Diagnostic Measures for Missing Covariate Data and Semiparametric Models for Neuroimaging

Xiaoyan Shi

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2008

Approved by:

Joseph G. Ibrahim, Advisor

Hongtu Zhu, Advisor

Haitao Chu, Reader

Martin Styner, Reader

Donglin Zeng, Reader

Abstract

XIAOYAN SHI: Diagnostic Measures for Missing Covariate Data and Semiparametric Models for Neuroimaging.
(Under the direction of Joseph G. Ibrahim and Hongtu Zhu.)

This dissertation is composed of two major topics: a) diagnostic measures for generalized linear models (GLMs) with missing covariate data, and b) semiparametric models for neuroimaging data.

The first topic, diagnostic measures for GLMs with missing covariate data, is covered in two thesis papers. In the first paper, we carry out an in-depth investigation for assessing the influence of observations and model misspecification in the presence of missing covariate data in GLMs. Our diagnostic measures include case-deletion measures and conditional residuals. We use the conditional residuals to construct goodness of fit statistics for testing possible misspecifications in model assumptions. We develop specific strategies for incorporating missing data into goodness of fit statistics in order to increase the power of detecting model misspecification, and employ a resampling method to approximate the $p-$value of the goodness of fit statistics. In the second paper, we formally set up a general local influence method to carry out sensitivity analyses of minor perturbations to GLMs with missing covariate data. We examine two types of perturbation schemes (the single-case and global perturbation schemes) and show that the metric tensor of a perturbation manifold provides useful information for selecting an appropriate perturbation. We also develop several local influence measures to identify influential points and test model misspecification.

The second topic, semiparametric models for neuroimaging data, also consists of two thesis papers. The main objective of the first paper is to develop an adjusted exponentially tilted empirical likelihood (ETEL) procedure for the analysis of neuroimaging data. We propose a likelihood ratio statistic to test hypotheses and construct goodness of fit statistics for testing possible model misspecifications and apply them to the classification of time-dependent covariates. Our semiparametric method avoids standard parametric assumptions and the adjustment to the ETEL method can dramatically improve its finite sample performance over the original ETEL. In the second paper, we develop a semiparametric framework for describing the variability of medial representation (m-rep) of subcortical subjects and its association with covariates in a Euclidean space. Because the elements of the m-rep do not form a vector space, applying classical multivariate regression techniques may be inadequate in establishing the association between an m-rep and covariates of interest. Our semiparametric model avoids specifying a probability distribution on a Riemannian manifold. We develop an estimation procedure based on the annealing evolutionary stochastic approximation Monte Carlo (AESAMC) algorithm to obtain parameter estimates and establish their limiting distributions. We use Wald statistics to carry out tests of hypotheses.

Acknowledgments

I would especially like to thank Dr. Joseph Ibrahim for his mentorship, encouragement, inspiration, and support during the preparation of this dissertation. Also, I would like to convey my thanks to Dr. Hongtu Zhu for his help, lectures, and abundant patience. I would also like to express my appreciation for the help and comments from committee members Dr. Donglin Zeng and Dr. Haitao Chu. Finally, many warm thanks go to Dr. Martin Styner for his important contributions and interesting discussions on my research.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AESAMC	Annealing Evolutionary Stochastic Approximation Monte Carlo
EL	Empirical Likelihood
ETEL	Exponentially Tilted Empirical Likelihood
EM	Expectation-maximization
EPEE	Empirical Processes of the Estimating Equations
FDR	False Discovery Rate
GEE	Generalized estimating equations
GLM	Generalized Linear Model
GMM	Generalized Method of Moments
LRM	Linear Regression Model
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimates
MRI	Magnetic Resonance Imaging
M-rep	Medial Representation
NMAR	Not Missing at Random
SD	Standard Deviation
MAD	Median Absolute Deviation

Chapter 1

Introduction and Literature Review

This dissertation is composed of two major topics: a) diagnostic measures for generalized linear models (GLMs) with missing covariate data, and b) semiparametric models for neuroimaging.

The first topic, diagnostic measures for GLMs with missing covariate data, is covered in two thesis papers. In the first paper, we carry out an in-depth investigation for assessing the influence of observations and model misspecification in the presence of missing covariate data in GLMs. In the second paper, we formally set up a general local influence method to carry out sensitivity analyses of minor perturbations to GLMs with missing covariate data. We also develop several local influence measures to identify influential points and test model misspecification.

The second topic, semiparametric models for neuroimaging, also consists of two thesis papers. The main objective of the first paper is to develop an adjusted exponentially tilted empirical likelihood (ETEL) procedure for the analysis of neuroimaging data. We propose a likelihood ratio statistic to test hypotheses and construct goodness of fit statistics for testing possible model misspecifications and apply them to the classification of time-dependent covariates. In the second paper, we develop a semiparametric framework for describing the variability of the medial representation (m-rep) of subcortical structures and its association with covariates in a Euclidean space. Our semiparametric model avoids specifying a probability distribution on a Riemannian manifold.

The dissertation is organized as follows. The next section presents a literature review for each of the two topics. The first covers a review on diagnostic measures for missing data, and the second reviews existing statistical methods for neuroimaging studies. Then we proceed to present each of the four papers: We assess the influence of observations and model misspecification in the presence of missing covariate data for GLMs in Chapter 2, and we formally develop a general local influence method to carry out sensitivity analyses of minor perturbations to GLMs with missing covariate data in Chapter 3. The next two chapters (4 and 5) are dedicated to semiparametric models for neuroimaging data. Chapter 4 develops an adjusted exponentially tilted empirical likelihood (ETEL) procedure for the analysis of neuroimaging data, and Chapter 5 examines a semiparametric framework for the medial representation (m-rep) of subcortical structures.

1.1 Diagnostic measures for GLMs with missing covariate data

Missing data may arise due to many circumstances, such as the loss of hospital records, survey nonresponse, study subjects failing to report for monthly evaluations, and etc. There are three classifications of missing data. Data are said to be MCAR (Missing Completely at Random) if the failure to observe a value does not depend on any data, either observed or missing, whereas data are said to be MAR (Missing at Random) if, conditional on the observed data, the failure to observe a value does not depend on the data that are unobserved. MAR also includes MCAR. The missing data mechanism is NMAR (Not Missing at Random) if the failure to observe a value depends on the value that would have been observed (Ibrahim et al., 2005). One simple method, known as a complete case analysis, is the technique most commonly used with missing data and still the default method in most software packages. The complete case analysis can be biased when the data are not MCAR. Further, even when the data are MCAR so that the complete case analysis is unbiased, as the percentage of missing data increases, deleting of all subjects with missing data is wasteful and very inefficient. Therefore, it is necessary to seek ways to incorporate incomplete cases into the analysis.

Methods for handling missing data strongly depend on the mechanism that generated the missing values and distributional and model assumptions at various stages. Therefore, the resulting estimates and tests may be sensitive to these assumptions. For this reason, sensitivity analyses are commonly done to check the sensitivity of the parameters of interest with respect to the model assumptions.

Diagnostic measures such as residuals and Cook's distance have been widely used to identify influential observations in various regression models, such as generalized linear models (Cox and Snell, 1968; Cook and Weisberg, 1982; Davison and Tsai, 1992; Zhu et al., 2001). In addition, diagnostic measures, such as residuals, can be used to construct goodness of fit statistics to detect any systematic discrepancies between the data and the fitted values obtained from the model (Stute, 1997; Lin et al., 2002). However, to the best of our knowledge, virtually no literature exists for developing diagnostic measures such as residuals, Cook's distance, and goodness of fit statistics in generalized linear models (GLMs) with missing covariate data.

Sensitivity analyses are often carried out in two consecutive steps: selection of perturbation schemes to various model assumptions and use of influence measures to quantify the effects of those perturbations. Some literature on sensitivity analysis for missing data problems includes Copas and Li (1997), Copas and Eguchi (2005), Troxel (1998), Jansen et al. (2003), Van Steen, Molenberghs, and Thijs (2001), Verbeke et al. (2001), Hens et al. (2005), Jansen et al. (2006), and Troxel, Ma, and Heitjan (2004). For instance, Copas and Eguchi (2005) proposed a general formulation for assessing the bias of maximum likelihood estimates due to incomplete data in the presence of small model uncertainty. Verbeke et al. (2001), Hens et al. (2005), and Jansen et al. (2006) developed local influence methods for assessing nonrandom dropout in incomplete longitudinal data.

Cook (1986) proposed a general approach for assessing the local influence of a minor perturbation to a statistical model, which has been applied to many types of models, such as mixed models (Beckman, Nachtsheim, and Cook, 1987), generalized linear models (Thomas and Cook, 1989), among others. Zhu and Lee (2001) extended Cook's approach for assessing local influence in a minor perturbation of statistical models for latent variable models. Recently, Zhu et al. (2007a) developed a perturbation manifold to select an appropriate perturbation for statistical models without missing data, which is central to the development of the local influence approach proposed here.

1.2 Semiparametric models for neuroimaging data

Neuroimaging data, including both anatomical and functional magnetic resonance imaging (MRI), have been/are being widely collected to understand the neural development of neuropsychiatric disorders, substance use disorders, and the normal brain in various cross-sectional and longitudinal studies. For instance, various morphometrical measures of the morphology of the cortical and subcortical structures (e.g., hippocampus) are extracted from anatomical MRI for understanding neuroanatomical differences in brain structure across different populations. Nowadays, studies of brain morphology have been conducted widely to characterize differences in brain structure across groups of healthy individuals and persons with various diseases, and across time (Thompson and Toga, 2002; Thompson et al., 2002; Styner et al., 2005). Moreover, functional MRI (fMRI) is a valuable tool for understanding functional integration of different brain regions in response to specific stimuli and behavioral tasks and detecting the association between brain function and covariates of interest, such as diagnosis, behavioral task, severity of disease, age, or IQ (Friston, 2007; Rogers et al., 2007; Huettel et al., 2004).

The statistical analysis of neuroimaging data typically fits a general linear model or a simple linear mixed model to the data from all subjects at each voxel and then generates a statistical parametric map that contains a statistic (or a $p-$value) at each voxel (Worsley et al., 2004; Friston, 2007; Lau et al., 2008). The general linear model used in the neuroimaging literature was mainly developed for cross-sectional studies, in which neuroimaging measures from different subjects are assumed to be independent (Ashburner and Friston, 2000; Styner et al., 2007; Wager et al., 2005; Worsley et al., 2004). Most existing neuroimaging software platforms including SPM, AFNI, and FSL do not have any valid methods to analyze neuroimaging data from longitudinal studies. In contrast, the primary goal of a longitudinal neuroimaging study is to characterize individual changes in neuroimaging measurements (e.g., volumetric and morphometric) over time, and covariates of interest, such as age, diagnostic status, and gender, that influence change. A distinctive feature of longitudinal neuroimaging data is that neuroimaging data have a temporal order. Imaging measurements of the same individual usually exhibit positive correlation and the strength of the correlation decreases with the time separation. Moreover, longitudinal data may provide crucial information for a causal role of time-dependent covariates (e.g., exposure) in the disease process. Improperly handling time-dependent covariates and ignoring (or incorrectly modeling) the temporal correlation structure in imaging measures likely would influence subsequent statistical inference, such as increasing the false positive and negative errors and yield misleading scientific inference (Diggle et al., 2002; Lai and Small, 2007). However, the linear mixed models used in the neuroimaging literature cannot properly handle time-dependent covariates (Lau et al., 2008).

