 5. Sample $\boldsymbol{\Gamma}_p$ from a normal distribution with mean and variance

$$\begin{aligned}\boldsymbol{m} &= v\left\{-\sum_{i=1}^{n}\boldsymbol{Z}_i^* \sum_{k=1,k\neq p}^{P} A_{i,pk}[\boldsymbol{\Gamma}_k \boldsymbol{Z}_i^* - (X_{i,k} - \mu_{X_{ik}})] + \sum_{i=1}^{n}\boldsymbol{Z}_i^* A_{i,pp}(X_{ip} - \mu_{x_{ip}})\right\}, \\ v &= \left\{\sum_{i=1}^{n} A_{i,pp}\boldsymbol{Z}_i\boldsymbol{Z}_i^T + \frac{I_Q}{\sigma_\gamma^2}\right\}^{-1},\end{aligned}$$

for $p = 1, \ldots, P$. Here $A_i \equiv \Sigma_{X_i}^{-1}$, and the $(p,k)^{th}$ component of A_i is denoted by $A_{i,pk}$.

Step 6. Precision parameter α_x is drawn from its full conditional

$$\pi(\alpha_x|\mathrm{rest}) \propto \frac{\Gamma(\alpha_x)}{\{\Gamma(\alpha_x/L)\}^L}(p_{1x})^{\alpha_x/L-1}\cdots(p_{Lx})^{\alpha_x/L-1}\pi(\alpha_x).$$

To draw α_x we shall use a Metropolis-Hastings algorithm with $\pi(\alpha_x)$ as the proposal density. Suppose that at the t^{th} iteration we draw $\alpha_x^{(new)}$ from $\pi(\alpha_x)$. Then

$$\alpha_x^{(t+1)} = \begin{cases} \alpha_x^{(new)} \text{ with probability } \rho(\alpha_x^{(new)}, \alpha_x^{(t)}) \\ \alpha_x^{(t)} \text{ otherwise} \end{cases},$$

where

$$\rho(\alpha_x^{(new)}, \alpha_x^{(t)}) = \frac{(p_{1x})^{\alpha_x^{(new)}/L-1} \times \cdots \times (p_{Lx})^{\alpha^{(new)x}/L-1}\Gamma(\alpha_x^{(new)})/\{\Gamma(\alpha_x^{(new)}/L)\}^L}{(p_{1x})^{\alpha_x^{(t)}/L-1} \times \cdots \times (p_{Lx})^{\alpha_x^{(t)}/L-1}\Gamma(\alpha_x^{(t)})/\{\Gamma(\alpha_x^{(t)}/L)\}^L}.$$

Step 7. Similarly, draw α_u from

$$\pi(\alpha_u|\text{rest}) \propto \frac{\Gamma(\alpha_u)}{\{\Gamma(\alpha_u/M)\}^M}(p_{1u})^{\alpha_u/M-1} \cdots (p_{Mu})^{\alpha_u/M-1}\pi(\alpha_u).$$

Step 8. Sample s_i from Multinomial$(p_{1x}^*, \cdots, p_{Lx}^*)$, where

$$(p_{1x}^*, \cdots, p_{Lx}^*) \propto (p_{1x} f_x(\boldsymbol{X}_i|\boldsymbol{Z}_i, \Gamma, \phi_1), \ldots, p_{Lx} f_x(\boldsymbol{X}_i|\boldsymbol{Z}_i, \Gamma, \phi_L))$$

and $f_x(X_i|Z_i, \Gamma, \phi_j) = (2\pi)^{-P/2}|\phi_{j,2}|^{-0.5}[\exp\{-0.5(\boldsymbol{X} - \boldsymbol{\phi}_{j,1} - \Gamma Z)^T \phi_{j,2}^{-1}(\boldsymbol{X} - \boldsymbol{\phi}_{j,1} - \Gamma Z)\}]$.

Step 9. Sample $(p_{1x}, \ldots, p_{Lx})$ from Dirichlet$(\alpha_x/L + m_1, \ldots, \alpha_x/L + m_L)$.

Step 10. If $m_j > 0$, sample ϕ_j from

$$\begin{aligned}\pi(\phi_j|\text{rest}) \propto \prod_{s_i=j} & \frac{1}{2(2\pi)^{P/2}|\phi_{j,2}|^{0.5}} \\ & \times \left[\exp\{-0.5(\boldsymbol{X} - \boldsymbol{\phi}_{j,1} - \Gamma Z)^T \phi_{j,2}^{-1}(\boldsymbol{X} - \boldsymbol{\phi}_{j,1} - \Gamma Z)\}\right] H_x(\phi_j),\end{aligned}$$

which is equivalent to drawing $\phi_{j,2}$ from

$$\mathrm{IW}\Bigg(\nu_x + m_j + 1, \qquad [\textstyle\sum_{i:s_i=j}(\boldsymbol{X}_i - \boldsymbol{\phi}_{j,1} - \Gamma Z_i)(\boldsymbol{X}_i - \boldsymbol{\phi}_{j,1} - \Gamma Z_i)^T + \\ \mathrm{Diag}(\boldsymbol{\tau}_x^{-1/2})(\phi_{j,1} - \boldsymbol{m}_x)(\boldsymbol{\phi}_{j,1} - \boldsymbol{m}_x)^T \mathrm{Diag}(\boldsymbol{\tau}_x^{-1/2}) + D_x^{-1}]^{-1}\Bigg)$$

and then drawing $\phi_{j,1}$ from a normal distribution with mean and variance

$$\begin{aligned} \boldsymbol{m} &= v\Bigg\{\mathrm{Diag}(\boldsymbol{\tau}_x^{-1/2})\phi_{j,2}^{-1}\mathrm{Diag}(\boldsymbol{\tau}_x^{-1/2})\boldsymbol{m}_x + \phi_{j,2}^{-1}\sum_{i:s_i=j}(\boldsymbol{X}_i - \Gamma Z_i)\Bigg\}, \\ v &= \{\mathrm{Diag}(\boldsymbol{\tau}_x^{-1/2})\phi_{j,2}^{-1}\mathrm{Diag}(\boldsymbol{\tau}_x^{-1/2}) + m_j\phi_{j,2}^{-1}\}^{-1}. \end{aligned}$$

If $m_j = 0$, sample ϕ_j from $\pi(\phi_j|\mathrm{rest}) \propto H_x(\phi_j)$ for $j = 1, \ldots, L$.

Step 11. Sample t_l from Multinomial$(p^*_{1u}, \ldots, p^*_{Mu})$, where

$$(p^*_{1u}, \ldots, p^*_{Mu}) \propto (p_{1u} f_u(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i|\varphi_j, \psi^{(r)_i}), \ldots, (p_{MU} f_u(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i|\varphi_M, \psi^{(r)_i}))$$

and

$$\begin{aligned} f_u(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i|\varphi_j, \psi_i^{(r)}) &= \frac{1}{(2\pi)^{P/2}|\varphi_{j,2}|^{1/2}} \\ &\quad \times \exp\Bigg\{-0.5(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i - \psi_i^{(r)}\boldsymbol{\varphi}_{j,1})^T \varphi_{j,2}^{-1}(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i - \psi_i^{(r)}\boldsymbol{\varphi}_{j,1})\Bigg\}. \end{aligned}$$

Step 12. Sample $(p_{1u}, \ldots, p_{Mu})$ from a Dirichlet$(\alpha_u/M + n_1, \ldots, \alpha_u/M + n_M)$.

Step 13. If $n_j > 0$, sample φ_j from

$$\begin{aligned} \pi(\varphi_j|\mathrm{rest}) &\propto \prod_{t_l=j} \frac{1}{(2\pi)^{P/2}|\varphi_{j,2}|^{1/2}} \\ &\quad \times \exp\Bigg\{-0.5(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i - \psi_i^{(r)}\boldsymbol{\varphi}_{j,1})^T \varphi_{j,2}^{-1}(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i - \psi_i^{(r)}\boldsymbol{\varphi}_{j,1})\Bigg\} H_u(\varphi_j), \end{aligned}$$

which is equivalent to drawing $\varphi_{j,2}$ from

$$\mathrm{IW}\Bigg(\nu_u + n_j + 1, \quad [\textstyle\sum_{i,r:t_i^{(r)}=j}(\boldsymbol{W}_i^{(r)} - \psi_i^{(r)}\boldsymbol{\varphi}_{j,1} - \boldsymbol{X}_i)(\boldsymbol{W}_i^{(r)} - \psi_i^{(r)}\boldsymbol{\varphi}_{j,1} - \boldsymbol{X}_i)^T + \mathrm{Diag}(\boldsymbol{\tau}_u^{-1/2})(\boldsymbol{\varphi}_{j,1} - \boldsymbol{m}_u)(\boldsymbol{\varphi}_{j,1} - \boldsymbol{m}_u)^T\mathrm{Diag}(\boldsymbol{\tau}_u^{-1/2}) + D_u^{-1}]^{-1}\Bigg)$$

and then drawing $\boldsymbol{\varphi}_{j,1}$ from a normal distribution with mean and variance

$$\begin{aligned}\boldsymbol{m} &= v\Bigg\{\mathrm{Diag}(\boldsymbol{\tau}_u^{-1/2})\varphi_{j,2}^{-1}\mathrm{Diag}(\boldsymbol{\tau}_u^{-1/2})\boldsymbol{m}_u + \varphi_{j,2}^{-1}\sum_{i,r:t_i^{(r)}=j}\psi_i^{(r)}(\boldsymbol{W}_i^{(r)} - \boldsymbol{X}_i)\Bigg\}, \\ v &= \{\mathrm{Diag}(\boldsymbol{\tau}_x^{-1/2})\varphi_{j,2}^{-1}\mathrm{Diag}(\boldsymbol{\tau}_u^{-1/2}) + n_j\varphi_{j,2}^{-1}\}^{-1}.\end{aligned}$$

If $n_j = 0$, sample φ_j from $\pi(\varphi_j|\mathrm{rest}) \propto H_u(\varphi_j)$ for $j = 1, \ldots, M$.

Step 14. For $p = 1, \ldots, P$, draw the p^{th} component of $\boldsymbol{\tau}_x$ from $\mathrm{IG}(g_x, h_x)$ and denote the value as τ_{px}^* and the new vector as $\boldsymbol{\tau}_x^*$, while the same notations without * are used for the current values. We accept the new vector with probability $\min\{1, f_{\boldsymbol{\tau}_x}(\boldsymbol{\tau}_x^*|\mathrm{rest})/f_{\boldsymbol{\tau}_x}(\boldsymbol{\tau}_x|\mathrm{rest})\}$, where

$$\begin{aligned}f_{\boldsymbol{\tau}_x}(\boldsymbol{\tau}_x|\mathrm{rest}) \propto & g_x^{-h_x}\tau_{px}^{-h_x-1}\frac{\exp\{-(g_x\tau_{px})^{-1}\}}{\Gamma(h_x)}\prod_{j=1}^{L}\frac{1}{(2\pi)^{P/2}|\boldsymbol{\tau}_x|^{0.5}} \\ & \times\Bigg[\exp\{-0.5(\phi_{j,1} - \boldsymbol{m}_x)^T\mathrm{Diag}(\boldsymbol{\tau}_x^{-1/2})\phi_{j,2}^{-1}\mathrm{Diag}(\boldsymbol{\tau}_x^{-1/2})(\phi_{j,1} - \boldsymbol{m}_x)\}\Bigg].\end{aligned}$$

Step 15. Draw D_x from the conditional distribution

$$\mathrm{IW}\Bigg(p + \nu_x\sum_{j=1}^{L}I(m^j > 0), [I_P + \sum_{j=1}^{L}I(m^j > 0)\phi_{j,2}^{-1}]^{-1}\Bigg).$$

Step 16. For $p = 1, \ldots, P$, draw the p^{th} component of $\boldsymbol{\tau}_u$ from $\mathrm{IG}(g_u, h_u)$ and denote the value as τ_{pu}^* and the new vector as $\boldsymbol{\tau}_u^*$, while the same notations without * are used for the current values. We accept the new vector with probability $\min\{1, f_{\boldsymbol{\tau}_u}(\boldsymbol{\tau}_u^*|\text{rest})/f_{\boldsymbol{\tau}_u}(\boldsymbol{\tau}_u|\text{rest})\}$, where

$$f_{\boldsymbol{\tau}_u}(\boldsymbol{\tau}_u|\text{rest}) \propto g_u^{-h_u}\tau_{pu}^{-h_u-1}\frac{\exp\{-(g_u\tau_{pu})^{-1}\}}{\Gamma(h_u)}\prod_{j=1}^{M}\frac{1}{(2\pi)^{P/2}|\boldsymbol{\tau}_u|^{0.5}}$$
$$\times\left[\exp\{-0.5(\varphi_{j,1}-\boldsymbol{m}_u)^T\text{Diag}(\boldsymbol{\tau}_u^{-1/2})\varphi_{j,2}^{-1}\text{Diag}(\boldsymbol{\tau}_u^{-1/2})(\varphi_{j,1}-\boldsymbol{m}_u)\}\right].$$

Step 17. Draw D_u from the conditional distribution

$$\text{IW}\left(p+\nu_u\sum_{j=1}^{M}I(n^j>0),[I_P+\sum_{j=1}^{M}I(n^j>0)\varphi_{j,2}^{-1}]^{-1}\right).$$

3.6 Simulation studies

3.6.1 Simulation design and method of analysis

Simulation studies were conducted by generating cohort data consisting of $(Y, Z, \boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)})$, where Z followed a Normal$(0,1)$ distribution. The erroneous surrogates were obtained by setting $\boldsymbol{W}^{(1)} = \boldsymbol{X}+\boldsymbol{U}^{(1)}$ and $\boldsymbol{W}^{(2)} = \boldsymbol{X}+\boldsymbol{U}^{(2)}$.