Statistical shape modeling and analysis have become important tools for understanding the geometric variability of the anatomical structures in various neuroimaging studies. Statistical shape models provide an efficient description (or measurement) of the morphology of the cortical and subcortical structures (e.g., hippocampus). For instance, linear shape models including the active shape model and landmark method describe shape changes as a combination of local translations (Bookstein, 1986; Cootes et al., 1995). The medial representation (m-rep) of shape provides a useful framework for describing shape variability in local thickness, bending, and widening (Fletch et al., 2004). Statistical analysis of these shape models is crucial for characterizing differences in brain structure across groups of healthy individuals and persons with various diseases, and changes of brain structure across time (Thompson and Toga, 2002; Thompson et al., 2002; Chung et al., 2005; Styner et al., 2005; Zhu et al., 2007b).

There are several important issues including multiple directions and correlation structure among different components of m-rep in developing regression models for m-rep models with a set of covariates. Although there is a sparse literature on regression modeling of a single directional observation from each subject (Mardia and Jupp, 1983; Jupp and Mardia 1989), these regression models of directional data are based on particular parametric distributions, such as the von Mises-Fisher distribution. For instance, existing circular regression models assume that the angular response follows the von Mises-Fisher distribution with either the angular mean η_i or the concentration parameter κ_i being associated with the covariates $\mathbf{x}_i$ (Gould, 1969; Johnson and Wehrly, 1978; Fisher and Lee, 1992). This circular regression model can be generalized to high-dimensional spherical data using the Fisher-Bingham family (Mardia, 1975; Mardia and Jupp, 1983). Furthermore, the spherically projected linear model for directional data assumes an offset normal distribution (Presnell, Morrison, and Littell, 1998). However, it remains unknown whether it is appropriate to use these parametric models for a single directional measure to simultaneously characterize the two spoke directions at each atom, which are correlated among themselves. Moreover, the two spoke directions may be correlated with other components of each atom and this provides further challenges in modeling the dependence structure of all components at each atom.

Chapter 2

Diagnostic Measures for GLMs with Missing Covariates

2.1 Introduction

Missing data are common in various settings, including surveys, clinical trials, and longitudinal studies. Methods for handling missing data strongly depend on the mechanism that generated the missing values as well as distributional and modeling assumptions at various stages. Therefore, the resulting estimates and tests may be sensitive to these assumptions. For this reason, sensitivity analyses are commonly done to check the sensitivity of the parameters of interest with respect to the model assumptions.

The aim of this paper is to systematically investigate various diagnostic measures for generalized linear models with Missing at Random (MAR) covariates as well as Not Missing at Random (NMAR) covariates, often referred to as nonignorably missing covariates. We propose two case-deletion measures, that is Cook's distance and the Q-displacement, based on the conditional expectation of the complete-data likelihood function in the expectation-maximization (EM) algorithm (Zhu et al., 2001). We formally define conditional residuals and examine their properties under different missing data mechanisms, such as MAR and NMAR, and then we develop conditional residual processes to construct goodness of fit statistics. Moreover, we develop specific strategies for incorporating missing covariate data into the goodness of fit statistics in order to increase the power of detecting model misspecification.

The model assessment methodology we develop here is crucial for missing data problems and the first of its kind. It is important since i) it often turns out that covariates with missing values may in fact lead to cases with influential observations and one cannot just delete the cases with missing values and carry out a complete case analysis to examine which cases are influential, ii) developing methods for assessing MAR and NMAR models and robustness is one of the most important problems in missing data, and the diagnostic and goodness of fit methodology we develop here is perfectly suited for this problem, iii) model assessment and goodness of fit in the presence of missing data is a very important problem whose development is quite different from methods based on complete data, as one needs to appropriately define residuals and other quantities in the context of missing data, and these statistics have very different small and large sample properties and operating characteristics than statistics based on complete data methods.

To motivate the proposed methodology, we consider data on 191 patients from two Eastern Cooperative Oncology Group clinical trials (Ibrahim, Chen, and Lipsitz, 1999), which is discussed in more detail in Section 2.5. The primary interest here was to find how the number of cancerous liver nodes (response) when entering the trials is predicted by six other baseline characteristics: time since diagnosis of the disease (in weeks); two biochemical markers (each classified as normal or abnormal), alpha fetoprotein and anti-hepatitis antigen; associated jaundice (yes, no); body mass index (weight in kilograms divided by the square of height in meters); and age (in years). From these six covariates, three had missing data and the remaining covariates were completely observed. The three with missing data were time since diagnosis of the disease, alpha fetoprotein, and anti-hepatitis antigen, with 8.9%, 5.8%, and 18.3% missingness percentages, yielding a total missingness percentage of 29% . Table 2.1 shows all the influential cases, where

Table 2.1: The five influential cases and the corresponding responses and covariates in the liver cancer data. The 65th and 131st observations have missing 'Anti-hepatitis antigen'.

Obs	Number of cancerous nodes	Time	Fetoprotein	Anti-hepatitis	Jaundice	BMI	Age
10	61	2.29	0	1	1	18.81	31.06
15	21	0.57	1	0	1	23.73	42.29
65	23	2.57	1	.	1	23.72	70.52
131	6	320.86	0	.	0	20.31	66.19
160	21	1.14	1	0	1	22.94	65.40

cases 10, 15, 65, and 160 have abnormally large response values, and case 131 has an extreme covariate value in time since diagnosis compared to the rest of the cases. In this paper, we will develop a formal methodology to assess such cases. In Section 2.5, we revisit this dataset and use our proposed methodology to determine whether these cases are influential or not.

The rest of this paper is organized as follows. In Section 2.2, we review the model assumptions of generalized linear models with missing covariates and related EM algorithm for calculating the maximum likelihood estimate. In Section 2.3, we develop new diagnostic measures including case-deletion diagnostics and conditional residuals and examine their properties. We construct goodness of fit statistics based on conditional residuals. We present several simulation studies in Section 2.4, and analyze the liver cancer dataset in Section 2.5. We conclude the paper with some final remarks in Section 2.6.

2.2 Preliminaries

Consider n independent observations $(\mathbf{x}_1, \mathbf{z}_1, y_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n, y_n)$, where y_i is the response variable, $\mathbf{x}_i$ is a p_1-dimensional vector of completely observed covariates, and $\mathbf{z}_i$ is a p_2-dimensional vector of partially observed covariates. Moreover let $\mathbf{z}_{m,i}$ and $\mathbf{z}_{o,i}$ denote the missing and observed components of $\mathbf{z}_i$ respectively. Let $\mathbf{r}_i$ be a p_2-dimensional random vector, whose k-th component, r_{ik}, equals 1 if z_{ik} is observed for subject i, and 0 if z_{ik} is missing, where z_{ik} is the k-th component of $\mathbf{z}_i$. Under the NMAR setting, we need to specify the joint distribution of $(\mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i, y_i)$ for each i. It is common to decompose $p(\mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i, y_i|\eta)$ into a product of three conditional distributions as follows:

$$p(\mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i, y_i|\eta) = p(y_i|\mathbf{x}_i, \mathbf{z}_i, \eta)p(\mathbf{z}_i|\mathbf{x}_i, \eta)p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \eta), \tag{2.1}$$

where η denotes the vector of all unknown parameters as defined below.

Modeling $(\mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i, y_i)$ usually involves three levels of assumptions. We assume a GLM for the conditional distribution of y_i given $(\mathbf{x}_i, \mathbf{z}_i)$ (Ibrahim, 1990; Little and Rubin, 2002; Lipsitz and Ibrahim, 1996; Ibrahim and Lipsitz, 1996). Specifically, y_i given $(\mathbf{x}_i, \mathbf{z}_i)$ has a density in the exponential family

$$p(y_i \mid \mathbf{x}_i, \mathbf{z}_i, \beta, \tau) = \exp\left\{a_i^{-1}(\tau)[y_i\theta_i(\beta) - b(\theta_i(\beta))] + c(y_i, \tau)\right\}, \tag{2.2}$$

$i = 1, \ldots, n$, indexed by the canonical parameter θ_i and the scale parameter τ, where the functions $b(\cdot)$ and $c(\cdot, \cdot)$ determine a particular distributional family in the class, such as the binomial, normal or Poisson distribution. The functions $a_i(\tau)$ are commonly of the form $a_i(\tau) = \tau^{-1}k_i^{-1}$, where the k_i's are known weights. Further, the θ_i's satisfy the equations $\theta_i = \theta(\mu_i)$, $i = 1, \ldots, n$, and $\mu_i = g((\mathbf{x}_i', \mathbf{z}_i')\beta)$ are the components of $\mu = E(\mathbf{y} \mid \mathbf{x}, \mathbf{z}, \beta, \tau)$, where $g(\cdot)$ is a known link function and $\beta = (\beta_1, \ldots, \beta_p)'$ is a p-dimensional vector of regression coefficients ($p = 1 + p_1 + p_2$). The GLMs include many well-known regression models, such as normal linear regression, logistic and probit regression, Poisson regression, gamma regression, and some proportional hazards models (McCullagh and Nelder, 1989).

We also need to specify a distribution for the missing covariates $\mathbf{z}_i$. For large p, modeling the covariates usually involves several assumptions. To reduce the number of parameters, we follow Lipsitz and Ibrahim (1996) and Ibrahim, Lipsitz, and Chen (1999) and write $p(\mathbf{z}_i|\mathbf{x}_i, \alpha)$ as a sequence of one-dimensional conditional distributions:

$$p(\mathbf{z}_i|\mathbf{x}_i, \alpha) = p(z_{ip_2}|z_{i(p_2-1)}, \cdots, z_{i1}, \mathbf{x}_i, \alpha) \cdots p(z_{i2}|z_{i1}, \mathbf{x}_i, \alpha) p(z_{i1}|\mathbf{x}_i, \alpha), \tag{2.3}$$

where α is a subvector of η. Furthermore, we typically assume specific parametric forms for these one dimensional conditional distributions. Since the $\mathbf{x}_i$'s are fully observed, it is not necessary to specify a distribution for $\mathbf{x}_i$.

One way of modeling the missing data mechanism $p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \xi)$ is to use a joint log-linear model (Lipsitz and Ibrahim, 1996). Following Ibrahim, Lipsitz, and Chen (1999), another way of modeling $p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i, \xi)$ is to assume that

$$\begin{aligned} p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i, \xi) &= p(r_{ip_2}|r_{i(p_2-1)}, \cdots, r_{i1}, \mathbf{x}_i, \mathbf{z}_i, y_i, \xi) \\ &\quad \cdots p(r_{i2}|r_{i1}, \mathbf{x}_i, \mathbf{z}_i, y_i, \xi) p(r_{i1}|\mathbf{x}_i, \mathbf{z}_i, y_i, \xi). \end{aligned} \tag{2.4}$$

It is common to use logistic regression models for the binary variables r_{ij}.