Here, we considered 7 different scenarios. For scenario 1, $\boldsymbol{X}|Z \sim \text{Normal}_2\{(0.1\ \ -0.1)^TZ, I_2)$, and $U_1, U_2 \sim \text{Normal}(0, \sqrt{0.5}^2)$, for scenario 2, $\boldsymbol{X}|Z \sim \text{Normal}_2\{(0.1\ \ -0.1)^TZ, I_2)$, and $U_1, U_2 \sim \text{Uniform}(-\sqrt{1.5}, \sqrt{1.5})$, for scenario 3, $\boldsymbol{X}|Z \sim (0.1\ \ -0.1)^TZ + [\text{Unif}(-\sqrt{3},\sqrt{3})\ \text{Unif}(-\sqrt{3},\sqrt{3})]$, and $U_1, U_2 \sim \text{Normal}(0, \sqrt{0.5}^2)$, for scenario 4, $\boldsymbol{X}|Z \sim (0.1\ \ -0.1)^TZ + [\text{Unif}(-\sqrt{3},\sqrt{3})\ \text{Unif}(-\sqrt{3},\sqrt{3})]$, and $U_1, U_2 \sim \text{Uniform}(-\sqrt{1.5}, \sqrt{1.5})$, for scenario 5, $\boldsymbol{X} \sim \text{Normal}_2\{(0.1\ \ -0.1)^TZ, \Sigma)$, and where

$$\Sigma = \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix},$$

and $\boldsymbol{U} \sim 0.5\text{Normal}_2[(0.5\ 0.5), 0.25\Sigma] + 0.5\text{Normal}_2[(-0.5\ \ -0.5), 0.25\Sigma]$, and scenarios 6 and 7 were the same as scenarios 1 and 2 with some differences that will be pointed later. Here Normal_2 and I_2 denote bivariate normal distribution and identity matrix of order 2, and $\text{Unif}(a, b)$ stands for uniform distribution between a and b. For all scenarios, the variance of the measurement error was 50% of the variance of $\boldsymbol{X}$ given Z.

Finally, Y was simulated from a Bernoulli distribution with success probability satisfying $\text{logit}\{\text{pr}(Y = 1|\boldsymbol{X}, Z) = -3 + X_1 + X_2 + 0.5X_1X_2 + Z$, resulting in about 10% cases on average. We considered the cohort size of $n = 5,000$.

Each simulated dataset was analyzed using the naive method (NV) method, where unobserved $\boldsymbol{X} = (X_1, X_2)^T$ was replaced by $\overline{\boldsymbol{W}} = (\overline{W}_1, \overline{W}_2) = (1/2)(\boldsymbol{W}^{(1)} + \boldsymbol{W}^{(2)})$, the regression calibration method [?] where the unobserved $\boldsymbol{X}$ is replaced by

$$\widehat{\boldsymbol{X}} = E(\boldsymbol{X}|\overline{\boldsymbol{W}}, \boldsymbol{Z}) = \widehat{\Gamma}\boldsymbol{Z} + \widehat{\Sigma_x}(\widehat{\Sigma_u}/2 + \widehat{\Sigma_x})^{-1}(\overline{\boldsymbol{W}} - \widehat{\Gamma}\boldsymbol{Z}), \tag{3.2}$$

where the quantities with $\widehat{}$ are empirically estimated, the simulation extrapolation (SIMEX) method [?], and the semiparametric Bayesian (SPB) method. In the first set of simulations (scenarios 1-5), we assumed both $\boldsymbol{W}^{(1)}$ and $\boldsymbol{W}^{(2)}$ are observed for all subjects. In the second set of simulations (scenarios 6 and 7), we evaluated performance of the methods with missing data, where $\boldsymbol{W}^{(2)}$ had 90% missing values. The missing values were created by the missing completely at random (MCAR) mechanism. For the naive method, we replaced $\boldsymbol{X}_i$ by $\overline{\boldsymbol{W}}^* = \sum_{r=1}^{2}(\boldsymbol{W}^{(r)}\Delta^{(r)})/\sum_{r=1}^{2}\Delta^{(r)}$. For the RC, $\boldsymbol{X}_i$ was replaced by $\widehat{\boldsymbol{X}}_i$ given in (3.2), where $\widehat{\Sigma}_u$ was obtained based on the subjects with both $\boldsymbol{W}^{(1)}$ and $\boldsymbol{W}^{(2)}$ observed, $\overline{\boldsymbol{W}}^*$ was used in place of $\overline{\boldsymbol{W}}$, and everything else was based on only $\boldsymbol{W}^{(1)}$. For the SIMEX, $\widehat{\Sigma}_u$ was obtained based on subjects where both $\boldsymbol{W}^{(1)}$ and $\boldsymbol{W}^{(2)}$ were observed and everything else was based on only $\boldsymbol{W}^{(1)}$.

For SPB, we used 10,000 MCMC iterations, and the posterior inference is based on the last 5,000 MCMC samples, and the mean and median of the MCMC chain were used as the point estimates. We used $a_\gamma = b_\gamma = 1,\sigma^2_{\beta_0} = \sigma^2_{\beta_1} = \sigma^2_{\beta_2} = \sigma^2_{\beta_3} = 5$, $\mu_\gamma = 0$, $\sigma^2_{\gamma 3} = 1$ $g_u = h_u = g_x = h_x = 1$, $\boldsymbol{m}_u = 0$, $\boldsymbol{m}_x = 0$, and $\mathrm{IW}(P, I_P)$ as prior for D_x and D_u. For all scenarios, we present mean bias (MB), median bias (MEDB), median absolute deviation (MAD), empirical standard error (SE), and root mean squared error (RMSE) based on 1,000 replications.

3.6.2 Results of the simulation study

Tables 3.1, 3.2, 3.3, and 3.4 contain the results for scenarios 1 and 2, scenarios 3 and 4, scenario 5, and scenarios 6 and 7, respectively.

In all scenarios, the SPB results based on the posterior mean and median were fairly close, and occasionally the median based results were slightly better than the mean based results. Overall, the naive method was obviously very biased. RC and SIMEX had smaller biases than the naive method. However, their biases were still highly significant in terms of the bias score $\sqrt{1,000}(\text{mean estimate} - \text{truth})/\text{SE}$. The proposed method had the smallest biases and was competitive in terms of RMSE. Now, we first examine the $\boldsymbol{X}$ related parameters.

When both replicates were observed for all observations, which was the case without missing data (Tables 3.1, 3.2, and 3.3), SPB was the clear winner with the smallest bias for each of the parameters than other methods. SPB is not expected to achieve the smallest standard errors, since we have to sacrifice some efficiency to obtain more robustness. However, comparisons based on RMSE, which incorporates both the variance of an estimator and its bias, might still give some insights. Based on RMSE, SPB seemed to be close to or better than SIMEX in the uniform $\boldsymbol{X}$ case (scenarios 3 and 4); in the bivariate normal $\boldsymbol{X}$ case (scenarios 1, 2, 5, 6, and 7), SPB had the smallest RMSE for the interaction effect. With partially missing data (Table 3.4), SPB was again the winner with the smallest biases for almost all parameters.

RC seemed competitive in estimating the $\boldsymbol{X}$ related main effect parameters, but it had large biases for the intercept and β_Z, and those biases were as bad as those of the naive method. Summarizing all results, the proposed method had the smallest biases for all parameters and often the smallest RMSEs for the intercept and β_Z.