The EM algorithm has been a popular technique for obtaining the maximum likelihood estimates (MLEs) of $\eta = (\beta, \tau, \alpha, \xi)'$ in GLMs with missing covariate data (Little and Schluchter, 1985; Little and Rubin, 2002; Schluchter and Jackson, 1989; Ibrahim, 1990; Ibrahim and Lipsitz, 1996; Lipsitz and Ibrahim, 1996, 1998). Let $D_c = \{(y_j, \mathbf{x}_j, \mathbf{z}_j, \mathbf{r}_j) : j = 1, \cdots, n\}$ be the complete data, $D_o = \{\mathbf{d}_{o,i} = (y_i, \mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i) : i = 1, \cdots, n\}$ be the observed data, and $D_m = (\mathbf{z}_{m,1}, \cdots, \mathbf{z}_{m,n})$ be the missing data. At the $s-$th step of the

EM algorithm, given $\eta^{(s)}$, the E-step involves evaluating the $Q-$function given by

$$\begin{aligned}
Q(\eta|\eta^{(s)}) &= E[L_c(\eta|D_c)|D_o, \eta^{(s)}] \\
&= \sum_{i=1}^{n} \int \log[p(y_i|\mathbf{x}_i, \mathbf{z}_i, \beta, \tau)] p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta^{(s)}) d\mathbf{z}_{m,i} \\
&+ \sum_{i=1}^{n} \int \log[p(\mathbf{z}_i|\mathbf{x}_i, \alpha)] p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta^{(s)}) d\mathbf{z}_{m,i} \\
&+ \sum_{i=1}^{n} \int \log[p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i, \xi)] p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta^{(s)}) d\mathbf{z}_{m,i}, \\
&= Q_1(\beta, \tau|\eta^{(s)}) + Q_2(\alpha|\eta^{(s)}) + Q_3(\xi|\eta^{(s)}), \qquad (2.5)
\end{aligned}$$

where $L_c(\eta|D_c) = \log p(D_c|\eta)$ is the complete-data log-likelihood function. The M-step consists of maximizing $Q_1(\beta, \tau|\eta^{(s)})$, $Q_2(\alpha|\eta^{(s)})$, and $Q_3(\xi|\eta^{(s)})$, separately (Ibrahim, Lipsitz, and Chen, 1999).

Our main interest is to make valid inferences about β, and this requires the correct specification of all three levels of assumptions in (2.1). Misspecifying some of those modeling assumptions may introduce serious bias in β. Thus, it is crucial to assess the potential degree of misspecification at each of the three levels of assumptions in (2.1).

2.3 Diagnostic measures

We define the following two types of diagnostic measures: case-deletion measures and conditional residuals for formal and informal examination of the adequacy of a GLM with missing covariates. The two case-deletion measures, Cook's distance and the Q-displacement, can be used to examine the effects of deleting individual observations on the estimate of η. The conditional residuals carry important information about the influence of observations. We use the conditional residuals to construct goodness of fit statistics for testing possible invalidity of particular model assumptions.

2.3.1 Case-deletion influence measures

To quantify the effects of deleting the $i-$th observation on the MLE, $\hat{\eta}$ of η, we define the MLE of η for a subsample $D_{c[i]}$, in which the $i-$th observation $\mathbf{d}_i = (y_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i)$ is deleted from D_c. For the subsample $D_{c[i]}$, we define $Q_{[i]}(\eta|\hat{\eta})$ as $Q_{[i]}(\eta|\hat{\eta}) = E[L_c(\eta|D_{c[i]})|D_o, \hat{\eta}]$, where the expectation is taken with respect to $p(D_m|D_o, \hat{\eta})$. Then we define $\hat{\eta}_{[i]}$ as the maximizer of $Q_{[i]}(\eta|\hat{\eta})$. Following Zhu et al. (2001), we calculate a one-step approximation $\hat{\eta}^1_{[i]}$ of $\hat{\eta}_{[i]}$ as follows:

$$\hat{\eta}^1_{[i]} = \hat{\eta} + \{-\partial^2_\eta Q(\eta|\hat{\eta})\}^{-1}\partial_\eta Q_{[i]}(\eta|\hat{\eta})\big|_{\eta=\hat{\eta}}, \tag{2.6}$$

where ∂_η and ∂^2_η represent the first-order and second-order derivatives with respect to η. In (2.6), several degrees of approximation are used, but this is usually adequate for diagnostic purposes (Cook and Weisberg, 1982; Zhu et al., 2001). Because $\partial_\eta Q(\eta|\hat{\eta})|_{\eta=\hat{\eta}} = \mathbf{0}$,

$$\begin{aligned} & \partial_\eta Q_{[i]}(\eta|\hat{\eta})|_{\eta=\hat{\eta}} \\ = \quad & -\int \partial_\eta \log\{p(\mathbf{d}_i|\hat{\eta})\} p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, y_i, \mathbf{r}_i, \hat{\eta}) d\mathbf{z}_{m,i} = -\partial_\eta \log p(\mathbf{x}_i, \mathbf{z}_{o,i}, y_i, \mathbf{r}_i|\hat{\eta}). \end{aligned}$$

We introduce two case-deletion measures to quantify the distance between the MLEs of η with and without the $i-$th observation deleted from the full sample (Cook and Weisberg, 1982; Zhu et al., 2001). Cook's distance, denoted by $\mathrm{CD}_i(M)$, in this setting is defined as

$$\mathrm{CD}_i(M) = (\hat{\eta}^1_{[i]} - \hat{\eta})' M (\hat{\eta}^1_{[i]} - \hat{\eta}), \tag{2.7}$$

where M is chosen to be a positive definite matrix. For simplicity, we use CD_i to denote $\mathrm{CD}_i(M)$, when $M = -\partial^2_\eta Q(\eta|\hat{\eta})|_{\eta=\hat{\eta}}$. Similar to the likelihood displacement (Cook, 1986), the Q-displacement (Zhu et al., 2001) is defined by

$$\mathrm{QD}_i = 2\{Q(\hat{\eta}|\hat{\eta}) - Q(\hat{\eta}^1_{[i]}|\hat{\eta})\}. \tag{2.8}$$

If the value of CD_i or QD_i is large, then the $i-$th observation is influential. Similarly, we can also quantify the effects of deleting two or more observations on $\hat{\eta}$ (Cook and Weisberg, 1982, Chapter 3). For simplicity, we omit those details here.

The diagnostic measures CD_i and QD_i can be decomposed as sums of three diagnostic measures for assumptions (2.2)-(2.4) due to the decomposition in (2.5). The matrix $\partial_\eta^2 Q(\eta|\hat{\eta})$ can be written as

$$\mathrm{diag}(\partial^2_{(\beta,\tau)} Q_1(\beta,\tau|\hat{\eta}), \partial^2_\alpha Q_2(\alpha|\hat{\eta}), \partial^2_\xi Q_3(\xi|\hat{\eta})).$$

Thus, equation (2.6) can be written as

$$\begin{aligned} (\hat{\beta}^1_{[i]}, \hat{\tau}^1_{[i]}) &= (\hat{\beta},\hat{\tau}) + \{-\partial^2_{(\beta,\tau)} Q_1(\hat{\beta},\hat{\tau}|\hat{\eta})\}^{-1} \partial_{(\beta,\tau)} Q_{1[i]}(\hat{\beta},\hat{\tau}|\hat{\eta}), \\ \hat{\alpha}^1_{[i]} &= \hat{\alpha} + \{-\partial^2_\alpha Q_2(\hat{\alpha}|\hat{\eta})\}^{-1} \partial_\alpha Q_{2[i]}(\hat{\alpha}|\hat{\eta}), \qquad (2.9) \\ \hat{\xi}^1_{[i]} &= \hat{\xi} + \{-\partial^2_\xi Q_3(\hat{\xi}|\hat{\eta})\}^{-1} \partial_\xi Q_{3[i]}(\hat{\xi}|\hat{\eta}). \end{aligned}$$

Finally, $\mathrm{CD}_i = \mathrm{CD}_{i,1} + \mathrm{CD}_{i,2} + \mathrm{CD}_{i,3}$, where

$$\begin{aligned} \mathrm{CD}_{i,1} &= \{\partial_{(\beta,\tau)} Q_{1[i]}(\hat{\beta},\hat{\tau}|\hat{\eta})\}'\{-\partial^2_{(\beta,\tau)} Q_1(\hat{\beta},\hat{\tau}|\hat{\eta})\}^{-1}\{\partial_{(\beta,\tau)} Q_{1[i]}(\hat{\beta},\hat{\tau}|\hat{\eta})\}, \\ \mathrm{CD}_{i,2} &= \{\partial_\alpha Q_{2[i]}(\hat{\alpha}|\hat{\eta})\}'\{-\partial^2_\alpha Q_2(\hat{\alpha}|\hat{\eta})\}^{-1}\{\partial_\alpha Q_{2[i]}(\hat{\alpha}|\hat{\eta})\}, \qquad (2.10) \\ \mathrm{CD}_{i,3} &= \{\partial_\xi Q_{3[i]}(\hat{\xi}|\hat{\eta})\}'\{-\partial^2_\xi Q_3(\hat{\xi}|\hat{\eta})\}^{-1}\{\partial_\xi Q_{3[i]}(\hat{\xi}|\hat{\eta})\}. \end{aligned}$$

Intuitively, $\mathrm{CD}_{i,1}$ is mainly associated with the effects of removing the $i-$th observation on assumption (2.2), $\mathrm{CD}_{i,2}$ is for assumption (2.3), and $\mathrm{CD}_{i,3}$ is for assumption (2.4). Similarly, it follows from (2.5) that

$$\mathrm{QD}_i = \mathrm{QD}_{i,1} + \mathrm{QD}_{i,2} + \mathrm{QD}_{i,3}, \qquad (2.11)$$

where $\mathrm{QD}_{i,1} = 2[Q_1(\hat{\beta},\hat{\tau}|\hat{\eta}) - Q_1(\hat{\beta}^1_{[i]}, \hat{\tau}^1_{[i]}|\hat{\eta})]$, $\mathrm{QD}_{i,2} = 2[Q_2(\hat{\alpha}|\hat{\eta}) - Q_2(\hat{\alpha}^1_{[i]}|\hat{\eta})]$, and $\mathrm{QD}_{i,3} =$

$2[Q_3(\hat{\xi}|\hat{\eta}) - Q_3(\hat{\xi}^1_{[i]}|\hat{\eta})]$. Thus, $\mathrm{QD}_{i,1}$ is mainly associated with the effects of removing the $i-$th observation on assumption (2.2), $\mathrm{QD}_{i,2}$ is for assumption (2.3), and $\mathrm{QD}_{i,3}$ is for assumption (2.4). Moreover, using a Taylor's series expansion, it can be shown that $\mathrm{QD}_{i,k}$ is asymptotically equivalent to $\mathrm{CD}_{i,k}$ for each of $k = 1, 2, 3$.

2.3.2 Conditional residuals

Residuals are key tools for revealing departures from assumptions (2.2)-(2.4). Since our primary interest is to make valid inferences on assumption (2.2), we define the residual for the $i-$th observation as

$$R_i(\hat{\eta}) = y_i - g((\mathbf{x}_i', \mathbf{z}_i')\hat{\beta}).$$

However, since $\mathbf{z}_{m,i}$ is missing, $R_i(\hat{\eta})$ cannot be directly calculated for those cases with missing covariates. Generally, there are many ways of 'integrating out' $\mathbf{z}_{m,i}$. Here we focus on two kinds of conditional residuals as follows:

$$\begin{aligned}
\mathrm{CR}_i^{(1)}(\eta) &= y_i - E[g((\mathbf{x}_i', \mathbf{z}_i')\beta)|\mathbf{x}_i, \mathbf{z}_{o,i}], &(2.12)\\
\mathrm{CR}_i^{(2)}(\eta) &= y_i - E[g((\mathbf{x}_i', \mathbf{z}_i')\beta)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i], &(2.13)
\end{aligned}$$

for $i = 1, \cdots, n$, where the expectations in (2.12) and (2.13) are taken with respect to $p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \eta)$ and $p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta)$ respectively. If there are no missing covariates in $\mathbf{z}_i$, then $\mathrm{CR}_i^{(1)}(\hat{\eta})$ and $\mathrm{CR}_i^{(2)}(\hat{\eta})$ reduce to $R_i(\hat{\eta})$. Thus, the conditional residuals $\mathrm{CR}_i^{(k)}(\hat{\eta})$ for $k = 1, 2$ can be regarded as generalizations of residuals in generalized linear models (Cook and Weisberg, 1982). The conditional residuals in (2.12) and (2.13) are computationally attractive because the conditional expectations involved can be easily evaluated using Markov chain Monte Carlo (MCMC) methods (Liu, 2003; Chen et al., 2000). We note that $\mathrm{CR}_i^{(1)}(\eta)$ does not account for the missing data mechanism.

We examine several properties of the proposed conditional residuals. Through a better understanding of the properties of conditional residuals, we may develop both formal and informal diagnostic tools for the examination of the adequacy of assumption (2.2). We derive the expectations and variances of the proposed conditional residuals in the following theorems, whose assumptions and detailed proofs can be found in the Appendix.

Proposition 1

Suppose that assumptions C3 and C4 in the Appendix are true. We have the following results.

(i) $E[CR_i^{(k)}(\eta_*)|\mathbf{x}_i] = E[CR_i^{(k)}(\eta_*)] = 0$ *for* $k = 1, 2$, *where* η_* *is the true value of* η. *However,* $E[CR_i^{(k)}(\eta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, \eta_*]$ *may not equal zero for* $k = 1, 2$.

(ii) If the missing data are MAR, then $CR_i^{(2)}(\eta) = y_i - E[g((\mathbf{x}_i', \mathbf{z}_i')\beta)|\mathbf{x}_i, \mathbf{z}_{o,i}, y_i]$ *and*

$$E\left[\left.\frac{CR_i^{(1)}(\eta_*)}{p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_{o,i}, y_i, \xi_*)}\right| \mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i\right] = 0. \tag{2.14}$$

(iii) If $p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \xi) = p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, \xi)$, *then*

$$E[CR_i^{(2)}(\eta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i] = 0, \quad \textit{but} \quad E[CR_i^{(1)}(\eta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i] \neq 0.$$

(iv) $CR_i^{(1)}(\hat{\eta}) = CR_i^{(1)}(\eta_*) - [\Delta_{i1}^{(1)\prime}(\hat{\beta} - \beta_*) + \Delta_{i2}^{(1)\prime}(\hat{\alpha} - \alpha_*)][1 + o_p(1)]$, *where* $\Delta_{i1}^{(1)} = E[\partial_\beta g((\mathbf{x}_i', \mathbf{z}_i')\beta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \alpha_*]$ *and*

$$\Delta_{i2}^{(1)} = E[g((\mathbf{x}_i', \mathbf{z}_i')\beta_*)\{\partial_\alpha \log p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \alpha_*)\}'|\mathbf{x}_i, \mathbf{z}_{o,i}, \alpha_*].$$

(v) $CR_i^{(2)}(\hat{\eta}) = CR_i^{(2)}(\eta_*) - [\Delta_{i1}^{(2)\prime}(\hat{\beta} - \beta_*) + \Delta_{i2}^{(2)\prime}(\hat{\eta} - \eta_*)][1 + o_p(1)]$, *where* $\Delta_{i1}^{(2)} = E[\partial_\beta g((\mathbf{x}_i', \mathbf{z}_i')\beta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta_*]$ *and*

$$\Delta_{i2}^{(2)} = E[g((\mathbf{x}_i', \mathbf{z}_i')\beta_*)\{\partial_\eta \log p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta_*)\}'|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta_*].$$

Proposition 1 (i) shows that $\mathrm{E}[\mathrm{CR}_i^{(k)}(\eta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i]$ for $k = 1, 2$ are biased, whereas

$\mathrm{E}[\mathrm{CR}_i^{(k)}(\eta_*)|\mathbf{x}_i]$ and $\mathrm{E}[\mathrm{CR}_i^{(k)}(\eta_*)]$ are unbiased. Proposition 1 (ii) shows that the missing data indicators can be dropped from $\mathrm{CR}_i^{(2)}(\eta)$ under MAR covariates. The inverse weighted residuals are unbiased only for $\mathrm{CR}_i^{(1)}(\eta)$. Proposition 1 (iii) shows that $\mathrm{E}[\mathrm{CR}_i^{(2)}(\eta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i]$ is unbiased when the missing mechanism is independent of y_i. Proposition 1 (iv) and (v) give the first order expansions of $\mathrm{CR}_i^{(1)}(\eta)$ and $\mathrm{CR}_i^{(2)}(\eta)$, respectively. In particular, the terms involving $\Delta_{i2}^{(1)}$ and $\Delta_{i2}^{(2)}$ are due to the presence of the missing data. The matrices $\Delta_{i1}^{(k)}$ and $\Delta_{i2}^{(k)}$ for $k = 1, 2$, can be calculated using MCMC methods (Chen et al., 2000). For instance,

$$\begin{aligned} & \partial_\eta\{\log p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta_*)\} \\ = \ & \partial_\eta\{\log p(\mathbf{d}_i|\eta_*)\} - E[\partial_\eta\{\log p(\mathbf{d}_i|\eta_*)\}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta_*]. \end{aligned} \tag{2.15}$$

Thus,

$$\begin{aligned} \Delta_{i2}^{(2)} = \ & E[g((\mathbf{x}_i', \mathbf{z}_i')\beta_*)\partial_\eta\{\log p(\mathbf{d}_i|\eta_*)\}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta_*] \\ & -E[g((\mathbf{x}_i', \mathbf{z}_i')\beta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta_*]E[\partial_\eta\{\log p(\mathbf{d}_i|\eta_*)\}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \eta_*]. \end{aligned}$$

We can use MCMC methods to generate random samples from $p(\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i, \hat{\eta})$ and construct a consistent estimate for $\Delta_{i2}^{(2)}$.

The values of the standardized $\mathrm{CR}_i^{(k)}(\hat{\eta})$ may be used to detect anomalous or influential observations (Cook and Weisberg, 1982). We define a standardized conditional residual as follows:

$$\mathrm{SCR}_i^{(k)}(\hat{\eta}) = \mathrm{CR}_i^{(k)}(\hat{\eta})/\sigma_{i;k}(\hat{\eta}), \tag{2.16}$$

where $\sigma_{i;1}(\eta)^2 = \mathrm{Var}[\mathrm{CR}_i^{(1)}(\eta)|\mathbf{x}_i, \mathbf{z}_{o,i}]$ and $\sigma_{i;2}(\eta)^2 = \mathrm{Var}[\mathrm{CR}_i^{(2)}(\eta)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i]$. When model (2.1) is correctly specified, $\mathrm{SCR}_i^{(k)}(\hat{\eta})$ and $\mathrm{CR}_i^{(k)}(\hat{\eta})$ should oscillate around 0. We consider the i-th observation as an 'outlier' if $|\mathrm{SCR}_i^{(k)}(\hat{\eta})|$ is significantly greater than some threshold, such as 3. Moreover, if many $|\mathrm{SCR}_i^{(k)}(\hat{\eta})|$'s are significantly greater than zero, then one should question whether assumption (2.2) is correct. It is also worthwhile to inspect $\mathrm{SCR}_i^{(k)}(\hat{\eta})$ against some function of the data, such as the observed responses and a specific covariate, which may provide an assessment of the adequacy of assumption (2.2).

2.3.3 Goodness of fit test without incorporating missing data

There is an extensive literature on developing test statistics to check the correct specification of the conditional mean (2.17) for generalized linear models with no missing data (Su and Wei, 1991; Stute, 1997; Lin et al., 2002; Stute and Zhu, 2002). However, to the best of our knowledge, no goodness of fit test statistics have ever been developed for GLMs with missing covariate data.

We may use the two types of conditional residuals proposed in the previous subsection to develop test statistics to formally check model assumptions in a GLM with missing covariates. However, for simplicity, we temporarily drop the superscript (k) in $\mathrm{CR}_i^{(k)}(\hat{\eta})$, because the results below hold for both types of conditional residuals. These test statistics are originally designed to test the following null and alternative hypotheses:

$$\begin{aligned} H_0^{(0)} &: E[y|\mathbf{x},\mathbf{z}] = g((\mathbf{x}',\mathbf{z}')\beta) \ \text{ for some } \eta, \qquad (2.17) \\ H_1^{(0)} &: E[y|\mathbf{x},\mathbf{z}] - g((\mathbf{x}',\mathbf{z}')\beta) \neq 0 \text{ for all } \eta. \end{aligned}$$

However, because some components of $\mathbf{z}$ are missing, we may wish to test the equality

$$h(\eta|\mathbf{x}) = E[y|\mathbf{x}] - E[g((\mathbf{x}',\mathbf{z}')\beta)|\mathbf{x}] = E\{\mathrm{CR}(\eta)|\mathbf{x}\} = 0.$$

Thus, instead of testing $H_0^{(0)}$ against $H_1^{(0)}$, we test the following null and alternative hypotheses:

$$H_0^{(1)} : h(\eta|\mathbf{x}) = 0 \text{ for some } \eta \qquad (2.18)$$
$$H_1^{(1)} : h(\eta|\mathbf{x}) \neq 0 \text{ for all } \eta.$$

Note that $h(\eta|\mathbf{x}) = 0$ is only a necessary condition of $E[y|\mathbf{x}, \mathbf{z}] = g((\mathbf{x}', \mathbf{z}')\beta)$. Thus, accepting $h(\eta|\mathbf{x}) = 0$ does not imply the acceptance of $H_0^{(0)}$.

We can construct statistics for testing $H_0^{(1)}$ as follows. Following Theorem 1 in Bierens (1992), $E\{\mathrm{CR}(\eta)|\mathbf{x}\} = 0$ is equivalent to $E\{\mathrm{CR}(\eta)|\mathbf{x}'\varphi\} = 0$ for any $\varphi \in R^{p_1}$. Thus, as shown in Lemma 1 of Escanciano (2006), $H_0^{(1)}$ is equivalent to

$$E\{\mathrm{CR}(\eta)\mathbf{1}(\mathbf{x}'\varphi \leq t)\} = 0 \qquad (2.19)$$

for almost every (φ, t). To test $H_0^{(1)}$, we may define a stochastic process as follows:

$$I_1((\varphi, t); \eta) = n^{-1/2} \sum_{i=1}^{n} \mathbf{1}(\mathbf{x}_i'\varphi \leq t)\mathrm{CR}_i(\eta),$$

where $(\varphi, t) \in \Pi = \{\varphi \in R^{p_1} : \|\varphi\| = 1\} \times [-\infty, \infty]$, in which $\|\cdot\|$ is the common L_2-norm in Euclidean space. Graphically, for a specific direction φ, we can plot $I_1((\varphi, t); \eta)$ as a function of t and use it as an exploratory tool for the detection of model specification along the direction φ (Lin et al., 2002). For instance, we may set $\varphi = (\hat{\beta}_1, \cdots, \hat{\beta}_{p_1})'$.

Theoretically, we regard $I_1((\varphi, t); \eta)$ as a stochastic process indexed by (φ, t) and then we use $I_1((\varphi, t); \eta)$ to construct two test statistics. We first define a conditional Kolmogorov test (CK) as

$$\mathrm{CK}_1 = \max_{(\varphi,t)} |I_1((\varphi, t); \hat{\eta})|. \qquad (2.20)$$

We also define a Cramer-von Mises test as follows:

$$\mathrm{CM}_1 = \int_{\Pi} |I_1((\varphi, t); \hat{\eta})|^2 F_{n,\varphi}(dt)d\varphi, \qquad (2.21)$$

where $F_{n,\varphi}(t)$ is the empirical distribution function of $\{\mathbf{x}_i'\varphi : i = 1, \cdots, n\}$ (Stute, 1997). Large values of CM_1 and CK_1 lead to rejection of $H_0^{(1)}$.

We note that CM_1 has several distinctive features (Escanciano, 2006). The statistic CM_1 has a closed form, whereas computing the Kolmogorov-type supremum statistic of the residual process involves high-dimensional maximizations (Stute, 1997; Lin et al., 2002). Particularly, when the dimension of the covariate vector is high or even moderate, it can be computationally demanding to compute the Kolmogorov supremum statistic. Thus, CM_1 avoids the problem of the curse of dimensionality. We are now led to the following theorem.