Table 3.1: Results of the simulation study based on 1,000 replications with sample size $n = 5,000$, where $\boldsymbol{X}|Z$ followed a bivariate normal distribution with uncorrelated components. Here MB, MEDB, MAD, SE, and RMSE denote the mean bias, median bias, median absolute deviation, empirical standard error, and root mean squared error, respectively. NV, RC, SIMEX, SPB-mean, and SPB-median denote the naive method, regression calibration, simulation extrapolation, the semiparametric Bayesian method based on the posterior mean and median, respectively.

		$U_1, U_2 \sim$ Normal					$U_1, U_2 \sim$ Uniform				
Method		β_{int}	β_Z	β_{X_1}	β_{X_2}	$\beta_{X_1X_2}$	β_{int}	β_Z	β_{X_1}	β_{X_2}	$\beta_{X_1X_2}$
NV	MB×10	1.99	−0.87	−2.23	−2.15	−2.45	2.00	−0.88	−2.24	−2.12	−2.49
	MEDB×10	2.00	−0.89	−2.26	−2.16	−2.45	2.05	−0.90	−2.25	−2.15	−2.50
	MAD×10	0.70	0.56	0.54	0.60	0.50	0.75	0.56	0.53	0.58	0.50
	SE×10	0.73	0.56	0.54	0.58	0.51	0.73	0.56	0.55	0.58	0.50
	RMSE×10	2.12	1.03	2.30	2.23	2.51	2.13	1.04	2.31	2.20	2.54
RC	MB×10	2.01	−0.86	−0.24	−0.21	−1.05	2.02	−0.87	−0.25	−0.17	−1.10
	MEDB×10	2.02	−0.89	−0.26	−0.24	−1.05	2.05	−0.89	−0.28	−0.18	−1.13
	MAD×10	0.69	0.56	0.67	0.75	0.78	0.76	0.57	0.67	0.74	0.80
	SE×10	0.73	0.56	0.68	0.73	0.80	0.73	0.57	0.69	0.73	0.78
	RMSE×10	2.14	1.03	0.72	0.76	1.31	2.14	1.04	0.73	0.75	1.35
SIMEX	MB×10	0.43	−0.21	−0.39	−0.35	−0.93	0.44	−0.23	−0.41	−0.31	−0.99
	MEDB×10	0.44	−0.23	−0.42	−0.37	−0.93	0.50	−0.26	−0.43	−0.35	−1.00
	MAD×10	0.85	0.63	0.73	0.84	0.81	0.89	0.63	0.71	0.81	0.84
	SE×10	0.87	0.62	0.74	0.80	0.82	0.87	0.63	0.74	0.79	0.81
	RMSE×10	0.97	0.66	0.84	0.87	1.24	0.97	0.67	0.85	0.85	1.28
SPB-mean	MB×10	−0.18	0.08	0.09	0.07	0.15	−0.13	0.06	0.04	0.09	0.09
	MEDB×10	−0.17	0.07	0.04	0.05	0.14	−0.08	0.02	0.02	0.05	0.04
	MAD×10	0.91	0.66	0.81	0.90	0.98	0.92	0.66	0.80	0.88	0.96
	SE×10	0.93	0.65	0.80	0.88	1.00	0.94	0.66	0.81	0.86	0.98
	RMSE×10	0.94	0.66	0.81	0.88	1.01	0.95	0.67	0.81	0.86	0.99
SPB-median	MB×10	−0.16	0.08	0.08	0.06	0.14	−0.11	0.06	0.03	0.08	0.08
	MEDB×10	−0.14	0.06	0.02	0.03	0.13	−0.07	0.01	0.01	0.05	0.03
	MAD×10	0.91	0.66	0.81	0.90	0.99	0.92	0.66	0.80	0.88	0.96
	SE×10	0.93	0.65	0.80	0.88	1.00	0.94	0.66	0.81	0.86	0.99
	RMSE×10	0.94	0.66	0.81	0.88	1.01	0.95	0.66	0.81	0.86	0.99

Table 3.2: Results of the simulation study based on 1,000 replications with sample size $n = 5,000$, where conditional distributions of X_1 and X_2 given Z were uniform. Here MB, MEDB, MAD, SE, and RMSE denote the mean bias, median bias, median absolute deviation, empirical standard error, and root mean squared error, respectively. NV, RC, SIMEX, SPB-mean, and SPB-median denote the naive method, regression calibration, simulation extrapolation, the semiparametric Bayesian method based on the posterior mean and median, respectively.

		$U_1, U_2 \sim$ Normal					$U_1, U_2 \sim$ Uniform				
Method		β_{int}	β_Z	β_{X_1}	β_{X_2}	$\beta_{X_1X_2}$	β_{int}	β_Z	β_{X_1}	β_{X_2}	$\beta_{X_1X_2}$
NV	MB×10	2.22	−0.91	−2.21	−2.12	−2.18	2.23	−0.92	−2.22	−2.13	−2.19
	MEDB×10	2.26	−0.91	−2.23	−2.13	−2.18	2.26	−0.90	−2.24	−2.14	−2.20
	MAD×10	0.75	0.56	0.55	0.54	0.48	0.71	0.55	0.55	0.57	0.49
	SE×10	0.75	0.56	0.54	0.55	0.47	0.74	0.56	0.54	0.55	0.49
	RMSE×10	2.35	1.07	2.28	2.19	2.23	2.35	1.08	2.28	2.20	2.24
RC	MB×10	2.24	−0.91	−0.20	−0.19	−0.62	2.25	−0.91	−0.21	−0.20	−0.63
	MEDB×10	2.29	−0.90	−0.21	−0.19	−0.62	2.28	−0.90	−0.23	−0.22	−0.64
	MAD×10	0.74	0.57	0.68	0.70	0.73	0.72	0.55	0.68	0.71	0.75
	SE×10	0.75	0.57	0.68	0.70	0.74	0.74	0.57	0.67	0.69	0.77
	RMSE×10	2.36	1.07	0.71	0.73	0.96	2.37	1.07	0.71	0.72	0.99
SIMEX	MB×10	0.51	−0.24	−0.37	−0.32	−0.67	0.52	−0.25	−0.38	−0.33	−0.70
	MEDB×10	0.57	−0.25	−0.39	−0.33	−0.66	0.56	−0.24	−0.40	−0.34	−0.72
	MAD×10	0.87	0.63	0.76	0.77	0.79	0.84	0.62	0.77	0.80	0.78
	SE×10	0.90	0.63	0.76	0.79	0.76	0.88	0.63	0.75	0.78	0.80
	RMSE×10	1.03	0.67	0.84	0.85	1.02	1.02	0.68	0.84	0.84	1.06
SPB-mean	MB×10	−0.22	0.03	0.14	0.07	0.16	−0.16	0.01	0.08	0.07	0.08
	MEDB×10	−0.16	0.04	0.13	0.03	0.13	−0.13	0.01	0.04	0.05	0.09
	MAD×10	0.97	0.65	0.86	0.86	0.98	0.94	0.64	0.83	0.91	1.00
	SE×10	0.98	0.65	0.88	0.91	0.99	0.96	0.66	0.87	0.90	1.03
	RMSE×10	1.00	0.65	0.89	0.92	1.00	0.98	0.66	0.87	0.90	1.03
SPB-median	MB×10	−0.20	0.02	0.13	0.06	0.16	−0.14	0.00	0.07	0.05	0.08
	MEDB×10	−0.14	0.03	0.12	0.01	0.12	−0.12	0.00	0.03	0.03	0.09
	MAD×10	0.96	0.65	0.86	0.88	0.98	0.93	0.65	0.84	0.91	1.00
	SE×10	0.98	0.65	0.88	0.91	0.98	0.96	0.66	0.87	0.89	1.02
	RMSE×10	1.00	0.65	0.88	0.91	1.00	0.97	0.66	0.87	0.90	1.03