Theorem 1

Suppose that assumptions C1-C7 in the Appendix are true. Under the null hypothesis $H_0^{(1)}$, we have the following results:

(i) $\sqrt{n}(\hat{\eta} - \eta_*) = n^{-1/2} \sum_{i=1}^n \psi_{n,i} + o_p(1)$, *with* $\psi_{n,i} = M_n(\eta_*)^{-1} \dot{\ell}_i(\eta_*)$, *where* $\dot{\ell}_i(\eta_*) = \partial_\eta \log p(\mathbf{d}_{o,i}|\eta_*)$ *and* $M_n(\eta_*) = n^{-1} \sum_{i=1}^n E[\partial_\eta^2 \log p(\mathbf{d}_{o,i}|\eta_*)]$.

(ii) $(I_1(\cdot;\eta_*), \sqrt{n}(\hat{\eta} - \eta_*)')'$ *converges in distribution to* $(G_1(\cdot), \nu_1')'$, *where* $(G_1(\cdot), \nu_1')$ *is a mean zero Gaussian process with covariance function*

$$C_1((\varphi_1, t_1), (\varphi_2, t_2)) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^n \begin{pmatrix} CR_i(\eta_*)\mathbf{1}(\mathbf{x}_i'\varphi_1 \le t_1) \\ \psi_{n,i}(\eta_*) \end{pmatrix} \begin{pmatrix} CR_i(\eta_*)\mathbf{1}(\mathbf{x}_i'\varphi_2 \le t_2) \\ \psi_{n,i}(\eta_*) \end{pmatrix}'.$$

(iii) CK_1 *and* CM_1 *converge in distribution to* $\sup_{(\varphi,t)} |G_1(\varphi, t) + \Delta_1(\varphi, t)'\nu_1|$ *and* $\int_\Pi |G_1(\varphi, t) + \Delta_1(\varphi, t)'\nu_1|^2 F_\varphi(dt)d\varphi$, *respectively, where* $F_\varphi(t)$ *is the limiting cumulative distribution function of* $F_{n,\varphi}(t)$ *and* $\Delta_1(\varphi, t)$ *is defined by*

$$\Delta_1(\varphi, t) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^n E\{\mathbf{1}(\mathbf{x}_i'\varphi \le t)\partial_\eta[CR_i(\eta_*)]\}.$$

Theorem 1 formally characterizes the asymptotic null distributions of CK_1 and CM_1.

Therefore, we may directly approximate those distributions in order to calculate the $p-$values of the test statistics CK_1 and CM_1.

The next result establishes the asymptotic distributions of CK_1 and CM_1 under a sequence of local alternatives converging to the null at a parametric rate $n^{-1/2}$. We consider the local alternatives such that $p(y_i|\mathbf{x}_i, \mathbf{z}_i)$ belongs to the exponential family (2.2) and

$$E[y_i|\mathbf{x}_i, \mathbf{z}_i] = g((\mathbf{x}_i', \mathbf{z}_i')\beta_*) + n^{-1/2} g_0(\mathbf{x}_i, \mathbf{z}_i) \tag{2.22}$$

for $i = 1, \cdots, n$, where $g_0(\mathbf{x}_i, \mathbf{z}_i)$ is a function of $(\mathbf{x}_i, \mathbf{z}_i)$. Let $\theta_i(t) = \dot{b}^{-1}(g((\mathbf{x}_i', \mathbf{z}_i')\beta_*) + t g_0(\mathbf{x}_i, \mathbf{z}_i))$, where $\dot{b}$ denotes $\partial_t b(t)$ and $\dot{b}^{-1}(\cdot)$ is the inverse function of $\dot{b}(\cdot)$. Then, we have

$$\theta_i = \theta_i(n^{-1/2}) = \dot{b}^{-1}(g((\mathbf{x}_i', \mathbf{z}_i')\beta_*) + n^{-1/2} g_0(\mathbf{x}_i, \mathbf{z}_i)).$$

Thus, the true distribution of y_i given $(\mathbf{x}_i, \mathbf{z}_i)$, denoted by $p(y_i|\mathbf{x}_i, \mathbf{z}_i, n^{-1/2})$, is

$$\exp\left\{a_i^{-1}(\tau_*)[y_i\theta_i(n^{-1/2}) - b(\theta_i(n^{-1/2}))] + c(y_i, \tau_*)\right\}. \tag{2.23}$$

Moreover, $p(\mathbf{x}_i, \mathbf{z}_i|\alpha_*)$ and $p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \xi_*)$ are unchanged. We are now led to the following results.

Theorem 2

Suppose that assumptions C1-C7 in the Appendix and the sequence of models in (2.23) are true. We have the following results:

(i) $\sqrt{n}(\hat{\eta} - \eta_*)$ converges in distribution to $\nu_1 + A_1$, where ν_1 is the same normal distribution as in Theorem 1 and

$$A_1 = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \psi_{n,i} E[a_i(\tau_*)^{-1} \partial_t \theta_i(0)(y_i - g((\mathbf{x}_i', \mathbf{z}_i')\beta_*))|\mathbf{d}_{o,i}].$$

(ii) $I_1(\cdot; \eta_*)$ converges in distribution to $G_1(\cdot) + A_2(\cdot)$, where $G_1(\cdot)$ is the same process as in Theorem 1. In addition, $A_2(\varphi, t) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbf{1}(\varphi'\mathbf{x}_i \leq t) E[g_0(\mathbf{x}_i, \mathbf{z}_i)|\mathbf{x}_i, \mathbf{z}_{o,i}]$

for $CR_i^{(1)}$*, whereas* $A_2(\varphi, t) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n \mathbf{1}(\varphi'\mathbf{x}_i \le t)E[g_0(\mathbf{x}_i, \mathbf{z}_i)|\mathbf{d}_{o,i}]$ *for* $CR_i^{(2)}$.

(iii) CK_1 *and* CM_1 *converge in distribution to* $\sup_{(\varphi,t)} |G_1(\varphi, t)+A_2(\varphi, t)+\Delta_1(\varphi, t)'(\nu_1+A_1)|$ *and* $\int_\Pi |G_1(\varphi, t) + A_2(\varphi, t) + \Delta_1(\varphi, t)'(\nu_1 + A_1)|^2 F_\varphi(dt)d\varphi$*, respectively.*

2.3.4 Goodness of fit test incorporating missing data

We propose to use the missing covariates $\mathbf{z}_i$ to improve the power of $I_1((\varphi, t); \eta)$ in detecting the misspecification of $g((\mathbf{x}', \mathbf{z}')\beta)$. Recall that $h(\eta|\mathbf{x}) = 0$ is only a necessary condition of $E[y|\mathbf{x}, \mathbf{z}] = g((\mathbf{x}', \mathbf{z}')\beta)$. Because $\mathbf{1}(\mathbf{x}'\varphi \le t)$ in $I_1((\varphi, t); \eta)$ does not involve the missing covariates $\mathbf{z}$, we may lose power in detecting the misspecification of $H_0^{(0)}$ in the missing covariate space. In particular, if the fraction of missing covariates is small, then it is very inefficient to drop all the information in $\mathbf{z}$.

We may test whether $H_0^{(0)}$ is true using the additional information contained in the missing covariates. Let $\mathbf{z}_{m,i}(\alpha) = E[\mathbf{z}_{m,i}|\mathbf{x}_i, \mathbf{z}_{o,i}, \alpha]$, we suggest replacing $\mathbf{z}_{m,i}$ by $\mathbf{z}_{m,i}(\hat\alpha)$, which is an imputed missing covariate vector. However, developing test statistics based on the imputed missing covariates depends on the specific missing data mechanism.

We first consider the case that $p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i)$ is independent of y_i. Using Proposition 1 (iii), we can show that

$$\begin{aligned} & E[\mathrm{CR}_i^{(2)}(\eta_*)\mathbf{1}(\mathbf{c}_{i*}'\tilde\varphi \le t)|\mathbf{x}_i, \mathbf{z}_{o,i}] \\ = \; & E\{E[\mathrm{CR}_i^{(2)}(\eta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i]\mathbf{1}(\mathbf{c}_{i*}'\tilde\varphi \le t)|\mathbf{x}_i, \mathbf{z}_{o,i}\} = 0, \end{aligned} \qquad (2.24)$$

for all $i = 1, \cdots, n$, where $(\tilde\varphi, t) \in \Pi = \{\tilde\varphi \in R^{p_1+p_2} : \|\tilde\varphi\| = 1\} \times [-\infty, \infty]$, and $\mathbf{c}_{i*} = \mathbf{c}_i(\alpha_*) = (\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{z}_{m,i}(\alpha_*))$. In addition, $\mathbf{c}_i(\alpha)$ is defined as

$$\mathbf{c}_i(\alpha) = (\mathbf{x}_i, r_{i1}z_{i1} + (1 - r_{i1})E[z_{i1}|\mathbf{x}_i, \mathbf{z}_{o,i}, \alpha], \cdots, r_{ip_2}z_{ip_2} + (1 - r_{ip_2})E[z_{ip_2}|\mathbf{x}_i, \mathbf{z}_{o,i}, \alpha])\,.$$

Let $\hat{\mathbf{c}}_i = \mathbf{c}_i(\hat\alpha)$. We are thus able to incorporate the additional information from $\mathbf{z}_{o,i}$ into the indicator function $\mathbf{1}(\hat{\mathbf{c}}_i'\tilde\varphi \le t)$. Following the reasoning in (2.24), we now propose the stochastic process:

$$I_2((\tilde{\varphi},t);\eta)=n^{-1/2}\sum_{i=1}^{n}\mathbf{1}(\hat{\mathbf{c}}_i'\tilde{\varphi}\leq t)\mathrm{CR}_i^{(2)}(\eta). \tag{2.25}$$

We first suggest plotting $I_2((\tilde{\varphi},t)$ against t for a specific $\tilde{\varphi}$ as an exploratory tool for detecting the form of misspecification of assumption (2.2). For instance, we may set $\tilde{\varphi}=\hat{\beta}$. Then, we develop the corresponding CK and CM statistics based on $I_2((\tilde{\varphi},t);\hat{\eta})$, denoted by CK_2 and CM_2. Large values of CK_2 and CM_2 lead to reject $E[\mathrm{CR}_i^{(2)}(\eta)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i]=0$.

Secondly, suppose that the missing data are MAR. Using Proposition 1 (ii), we can show that for $i=1,\cdots,n$,

$$\begin{aligned}&E[\frac{\mathrm{CR}_i^{(1)}(\eta_*)}{p(\mathbf{r}_i|\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\xi_*)}\mathbf{1}(\mathbf{c}_{i*}'\tilde{\varphi}\leq t)|\mathbf{x}_i,\mathbf{z}_{o,i}]\\ =\;&E\{E[\frac{\mathrm{CR}_i^{(1)}(\eta_*)}{p(\mathbf{r}_i|\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\xi_*)}|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i]\mathbf{1}(\mathbf{c}_{i*}'\tilde{\varphi}\leq t)|\mathbf{x}_i,\mathbf{z}_{o,i}\}=0.\end{aligned}$$

Then, we propose an inverse weighted process as follows:

$$I_3((\tilde{\varphi},t);\eta)=n^{-1/2}\sum_{i=1}^{n}\mathbf{1}(\hat{\mathbf{c}}_i'\tilde{\varphi}\leq t)\mathrm{CR}_i^{(1)}(\eta)/p(\mathbf{r}_i|\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\xi). \tag{2.26}$$

We may plot $I_3((\tilde{\varphi},t)$ against t for a specific $\tilde{\varphi}$ as an exploratory tool for detecting the assumption of MAR. Similar to (2.21) and (2.20), we can develop the corresponding CK and CM statistics based on $I_3((\tilde{\varphi},t);\hat{\eta})$ and denote them by CK_3 and CM_3. Large values of CK_3 and CM_3 lead to reject $E[\mathrm{CR}_i^{(1)}(\eta_*)/p(\mathbf{r}_i|\mathbf{x}_i,\mathbf{z}_{o,i},y_i,\xi_*)|\mathbf{x}_i,\mathbf{z}_{o,i},\mathbf{r}_i]=0$.