Table 3.3: Results of the simulation study based on 1,000 replications with sample size $n = 5,000$, where $\boldsymbol{X}$ given Z followed a bivariate normal distribution with correlated components and $\boldsymbol{U}$ followed a mixture of two bivariate normal distributions each with correlated components. Here MB, MEDB, MAD, SE, and RMSE denote the mean bias, median bias, median absolute deviation, empirical standard error, and root mean squared error, respectively. NV, RC, SIMEX, SPB-mean, and SPB-median denote the naive method, regression calibration, simulation extrapolation, the semiparametric Bayesian method based on the posterior mean and median, respectively.

Method		β_{int}	β_Z	β_{X_1}	β_{X_2}	$\beta_{X_1X_2}$
NV	MB×10	3.07	−1.29	−2.95	−2.79	−2.84
	MEDB×10	3.09	−1.30	−2.98	−2.79	−2.84
	MAD×10	0.68	0.53	0.55	0.53	0.49
	SE×10	0.67	0.54	0.52	0.54	0.47
	RMSE×10	3.14	1.40	2.99	2.84	2.88
RC	MB×10	3.37	−1.29	−0.48	−0.39	−1.50
	MEDB×10	3.38	−1.30	−0.52	−0.39	−1.49
	MAD×10	0.68	0.53	0.68	0.65	0.79
	SE×10	0.67	0.54	0.66	0.69	0.76
	RMSE×10	3.44	1.40	0.81	0.79	1.68
SIMEX	MB×10	0.88	−0.46	−0.84	−0.74	−1.28
	MEDB×10	0.91	−0.48	−0.90	−0.75	−1.27
	MAD×10	0.86	0.61	0.75	0.72	0.80
	SE×10	0.85	0.62	0.71	0.75	0.79
	RMSE×10	1.23	0.77	1.11	1.06	1.50
SPB-mean	MB×10	−0.24	0.04	0.31	0.03	0.05
	MEDB×10	−0.20	0.01	0.27	0.01	0.07
	MAD×10	1.01	0.65	0.85	0.81	0.98
	SE×10	1.00	0.68	0.84	0.86	1.01
	RMSE×10	1.03	0.68	0.90	0.86	1.01
	RMSE×10	1.03	0.68	0.90	0.86	1.01
SPB-median	MB×10	−0.22	0.03	0.30	0.01	0.05
	MEDB×10	−0.18	0.01	0.25	−0.00	0.06
	MAD×10	1.01	0.65	0.85	0.81	0.99
	SE×10	1.00	0.67	0.84	0.85	1.01
	RMSE×10	1.03	0.67	0.89	0.85	1.01
	RMSE×10	1.03	0.68	0.89	0.86	1.01

Table 3.4: Results of the simulation study based on 1,000 replications with sample size $n = 5,000$, where $\boldsymbol{X}|Z$ followed a bivariate normal distribution with uncorrelated components and the second replicate had a missingness probability of 90%. Here MB, MEDB, MAD, SE, and RMSE denote the mean bias, median bias, median absolute deviation, empirical standard error, and root mean squared error, respectively. NV, RC, SIMEX, SPB-mean, and SPB-median denote the naive method, regression calibration, simulation extrapolation, the semiparametric Bayesian method based on the posterior mean and median, respectively.

		$\boldsymbol{U} \sim$ Bivariate Normal					$\boldsymbol{U} \sim$ Bivariate Uniform				
Method		β_{int}	β_Z	β_{X_1}	β_{X_2}	$\beta_{X_1X_2}$	β_{int}	β_Z	β_{X_1}	β_{X_2}	$\beta_{X_1X_2}$
NV	MB×10	3.06	−1.30	−3.52	−3.42	−3.41	3.06	−1.32	−3.52	−3.39	−3.47
	MEDB×10	3.08	−1.31	−3.53	−3.45	−3.39	3.12	−1.33	−3.51	−3.41	−3.47
	MAD×10	0.67	0.54	0.46	0.52	0.39	0.65	0.53	0.49	0.51	0.39
	SE×10	0.67	0.53	0.48	0.51	0.41	0.70	0.55	0.49	0.52	0.39
	RMSE×10	3.14	1.40	3.56	3.46	3.43	3.14	1.43	3.56	3.43	3.49
RC	MB×10	3.17	−1.32	−0.39	−0.35	−1.60	3.18	−1.34	−0.38	−0.28	−1.73
	MEDB×10	3.18	−1.34	−0.40	−0.37	−1.60	3.21	−1.34	−0.41	−0.31	−1.77
	MAD×10	0.68	0.55	0.76	0.87	0.88	0.68	0.55	0.79	0.82	0.84
	SE×10	0.68	0.54	0.79	0.84	0.89	0.70	0.56	0.83	0.86	0.85
	RMSE×10	3.25	1.43	0.88	0.91	1.83	3.25	1.45	0.91	0.90	1.92
SIMEX	MB×10	1.27	−0.62	−1.28	−1.19	−2.17	1.25	−0.65	−1.26	−1.11	−2.30
	MEDB×10	1.30	−0.63	−1.29	−1.20	−2.18	1.31	−0.64	−1.29	−1.14	−2.32
	MAD×10	0.82	0.60	0.70	0.77	0.72	0.83	0.64	0.76	0.78	0.68
	SE×10	0.83	0.61	0.72	0.78	0.74	0.87	0.63	0.75	0.80	0.69
	RMSE×10	1.52	0.87	1.47	1.42	2.29	1.52	0.91	1.47	1.37	2.40
SPB-mean	MB×10	−0.39	0.18	0.22	0.30	0.37	−0.33	0.12	0.24	0.29	0.16
	MEDB×10	−0.33	0.16	0.21	0.26	0.32	−0.22	0.09	0.22	0.30	0.09
	MAD×10	1.08	0.71	1.03	1.17	1.32	1.07	0.73	1.06	1.11	1.19
	SE×10	1.09	0.71	1.06	1.15	1.34	1.10	0.73	1.08	1.15	1.22
	RMSE×10	1.16	0.73	1.08	1.19	1.39	1.15	0.74	1.11	1.19	1.23
SPB-median	MB×10	−0.35	0.17	0.18	0.26	0.33	−0.30	0.11	0.21	0.26	0.13
	MEDB×10	−0.31	0.15	0.19	0.23	0.29	−0.18	0.08	0.18	0.28	0.06
	MAD×10	1.08	0.70	1.03	1.16	1.31	1.04	0.73	1.06	1.12	1.21
	SE×10	1.09	0.71	1.05	1.15	1.34	1.09	0.73	1.08	1.14	1.23
	RMSE×10	1.14	0.73	1.07	1.18	1.38	1.13	0.73	1.10	1.17	1.23

3.7 Analysis of the NHANES Data

3.7.1 Background and method of analysis

In the National Health and Nutrition Examination Survey, participants were recruited as a representative sample of the noninstitutionalized US population. In this analysis, we focus on data from 2003 to 2010.

To avoid confounding, we looked at a homogeneous group of non-Hispanic White male who at least 20 years of age with an education level of above high school. High cholesterol level was considered as the binary response variable Y. The predictors of interest were calorie-adjusted total fat (gm per kcal) and calorie-adjusted dietary protein (gm per kcal), both were measured with errors. For numerical stability, we applied the following transformation: $100 \times \log(1 + \text{predictor})$ on the 24 hours recalls; the corresponding re-centered true values are denoted as pFat and pProtein. Also, we used age as the control variable Z.

The data set was analyzed by the four methods described in the simulation section, and the results are summarized in Table 3.5. For SPB, we used the same priors as used in the simulation study.

3.7.2 Results of analysis

Out of the 2025 subjects included in our analysis, 910 (45%) reported having ever been told by a doctor or other health professional that their blood cholesterol level was high. In the logistic models, the interaction parameter between pFat and pProtein as well as the main effect parameters came out to be statistically significant in all methods. In addition, the results of all methods indicate that higher risks were associated with older age. However, caution should be exercised in concluding causal relationships as the dataset was obtained from an observational study.

Table 3.5: Results of the analysis of the NHANES data (2003–2010). Here Est., L, and U stand for the estimate, and lower and upper bound of the 95% bootstrap percentile confidence intervals based on 2,000 bootstrap samples for the NV, the RC, and the SIMEX method . For the SPB, L and U denote the lower and upper bound of the 95% equal tail credible interval, and PMN and PMD stand for the posterior mean and median, respectively.

	NV			RC			SIMEX			SPB			
	Est.	L	U	Est.	L	U	Est.	L	U	PMN	PMD	L	U
Age	0.025	0.020	0.030	0.025	0.020	0.031	0.026	0.020	0.031	0.025	0.025	0.020	0.030
pFat	0.022	−0.095	0.134	−0.021	−0.298	0.218	0.018	−0.167	0.180	−0.002	0.000	−0.265	0.248
pProtein	0.135	0.039	0.238	0.321	0.036	0.731	0.228	0.061	0.388	0.269	0.263	0.025	0.539
pFat×pProtein	0.175	0.077	0.292	0.709	0.234	1.370	0.316	0.136	0.548	0.442	0.419	0.082	0.915

Since the interaction effect was significant, the main effects could not be interpreted on their own. Instead, the effect of increasing pFat from its first quartile to the third quartile should be examined with the value of pProtein fixed. With pProtein fixed at its first or second quartile, the effect of increasing pFat was not significant as all confidence/credible intervals of the odds ratios contained one (left panel of Figure 3.2) except for the confidence interval by RC with pProtein fixed at its first quartile. However, with pProtein fixed at its third quartile, the effect of increasing pFat was significant with all confidence/credible intervals above one. Similarly, the effect of increasing pProtein from its first quartile to the third quartile should be examined with the value of pFat fixed. The effect of increasing pProtein was not statistically significant judged by all methods when pFat was at its first quartile (right panel of Figure 3.2), but it was significant when pFat was at its second or third quartile. In summary, pFat's and pProtein's associations with the outcome seem to reinforce each other.

In addition to parameter estimation and statistical inference, the SPB method can provide other useful information. For example, we can obtain the posterior distribution of $(\boldsymbol{\mu}_x, \Sigma_x)$. To get that distribution we randomly draw $(\boldsymbol{\mu}_{x,n+1}, \Sigma_{x,n+1})$ from

$$\frac{\alpha_x\{1 - \sum_{j=1}^{L} I(m^j > 0)/L\}}{(\alpha_x + n)} H_x(\boldsymbol{\mu}_{x,n+1}, \Sigma_{x,n+1}) \\ + \sum_{j=1}^{L} \frac{(m^j + \alpha_x/L)}{(\alpha_x + n)} I(m^j > 0)\delta_{(\phi_{j,1},\phi_{j,2})}(\boldsymbol{\mu}_{x,n+1}, \Sigma_{x,n+1})$$

in every iteration of the MCMC chain. Similarly, we obtain the posterior distribution of $(\boldsymbol{\mu}_u, \Sigma_u)$ by simulating $(\boldsymbol{\mu}_{u,2n+1}, \Sigma_{u,2n+1})$ (as we condition on 2 replicated surrogate for each subject) from the following mixture distribution

$$\frac{\alpha_u\{1-\sum_{j=1}^{L} I(n^j>0)/M\}}{(\alpha_u+2n)} H_u(\boldsymbol{\mu}_{u,2n+1},\Sigma_{u,2n+1})$$
$$+\sum_{j=1}^{M}\frac{(n^j+\alpha_u/M)}{(\alpha_u+2n)} I(n^j>0)\delta_{(\varphi_{j,1},\varphi_{j,2})}(\boldsymbol{\mu}_{u,2n+1},\Sigma_{u,2n+1})$$

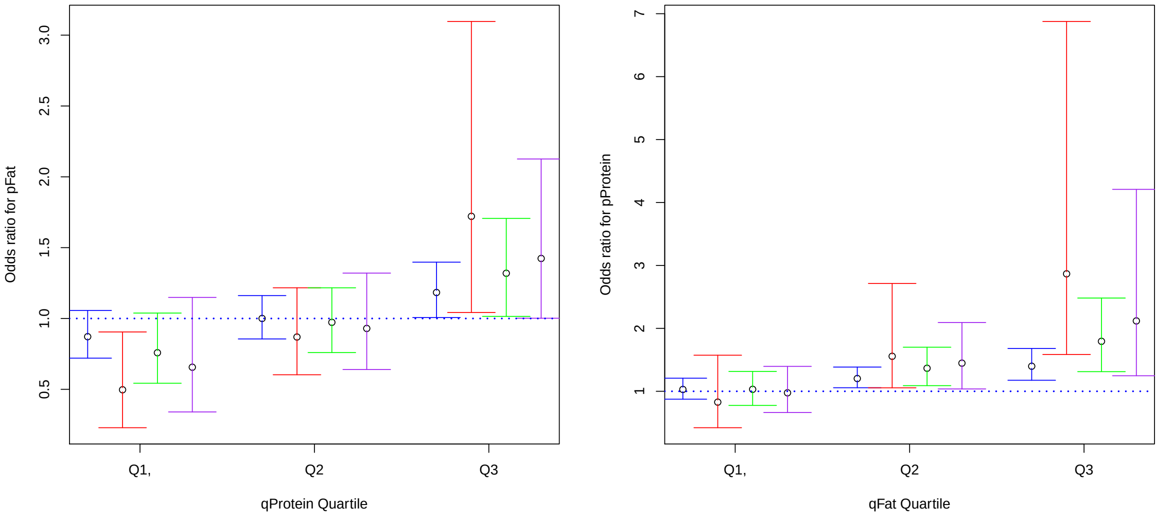
in every iteration of the MCMC chain. The posterior distributions of $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_u$ are given in the 3D plots of Figure 3.3 along with the marginals of each components. The plots clearly indicate that a single biviariate normal distribution would not adequately capture the features of the distributions, which suggests that flexible models like ours are needed to avoid misspecification and the related bias.