Similar to Theorems 1 and 2, we can establish the asymptotic distributions of CK_k and CM_k and their power behavior under local alternatives for $k=2,3$. For simplicity, we only include the asymptotic null distributions of $I_2((\tilde{\varphi},t);\eta_*)$ below.

Theorem 3

Suppose that assumptions C1-C8 in the Appendix are true. Under the null hypothesis $H_0^{(0)}$, $I_2(\cdot;\eta_*)$ converges in distribution to $G_2(\cdot)$, where $G_2(\cdot)$ is a mean zero Gaussian *process with covariance function*

$$C_2((\tilde{\varphi}_1, t_1), (\tilde{\varphi}_2, t_2)) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} [CR_i^{(2)}(\eta_*)]^2 \mathbf{1}(\mathbf{c}'_{i*}\tilde{\varphi}_1 \le t_1)\mathbf{1}(\mathbf{c}'_{i*}\tilde{\varphi}_2 \le t_2).$$

The CK_k and CM_k for $k = 2, 3$ differ from CK_1 and CM_1 in several aspects. The CK_1 and CM_1 focus on testing $H_0^{(1)}$ regardless of the missing data mechanism and the type of conditional residual, whereas large values of CK_k and CM_k for $k = 2, 3$ can be caused by the misspecification of the missing data mechanism. For instance, CK_2 and CM_2 test whether $E[\mathrm{CR}_i^{(2)}(\eta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i]$ equals zero or not, whereas CK_3 and CM_3 test whether $E[\mathrm{CR}_i^{(1)}(\eta_*)/p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \xi_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i]$ equals zero or not. The rejection of $E[\mathrm{CR}_i^{(2)}(\eta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i] = 0$ may be caused by the dependence of $p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \xi)$ on the response y_i, while the rejection of $E[\mathrm{CR}_i^{(1)}(\eta_*)/p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \xi_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i] = 0$ can be caused by NMAR covariate data. Thus CK_k and CM_k, $k = 2, 3$ are useful goodness of fit statistics for testing the missing data mechanism.

2.3.5 Resampling method

In the following, we devise a resampling method to approximate the $p-$value of CK_1. We can develop similar methods for CK_k, $\mathrm{CM}_j, k = 2, 3, j = 1, 2, 3$. There are four steps in generating the stochastic processes that have the same asymptotic distributions as $I_1((\varphi, t); \hat{\eta})$.

Step 1. Generate independent and identically distributed random samples, $\{v_i^{(q)} : i = 1, \cdots, n\}$, from a $N(0, 1)$ distribution for $q = 1, \cdots, Q$, where Q is the number of replications, say $Q = 1000$.

Step 2. Calculate

$$\mathrm{I}_1((\varphi, t); \hat{\eta})^{(q)} = n^{-1/2} \sum_{i=1}^{n} v_i^{(q)} \{\mathrm{CR}_i(\hat{\eta})\mathbf{1}(\mathbf{x}'_i\varphi \le t) - \hat{\Delta}_1(\varphi, t)\psi_{ni}\}$$

where $\hat{\Delta}_1(\varphi, t) = n^{-1}\sum_{i=1}^{n} \partial_\eta \mathrm{CR}_i(\hat{\eta})\mathbf{1}(\mathbf{x}_i'\varphi \leq t)$. Note that conditional on the observed data, since $I_1((\varphi, t); \hat{\eta})^{(q)}$ is the sum of independent but not identically distributed stochastic process, it follows from some mild conditions that $I_1((\varphi, t); \hat{\eta})^{(q)}$ converges weakly to the desired Gaussian process in Theorem 1 as $n \to \infty$ (Kosorok, 2003; van der Vaart and Wellner, 1996; Stute et al., 1998).

Step 3. Calculate the test statistics $\mathrm{CK}_1^{(q)} = \sup_{(\varphi,t)} |\mathrm{I}_1((\varphi, t); \hat{\eta})^{(q)}|$ and obtain $\{\mathrm{CK}_1^{(q)} : q = 1, \cdots, Q\}$.

Step 4. Calculate the $p-$value of CK_1 using $\{\mathrm{CK}_1^{(q)} : q = 1, \cdots, Q\}$.

2.4 Simulation studies

We conducted Monte Carlo simulations to examine the finite sample performance of the various diagnostic measures proposed here. First, we applied case-deletion measures and standardized conditional residuals to a simulated dataset based on a linear model, in which an 'outlier' was added. We expected that the diagnostic measures would detect the 'outlier'. Secondly, we evaluated the rejection rates of the Type I and Type II errors for CM_1 based on the conditional residuals $\mathrm{CR}_i^{(1)}$ and $\mathrm{CR}_i^{(2)}$, for CM_2, and for CM_3 respectively. For the sake of simplicity, we omitted the results based on CK_k for $k = 1, 2, 3$ to save space, since they have similar Type I and Type II errors as CM_k. Furthermore, we evaluated the rejection rates for CM_1 based on the conditional residuals $\mathrm{CR}_i^{(1)}$ and $\mathrm{CR}_i^{(2)}$ and for CM_2 for a simulated logistic regression model.

2.4.1 Case-deletion measures and conditional residuals for a linear model

We considered the linear model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i, \tag{2.27}$$

where the ϵ_i's are iid and $\epsilon_i \sim N(0, \tau), i = 1, \ldots, n$. We assume that y_i and x_i are completely observed for $i = 1, \ldots, n$, but the covariate z_i may be missing for some cases. We set $n = 100$, $\beta_0 = \beta_1 = \beta_2 = 1$ and $\tau = 1$. Moreover, we independently generated 100 random vectors (x_i, z_i) from a $N_2(\mathbf{0}, \mathbf{I}_2)$ distribution. We also assumed the covariates are MAR,

$$p(r_i = 1 | x_i, z_i, y_i) = \frac{\exp(\xi_0 + \xi_1 x_i)}{1 + \exp(\xi_0 + \xi_1 x_i)}, \tag{2.28}$$

with $\xi_0 = -1.5$ and $\xi_1 = 1.0$ to obtain an average missingness fraction of 20%.

We first changed the last response y_{100} to $y_{100} + 5.0$ in order to add an 'outlier' to the dataset. We fit the linear model assuming a MAR z_i for the simulated data and a normal distribution for z_i. Then we calculated case-deletion measures and conditional residuals for each observation. The last observation was classified as the most influential observation by CD_i and $\mathrm{CD}_{i,1}$, but not $\mathrm{CD}_{i,2}$, because we only changed y_{100} in the response space (Figure 2.1 a, b, c). Specifically, $\mathrm{CD}_{100} = 4.378$ is much larger than the second largest $\mathrm{CD}_i = 0.33$ (Figure 2.1 a). Moreover, we obtained similar findings based on QD_i, $\mathrm{QD}_{i,1}$ and $\mathrm{QD}_{i,2}$ (not presented here). The standardized conditional residuals with $\mathrm{SCR}_{100}^{(1)} = 3.256$ also identified y_{100} as an influential observation (Figure 2.1 d).

Now, instead of changing the last response y_{100}, we changed z_{100} to $z_{100} + 5.0$ to add an outlier in the covariate space, and fit the same linear model assuming a MAR z_i. The last observation was classified as the most influential observation by CD_i, $\mathrm{CD}_{i,1}$, and $\mathrm{CD}_{i,2}$ (Figure 2.1 e, f, g). In contrast to the previous case in which y_{100} was changed, both $\mathrm{CD}_{i,1}$ and $\mathrm{CD}_{i,2}$ detected the influential observation z_{100} (Figure 2.1 f, g), because changing z_{100} affected the first two components of (2.1). The standardized conditional residuals $\mathrm{SCR}_i^{(k)}$ for $k = 1, 2$ identified the last observation as influential (Figure 2.1 h).

2.4.2 Goodness of fit statistics for the linear model

We systematically assessed the goodness of fit statistics based on the conditional residuals developed in Section 2.3 under various scenarios. We used 500 replications to calculate

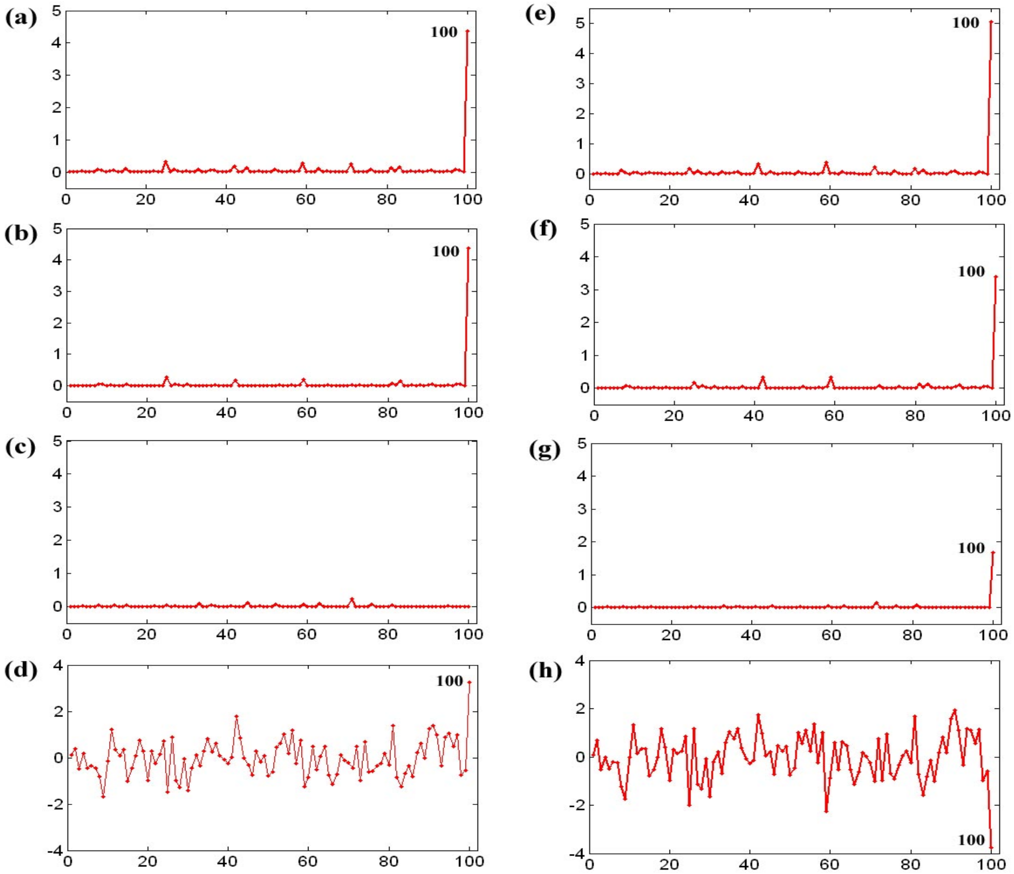

Figure 2.1: Index plots of diagnostic measures from two simulated datasets: (a) CD_i; (b) $\mathrm{CD}_{i,1}$; (c) $\mathrm{CD}_{i,2}$; (d) $\mathrm{SCR}_i^{(1)}$; (e) CD_i; (f) $\mathrm{CD}_{i,1}$; (g) $\mathrm{CD}_{i,2}$; (h) $\mathrm{SCR}_i^{(1)}$. Column one shows the results from the simulated data with y_{100} as an influential point, whereas column two shows the results from the simulated data with z_{100} as an influential point.

the $p-$values of all test statistics. The significance level was always fixed at 0.05.