3.8 Discussion

In this section, we have proposed a semiparametric Bayesian method to handle measurement error of covariates in a logistic regression model. The novelty of the method is the estimation of the interaction effect along with the mail effects while the predictors are measured with error. In addition, the method is able to handle missing values in the variables measured with error in an easy and natural way.

The simulation studies suggest that the proposed method outperforms RC and the SIMEX in terms of bias in all the simulation settings. In particular, we see that the SIMEX approach has a large bias in estimating the interaction term. In estimating the intercept and β_Z, the proposed method consistently beat the RC. Although any method that properly takes into account measurement error is accompanied by relatively large uncertainties, the uncertainties (as measured via standard errors) of the proposed method are comparable with the SIMEX approach.

As a Bayesian method, the SPB can easily incorporate prior knowledge when such knowledge exists and thus lead to better estimates. Some of the byproducts can



(a)

(b)

Figure 3.2: Estimates and confidence/credible intervals of odds ratios for the naive (blue), the RC (red), the SIMEX (green), and the SPB method (purple). The plot on the left shows the odds ratios corresponding to increasing pFat from its first quartile to its third quartile with the value of pProtein fixed to the first, second, and third quartiles. The plot on the right shows the odds ratios corresponding to increasing pProtein from its first quartile to its third quartile with the value of pFat fixed to the first, second, and third quartiles.

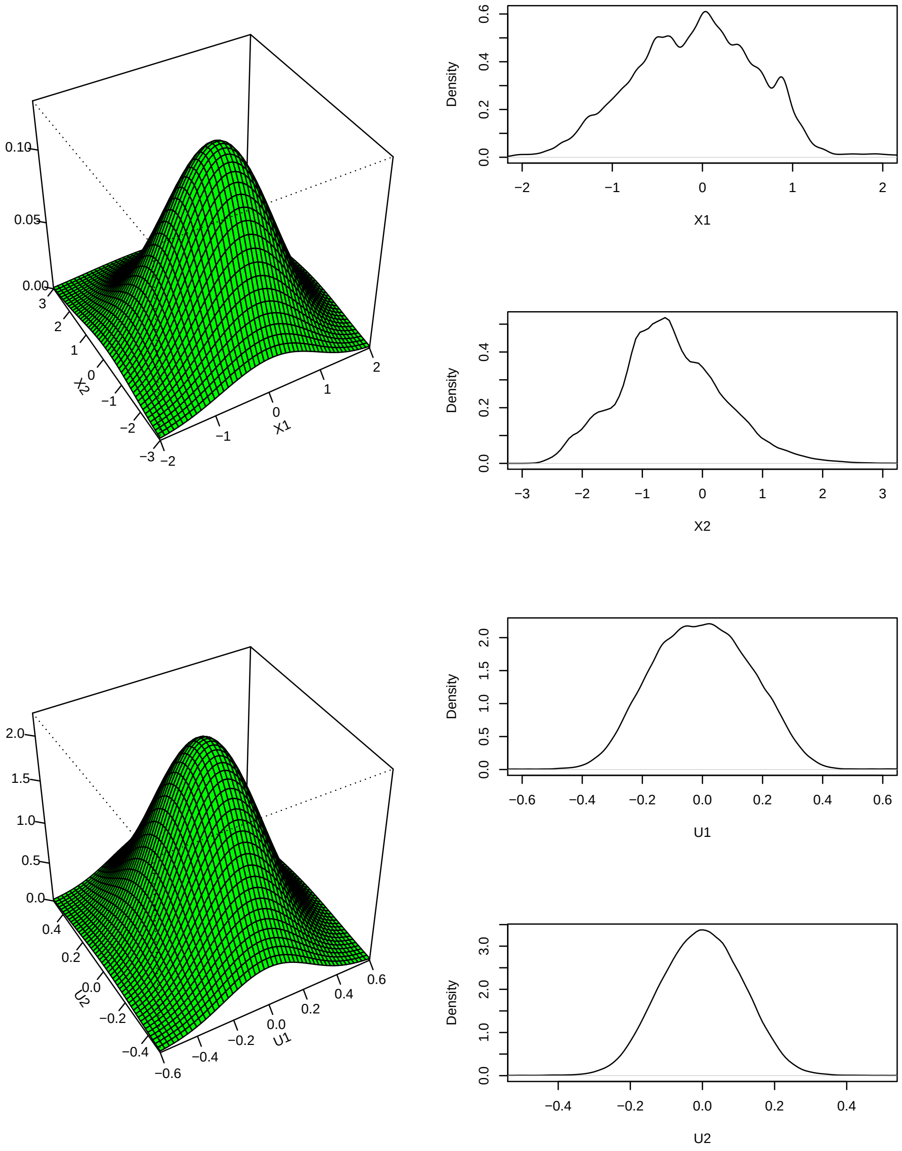


Figure 3.3: Posterior distributions and the corresponding marginals for $\boldsymbol{\mu}_x$ (top) and $\boldsymbol{\mu}_u$ (bottom).

be useful. For example, we can obtain information on the distributions on the true unobserved variables, the latent clusters, and get a clear idea about the measurement error distribution.

The proposed approach tries to avoid model misspecification by using the Dirichlet process mixture on the distribution of the unobserved covariates and the measurement error, which is flexible enough to model otherwise difficult distributions. At the same time, to lighten the computational burden we use a finite dimensional approximation of the full Dirichlet process.

One of the most important components of the current paper is the analysis of the NHANES data. To the best of our knowledge, this represents the first attempt to simultaneously address the main effects of fat and protein, their interaction effect, and the measurement error problem. The interaction was significant by all four methods, and we recommend that the interaction between fat and protein be considered in future research.

4. CONCLUSIONS

4.1 Summary

In Section 2, we have used the two-stage model to incorporating breast cancer trait information into etiologic investigations. In addition to reducing the dimension of the polytomous logistic regression, the model provides a convenient evaluation of the heterogeneity of the odds ratios.

For parameter estimation of the second-stage model, we have proposed a pseudo conditional likelihood method, which artificially matches each case with all the controls. As a result of the matching process, our method reduces the effects of the intercepts on the estimation of the regression parameters, which explains its robustness against the misspecification of the intercept model.

We've established the method's unbiasedness and asymptotic properties, which have been demonstrated by simulation studies. Compared with alternative methods of parameter estimation (i.e., the MLE and the CMLE), our method usually has smaller biases when the second-stage intercept model is misspecified. In terms of efficiency, our method is superior to the CMLE and very close to the MLE method.

Although the motivating example is a breast cancer classification problem, our methodology may have applications in other areas. For example, the method may be used to identify factors associated with whether a student can finish college education within four years, and the possible traits of interest may include honor status (no honor, Cum Laude, Magna Cum Laude, or Summa Cum Laude), job seeking outcome within 3 months of graduation (no job offer, 1-2 job offers, 3 or more job offers), etc.

In Section 3, we have addressed a measurement error problem, which is common in nutritional epidemiological studies. A few features make the problem challenging, including the missing data problem, the non-linearity of the model, and the potential interaction between nutrients measured with errors.