We considered three groups of simulation studies. The first group of simulation studies was to compare the finite sample performance of CM_1 using either $\mathrm{CR}_i^{(1)}$ or $\mathrm{CR}_i^{(2)}$ under two scenarios. In the first scenario, we simulated 500 data sets from $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2 + \epsilon_i$ for $i = 1, \cdots, 100$, where (x_i, z_i) were generated from a $N_2(\mathbf{0}, \mathbf{I}_2)$ distribution, the ϵ_i's are iid and $\epsilon_i \sim N(0, \tau), i = 1, \ldots, n$, and c is in the range [0, 1]. We set $\beta_0 = \beta_1 = \beta_2 = 1$. We assumed that the covariate z_i has a normal distribution. We considered two missing data mechanisms: MCAR and MAR. For MAR, the missing data mechanism was given by (2.28), in which we set $\xi_1 = 1.0$

Table 2.2: Rejection rates for CM_1 using either $\mathrm{CR}_i^{(1)}$ or $\mathrm{CR}_i^{(2)}$ at the 5% significance level for the linear model. The first half shows the results for the first scenario where the misspecification is due to x_i, and the second half of the table for the second scenario where the misspecification is due to z_i.

	$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2 + \epsilon_i$											
	MCAR						MAR					
	20%		40%		60%		20%		40%		60%	
c	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$
0	0.054	0.046	0.036	0.046	0.046	0.044	0.062	0.040	0.044	0.046	0.038	0.042
0.2	0.312	0.424	0.192	0.316	0.188	0.298	0.248	0.344	0.234	0.324	0.212	0.282
0.4	0.786	0.874	0.652	0.818	0.658	0.802	0.798	0.864	0.700	0.826	0.706	0.772
0.6	0.966	0.974	0.928	0.972	0.922	0.938	0.968	0.980	0.952	0.978	0.934	0.936
0.8	0.986	0.992	0.978	0.990	0.976	0.986	0.992	0.996	0.990	0.994	0.958	0.964
1.0	1.000	1.000	0.984	0.994	0.986	0.992	1.000	1.000	1.000	1.000	0.996	1.000
	$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2 + \epsilon_i$											
	MCAR						MAR					
	20%		40%		60%		20%		40%		60%	
c	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$	$\mathrm{CR}_i^{(1)}$	$\mathrm{CR}_i^{(2)}$
0	0.054	0.046	0.036	0.046	0.046	0.044	0.062	0.040	0.044	0.046	0.038	0.042
0.2	0.062	0.062	0.034	0.056	0.032	0.034	0.058	0.048	0.032	0.028	0.038	0.042
0.4	0.044	0.048	0.058	0.058	0.018	0.028	0.040	0.036	0.030	0.044	0.036	0.040
0.6	0.046	0.052	0.046	0.050	0.034	0.048	0.032	0.036	0.028	0.034	0.030	0.036
0.8	0.050	0.062	0.040	0.054	0.048	0.056	0.040	0.042	0.038	0.040	0.032	0.042
1.0	0.052	0.068	0.048	0.058	0.044	0.052	0.046	0.050	0.034	0.052	0.038	0.044

and ξ_0 with values $-1.5,\ -0.5,\ 0.5$ to obtain average missing data fractions of 20%, 40%, 60%, respectively. Then we fit $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ with MAR, and thus the fitted model would be misspecified if $c \neq 0$ and the misspecification is due to the fully observed covariate x_i.

The top half of Table 2.2 shows the rejection rates of CM_1 based on both $\mathrm{CR}_i^{(1)}$ and $\mathrm{CR}_i^{(2)}$ for this scenario. The Type I error rates are accurate across all missingness fractions. The CM_1 based on $\mathrm{CR}_i^{(2)}$ is uniformly more powerful than that based on $\mathrm{CR}_i^{(1)}$. Consistent with our expectations, the power for detecting misspecification of the model increases with $|c|$ for CM_1. The missing data fraction slightly influences the power of detecting model misspecification for CM_1.

In the second scenario, we generated 500 data sets from $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2 + \epsilon_i$, whereas the rest of the setup remained the same as in the first scenario described earlier. We fit $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ assuming MAR, and thus the model would be misspecified if $c \neq 0$ and the misspecification is due to the missing covariate z_i. The rejection rates are shown in the second half of Table 2.2. We found that CM_1 based on both $\text{CR}_i^{(1)}$ and $\text{CR}_i^{(2)}$ cannot detect the misspecification of $c z_i^2$, because CM_1 did not incorporate the missing covariate z_i. Comparing the top half with the second half in Table 2.2 reveals the importance of incorporating the misspecified covariate in the indicator function $\mathbf{1}(\mathbf{x}_i' \varphi \leq t)$.

The second group of simulation studies was to assess the finite sample performance of CM_2. Firstly, we evaluated the power of CM_2 in detecting the misspecification of $E[y_i | x_i, z_i]$. We used the same two scenarios in the first group of simulations, and in each case we fit the linear model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ assuming z_i is MAR.

The first half of Table 2.3 shows the results for the first scenario where the misspecification is due to x_i, and the second half of the table for the second scenario where the misspecification is due to z_i. The Type I errors rates of CM_2 are accurate across all missingness fractions. For both scenarios, the power for detecting misspecification of the model increased with $|c|$ for CM_2 and the missing data fraction influences the power in detecting model misspecification (i.e., $|c| \neq 0$). Compared to Table 2.2, when $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2 + \epsilon_i$ is the true model (i.e. the first scenario), CM_1 based on $\text{CR}_i^{(2)}$ is slightly more powerful than CM_2 in detecting the presence of $c x_i^2$. However, if $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2 + \epsilon_i$ is the true model (i.e. the scecond scenario), then CM_2 is much more powerful than CM_1 based on $\text{CR}_i^{(2)}$. This indicates that incorporating the missing data can increase the power of detecting model misspecification due to $c z_i^2$.

We checked the influence of the misspecified parametric assumptions for the missing covariate distribution on the finite sample performance of CM_2. Again, we used the same two settings as before except for one change: z_i were generated from $U[-3, 3]$,

Table 2.3: Rejection rates for CM_2 at the 5% significance level for the linear model. The first half shows the results for the first scenario where the misspecification is due to x_i, and the second half of the table for the second scenario where the misspecification is due to z_i.

$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2 + \epsilon_i$						
	MCAR			MAR		
c	20%	40%	60%	20%	40%	60%
0	0.044	0.046	0.044	0.046	0.044	0.042
0.2	0.264	0.164	0.170	0.222	0.210	0.204
0.4	0.756	0.648	0.622	0.716	0.678	0.630
0.6	0.938	0.906	0.854	0.946	0.906	0.890
0.8	0.972	0.970	0.966	0.990	0.970	0.938
1.0	1.000	0.984	0.980	1.000	0.986	0.980

$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2 + \epsilon_i$						
	MCAR			MAR		
c	20%	40%	60%	20%	40%	60%
0	0.044	0.046	0.044	0.046	0.044	0.042
0.2	0.176	0.086	0.058	0.182	0.080	0.058
0.4	0.504	0.254	0.092	0.464	0.262	0.112
0.6	0.730	0.336	0.130	0.756	0.408	0.146
0.8	0.790	0.540	0.192	0.820	0.552	0.188
1.0	0.882	0.558	0.208	0.850	0.600	0.280

Table 2.4: Rejection rates for CM_2 at the 5% significance level for the linear model with a misspecified covariate. The missing covariates z_i are generated from a uniform distribution, whereas a normal distribution is assumed for z_i when we fit the data. The first half shows the results for the first scenario where the misspecification is due to x_i, and the second half of the table is for the second scenario where the misspecification is due to z_i.

$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2 + \epsilon_i$						
	MCAR			MAR		
c	20%	40%	60%	20%	40%	60%
0	0.040	0.053	0.053	0.040	0.042	0.038
0.2	0.262	0.246	0.244	0.298	0.176	0.210
0.4	0.766	0.760	0.664	0.782	0.728	0.698
0.6	0.962	0.936	0.946	0.954	0.942	0.880
0.8	0.980	0.964	0.964	0.994	0.980	0.974
1.0	0.998	0.982	0.980	0.994	0.984	0.982

$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2 + \epsilon_i$						
	MCAR			MAR		
c	20%	40%	60%	20%	40%	60%
0	0.040	0.054	0.044	0.040	0.038	0.042
0.2	0.090	0.064	0.056	0.078	0.076	0.042
0.4	0.178	0.120	0.058	0.164	0.078	0.050
0.6	0.420	0.164	0.088	0.446	0.178	0.088
0.8	0.680	0.358	0.102	0.664	0.332	0.130
1.0	0.840	0.510	0.152	0.852	0.508	0.150

a uniform distribution, instead of a $N(0,1)$ distribution. We fit the linear model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ and assumed that z_i is MAR and has a normal distribution. The first half of Table 2.4 shows the results for the first scenario where the misspecification is due to x_i, and the second half of the table is for the second scenario where the misspecification is due to z_i. Compared to Table 2.3, When $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2 + \epsilon_i$ is the true model (i.e. the first scenario), the misspecified covariate distribution for z_i has little effect on the statistical power of detecting the presence of $c x_i^2$. However, if $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2 + \epsilon_i$ is the true model (i.e. the scecond scenario), the misspecified covariate distribution for z_i has a clear effect on the statistical power of detecting the presence of $c z_i^2$, especially when the missing data fraction is large. This indicates that the covariate distribution may have a profound effect on the finite sample performance of our goodness of fit tests.

Moreover, we assessed the power of CM$_2$ in detecting whether the missing data mechanism dependeds on the response variable. Specifically, 500 data sets were generated from $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ assuming z_i is MAR,

$$p(r_i = 1|x_i, z_i, y_i) = \frac{\exp(\xi_0 + \xi_1 x_i + a y_i)}{1 + \exp(\xi_0 + \xi_1 x_i + a y_i)},$$

for $i = 1, \cdots, 100$, where $\beta_0 = \beta_1 = \beta_2 = 1$ and $\epsilon_i \sim N(0, 1)$. We fit the linear model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ under (2.28). The rejection rates were $0.045, 0.198, 0.328$ and 0.358 for $a = 0.0, 1.0, 1.5$, and 2.0 respectively. Thus, CM$_2$ can detect the dependence of the missing data mechanism on the response for large values of $|a|$.

The third group of simulation studies was to assess the finite sample performance of CM$_3$. Firstly, we evaluated the power of CM$_3$ in detecting the misspecification of $E[y_i|x_i, z_i]$ when the missing data mechanism is dependent on the response variable. We simulated 500 datasets using the second scenario in the first group of simulation studies, and then we fit the linear model assuming an MAR missing data mechanism

$$p(r_i = 1|x_i, z_i, y_i) = \frac{\exp(\xi_0 + \xi_1 y_i)}{1 + \exp(\xi_0 + \xi_1 y_i)}, \tag{2.29}$$

with various values of ξ_0 and ξ_1 to obtain the desired average missing data fractions. The rejection rates of CM$_3$ were $0.051, 0.380, 0.514$, and 0.594 for $c = 0.0, 0.5, 1.0$, and 1.5, respectively, assuming a 60% missingness fraction for z_i.