Compared with functional methods, structural methods are flexible enough to handle various problems but at the same time susceptible to mode misspecification. To guard against misspecification, we've proposed a semiparametric Bayesian method that uses flexible models for the unobserved true nutrient intakes as well as the measurement errors. Flexibility is achieved by using the finite dimensional Dirichlet process mixture distributions, which can well approximate any continuous distributions in principle.

We have compared the finite sample performance of the proposed method with the popular RC and SIMEX methods through simulation studies. The results indicate that our method has the smallest biases in parameter estimation. We have also applied the method to the NHANES data to assess the effects of total fat and protein on high cholesterol. Our results point to two possible reasons why some studies failed to find associations between nutrient intakes and certain diseases: they may have (1) ingnored the measurement errors or (2) neglected the interactions between nutrients.

4.2 Future Research

In Section 2, we assumed that the second- or higher-order contrasts in the second stage model were zero, which could be formally tested as hypotheses. But there could be more convenient approaches. One possibility is to use model selection methods. Instead of making assumptions, we can use lasso (Tibshirani 1997) or other types of model selection procedures to decide which of the higher-order contrasts should be included in the model. The advantage of modeling selection is that the whole process can be automated, and the price to pay is that the computational burden might be heavier.

In Section 3, we assumed that the observed surrogates were missing completely at random, which is a fairly strong assumption. It would be interesting to see whether this assumption can be relaxed to allow for some dependence between missingness and the covariates and/or the binary response variable.

REFERENCES

Appel LJ, et al. (2005). Effects of protein, monounsaturated fat, and carbohydrate intake on blood pressure and serum lipids: results of the OmniHeart randomized trial. *Jama*, **294**(19): 2455–2464.

Bolfarine H and Lachos VH. (2007). Skew-probit measurement error models. *Statistical Methodology*, **4**(1): 1–12.

Calle EE, et al. (2002). The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer*, **94**(9): 500–511.

Carroll RJ. (1998). Measurement error in epidemiologic studies. In: *Armitage P, Colton T, eds. Encyclopedia of Biostatistics.* Vol 3. Chichester, England: John Wiley & Sons, Ltd.

Carroll RJ, Ruppert D, Stefanski LA, and Crainiceanu CM. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective.* London, UK: Chapman and Hall/CRC.

Chatterjee N. (2004). A two-stage regression model for epidemiological studies with multivariate disease classification data. *Journal of the American Statistical Association*, **99**(465): 127–138.

Chatterjee N, Sinha S, Diver R, and Feigelson HS. (2010). Analysis of cohort studies with multivariate, partially observed, disease classification data. *Biometrika*, **97**(3): 683–698.

Dodge, Y (Ed). (2006). *The Oxford Dictionary of Statistical Terms.* Oxford, UK: Oxford University Press.

Engel J. (1998). Polytomous logistic regression. *Statistica Neerlandica*, **42**(4): 233–252.

Fagerland MW, Hosmer DW, and Bofin AM. (2008). Multinomial goodness of fit tests for logistic regression models. *Statistics in Medicine*, **27**(21): 4238–4253.

Fuller WA. (1987). *Measurement Error Models.* New York, NY: Wiley.

Garfinkel L. (1985). Selection, follow-up, and analysis in the American Cancer Society prospective studies: In: *Selection, Follow-up, and Analysis in Prospective Studies: A Workshop. National Cancer Institute Monograph 67*; Bethesda, MD.: National Cancer Institute: 49–52.

Goeman JJ and le Cessie S. (2006). A Goodness-of-Fit Test for Multinomial Logistic Regression. *Biometrics*, **62**(4): 980–985.

Goetghebeur E and Ryan L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika*, **82**(4): 821–833.

Goldman MB and Hatch MC. (2000). Breast cancer epidemiology, treatment, and prevention. In:*Ursin G, Spicer D (Eds) Women and Health*; London, UK: Academic Press: 871–883.

Gustafson P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments.* London, UK: Chapman and Hall, CRC.

Hosmer DW and Lemeshow S. (2000). *Applied Logistic Regression, 2nd edition.* New York, NY: Wiley.

Hossain S and Gustafson P. (2009). Bayesian adjustment for covariate measurement errors: a flexible parametric approach. *Statistics in medicine*, **28**(11): 1580–1600.

Huang X, Stefanski LA, and Davidian M. (2006). Latent-model robustness in structural measurement error models. *Biometrika*, **93**(1): 53–64.

Ishwaran H and James LF. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**(453): 161–173.

Ishwaran H and Zarepour M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, **30**(2): 269–283.

Johnson BA, Herring AH, Ibrahim JG, and Siega-Riz AM. (2007). Structured measurement error in nutritional epidemiology: Applications in the pregnancy, infection, and nutrition (PIN) study. *Journal of the American Statistical Association*, **102**(479): 856–866.

Little RJ. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, **83**(404): 1198–1202.

Mensink RP and Katan MB. (1989). Effect of a diet enriched with monounsaturated or polyunsaturated fatty acids on levels of low-density and high-density lipoprotein cholesterol in healthy women and men. *New England Journal of Medicine*, **321**(7): 436–441.

McCullough LE, et al. (2012). Fat or fit: The joint effects of physical activity, weight gain, and body size on breast cancer risk. *Cancer*, **18**(19): 4860–4868

Morton LM, et al. (2008). Etiologic heterogeneity among non-Hodgkin lymphoma subtypes. *Blood*, **112**(13): 5150–5160.

Nakamura T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, **77**(1): 127–137.

National Center for Health Statistics. National Health and Nutrition Examination Survey Overview.http://www.cdc.gov/nchs/nhanes.htm. Access on January 10, 2013.

Orgéas CC, et al. (2008). The influence of menstrual risk factors on tumor characteristics and survival in postmenopausal breast cancer. *Breast Cancer Res*, **10**(6): R107.

Richardson S, Leblond L, Jaussent I, and Green PJ. Mixture models in measurement error problems with reference to epidemiological studies. *Journal of the Royal Statistical Society, Series A*, **165**(3): 549–566.

Roeder K, Carroll RJ, and Lindsay BG. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, **91**(434): 722-732.

Rubin DB. (1976). Inference and missing data. *Biometrika*, **63**(3): 581–592.

Schwartz, JE. (1985). The neglected problem of measurement error in categorical data. *Sociological Methods & Research*, **13**(4): 435–466.

Sinha S, Mallick BK, Kipnis V, and Carroll RJ. (2010). Semiparametric bayesian analysis of nutritional epidemiology Data in the presence of measurement error. *Biometrics*, **66**(2): 444–454.

Spiegelman D, McDermott A, and Rosner B. (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *The American Journal of Clinical Nutrition*, **65**(4): 1179S–1186S.

Stefanski LA and Carroll RJ. (1987). Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika*, **74**(4): 703–716.

Stefanski LA and Cook JR. (1995). Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, **90**(432): 1247–1256.

Tibshirani R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**(4): 385–395.

van der Vaart AW. (1998). *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

Zaman K, Bodmer A, Pralong F, Castiglione-Gertsch M. (2010). Breast cancer and obesity, a dangerous relation. *Rev Med Suisse*, **8**(342): 1101–1104.