Furthermore, we assessed the power of CM$_3$ in detecting whether the missing data mechanism is nonignorable. The 500 data sets were generated from $y_i = 1 + x_i + z_i + \epsilon_i$ for $i = 1, \cdots, 100$, where $\epsilon_i \sim N(0, 1)$, and the missing data mechanism is

$$p(r_i = 1|x_i, z_i, y_i) = \frac{\exp(\xi_0 + \xi_1 y_i + a z_i)}{1 + \exp(\xi_0 + \xi_1 y_i + a z_i)}.$$

Three average missingness fractions of 20%, 40%, and 60% were used. We fit the linear model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ assuming (2.29). The rejection rates of CM$_3$ were

$0.046, 0.16, 0.262, 0.422$ for $a = 0.0, 1.0, 1.5, 2.0$, respectively, for the 60% missingness fraction.

2.4.3 Goodness of fit statistics for a logistic model

We considered a logistic model. The first group of simulation studies was to compare the finite sample performance of CM_1 using either $\mathrm{CR}_i^{(1)}$ or $\mathrm{CR}_i^{(2)}$ under the two similar scenarios as in the previous section. In the first scenario, we simulated 500 data sets from

$$p(y_i = 1) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2)},$$

for $i = 1, \cdots, 200$, where the c was in the range [0, 1]. We set $\beta_0 = \beta_1 = \beta_2 = 1$. We considered two missing data mechanisms: MCAR and MAR. For MAR, the missing data mechanism was given by (2.28), in which we set $\xi_1 = 1.0$ and ξ_0 with values $-1.5,\ -0.5,\ 0.5$ to obtain average missing data fractions of 20%, 40%, 60%, respectively. Then we fit

$$p(y_i = 1) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 z_i)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 z_i)},$$

assuming an MAR mechanism, and thus the fitted model would be misspecified if $c \neq 0$ and the misspecification is due to x_i.

The results shown in the first half of Table 2.5 are similar to those from the linear model. The Type I error rates of CM_1 based on both $\mathrm{CR}_i^{(1)}$ and $\mathrm{CR}_i^{(2)}$ are accurate across all missingness fractions. The CM_1 based on $\mathrm{CR}_i^{(2)}$ is uniformly more powerful than that based on $\mathrm{CR}_i^{(1)}$. The power for detecting misspecification of the model increased with $|c|$ for CM_1. The missing data fraction slightly influences the power of detecting model misspecification for CM_1.

In the second scenario, we generated 500 data sets from

$$p(y_i = 1) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2)},$$

Table 2.5: Rejection rates for CM_1 using either $\text{CR}_i^{(1)}$ or $\text{CR}_i^{(2)}$ at the 5% significance level for a logistic regression model. The first half shows the results for the first scenario where the misspecification is due to x_i, and the second half of the table is for the second scenario where the misspecification is due to z_i.

$\text{logit}(p(y_i = 1)) = \beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2$												
	MCAR						MAR					
	20%		40%		60%		20%		40%		60%	
c	$\text{CR}_i^{(1)}$	$\text{CR}_i^{(2)}$	$\text{CR}_i^{(1)}$	$\text{CR}_i^{(2)}$	$\text{CR}_i^{(1)}$	$\text{CR}_i^{(2)}$	$\text{CR}_i^{(1)}$	$\text{CR}_i^{(2)}$	$\text{CR}_i^{(1)}$	$\text{CR}_i^{(2)}$	$\text{CR}_i^{(1)}$	$\text{CR}_i^{(2)}$
0	0.062	0.056	0.060	0.054	0.060	0.052	0.058	0.056	0.064	0.058	0.062	0.058
0.2	0.282	0.346	0.162	0.242	0.144	0.208	0.278	0.324	0.212	0.316	0.200	0.282
0.4	0.592	0.662	0.486	0.520	0.448	0.480	0.584	0.632	0.560	0.618	0.546	0.602
0.6	0.804	0.856	0.896	0.892	0.876	0.884	0.818	0.826	0.822	0.834	0.804	0.826
0.8	0.952	0.980	0.922	0.932	0.896	0.906	0.960	0.986	0.920	0.944	0.906	0.914
1.0	1.000	1.000	0.940	0.964	0.906	0.942	1.000	1.000	0.938	0.958	0.900	0.936

$\text{logit}(p(y_i = 1)) = \beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2$												
	MCAR						MAR					
	20%		40%		60%		20%		40%		60%	
c	$\text{CR}_i^{(1)}$	$\text{CR}_i^{(2)}$	$\text{CR}_i^{(1)}$	$\text{CR}_i^{(2)}$	$\text{CR}_i^{(1)}$	$\text{CR}_i^{(2)}$	$\text{CR}_i^{(1)}$	$\text{CR}_i^{(2)}$	$\text{CR}_i^{(1)}$	$\text{CR}_i^{(2)}$	$\text{CR}_i^{(1)}$	$\text{CR}_i^{(2)}$
0	0.034	0.052	0.036	0.050	0.046	0.054	0.056	0.054	0.054	0.056	0.038	0.062
0.2	0.042	0.054	0.054	0.050	0.058	0.054	0.050	0.058	0.062	0.048	0.044	0.062
0.4	0.050	0.060	0.052	0.054	0.052	0.058	0.054	0.054	0.046	0.058	0.052	0.054
0.6	0.048	0.054	0.048	0.056	0.044	0.068	0.032	0.058	0.048	0.060	0.058	0.066
0.8	0.056	0.058	0.062	0.054	0.058	0.054	0.058	0.054	0.044	0.064	0.062	0.058
1.0	0.062	0.054	0.068	0.068	0.062	0.066	0.058	0.060	0.064	0.062	0.068	0.068

whereas the rest of the setup remained the same as in the first scenario. We fit the model ignoring the term cz_i^2, and thus the model would be misspecified if $c \neq 0$ and the misspecification is due to z_i. The results are shown in the second half of Table 2.5. Similar to the linear model, CM_1 based on both $\mathrm{CR}_i^{(1)}$ and $\mathrm{CR}_i^{(2)}$ cannot detect the misspecification of cz_i^2, because CM_1 did not incorporate the missing covariate z_i.

Similarly to the linear model, we assessed the finite sample performance of CM_2 using the same two scenarios. The first half of Table 2.6 shows the results for the first scenario where the misspecification is due to x_i, and the second half of the table for the second scenario where the misspecification is due to z_i. The Type I errors rates of CM_2 are accurate across all missingness fractions. And for both scenarios, the power for detecting misspecification of the model increased with $|c|$ for CM_2 and the missing data fraction influences the power in detecting model misspecification (i.e., $|c| \neq 0$). Compared to Table 2.5, for the first scenario, CM_1 based on $\mathrm{CR}_i^{(2)}$ is slightly more powerful than CM_2 in detecting the presence of cx_i^2. However, for the scecond scenario, CM_2 is much more powerful than CM_1 based on $\mathrm{CR}_i^{(2)}$. This indicates that incorporating the missing data can increase the power of detecting model misspecification due to cz_i^2.

2.5 Liver cancer data

To illustrate our proposed methods, we considered data on 191 patients from two Eastern Cooperative Oncology Group clinical trials as mentioned in Section 2.1 (Ibrahim, Chen, and Lipsitz, 1999). We are interested in how the number of cancerous liver nodes (y) when entering the trials is predicted by six other baseline characteristics: time since diagnosis of the disease (in weeks) (z_1); two biochemical markers (each classified as normal or abnormal), alpha fetoprotein (z_2) and anti-hepatitis antigen (z_3); associated jaundice (yes, no) (x_1); body mass index (weight in kilograms divided by the square of height in meters) (x_2); and age (in years) (x_3).

We used a Poisson regression model, thus $p(y_i|\mathbf{x}_i, \mathbf{z}_i, \beta) \propto \exp[y_i(\mathbf{v}_i'\beta) - \exp(\mathbf{v}_i'\beta)]$

Table 2.6: Rejection rates for CM_2 at the 5% significance level for a logistic regression model. The first half shows the results for the first scenario where the misspecification is due to x_i, and the second half of the table is for the second scenario where the misspecification is due to z_i.

$\text{logit}(p(y_i = 1)) = \beta_0 + \beta_1 x_i + \beta_2 z_i + c x_i^2$						
	MCAR			MAR		
c	20%	40%	60%	20%	40%	60%
0	0.052	0.044	0.042	0.056	0.056	0.058
0.2	0.200	0.164	0.166	0.224	0.216	0.204
0.4	0.480	0.322	0.306	0.446	0.336	0.322
0.6	0.736	0.604	0.584	0.740	0.628	0.592
0.8	0.800	0.774	0.764	0.856	0.776	0.778
1.0	0.912	0.884	0.876	0.910	0.876	0.870

$\text{logit}(p(y_i = 1)) = \beta_0 + \beta_1 x_i + \beta_2 z_i + c z_i^2$						
	MCAR			MAR		
c	20%	40%	60%	20%	40%	60%
0	0.054	0.056	0.058	0.058	0.054	0.058
0.2	0.172	0.066	0.056	0.180	0.066	0.068
0.4	0.324	0.222	0.090	0.326	0.210	0.092
0.6	0.562	0.306	0.118	0.540	0.288	0.144
0.8	0.642	0.442	0.176	0.666	0.454	0.174
1.0	0.884	0.508	0.200	0.856	0.544	0.242

where $\mathbf{v}_i^{'} = (1, x_{i1}, x_{i2}, x_{i3}, z_{i1}, z_{i2}, z_{i3})$ is the 1×7 vector of covariates including an intercept, and $\beta = (\beta_0, \beta_1, \cdots, \beta_6)^{'}$ are the corresponding regression coefficients. Logarithm of the time since diagnosis was used to achieve approximate normality. Since only $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3})$ have missing values, we need to consider a joint distribution only for these covariates given $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$. Since z_{i2} and z_{i3} were both dichotomous, it was reasonable to model their conditional univariate distributions by means of logistic regressions. Thus

$$p(z_{i1}, z_{i2}, z_{i3}|\mathbf{x}_i, \alpha) = p(z_{i3}|z_{i1}, z_{i2}, \mathbf{x}_i, \alpha_3)p(z_{i2}|z_{i1}, \mathbf{x}_i, \alpha_2)p(z_{i1}|\mathbf{x}_i, \alpha_1),$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ and $(z_{i3}|z_{i1}, z_{i2}, \mathbf{x}_i)$ is a logistic regression with probability of success

$$p(z_{i3} = 1|z_{i1}, z_{i2}, \mathbf{x}_i, \alpha_3) = \frac{\exp(\alpha_{30} + \alpha_{31}z_{i1} + \alpha_{32}z_{i2} + \alpha_{3x}^{'}\mathbf{x}_i)}{1 + \exp(\alpha_{30} + \alpha_{31}z_{i1} + \alpha_{32}z_{i2} + \alpha_{3x}^{'}\mathbf{x}_i)},$$

and $\alpha_{3x}^{'} = (\alpha_{33}, \alpha_{34}, \alpha_{35})$. Similarly,

$$p(z_{i2} = 1|z_{i1}, \mathbf{x}_i, \alpha_2) = \frac{\exp(\alpha_{20} + \alpha_{21}z_{i1} + \alpha_{2x}^{'}\mathbf{x}_i)}{1 + \exp(\alpha_{20} + \alpha_{21}z_{i1} + \alpha_{2x}^{'}\mathbf{x}_i)},$$

and $\alpha_{2x}^{'} = (\alpha_{22}, \alpha_{23}, \alpha_{24})$. In addition, we took a normal distribution for the missing covariate z_1, specifically, $z_{i1} \sim N(\alpha_{11}, \alpha_{12}), i = 1, \cdots n$ and $\alpha_1^{'} = (\alpha_{11}, \alpha_{12})$.

We assumed the missing covariates are MAR and calculated the maximum likelihood estimate of (β, α) using the EM algorithm. The case-deletion diagnostic measures CD_i identified cases 10, 15, 65, 131, and 160 as influential, among which $\text{CD}_{i,1}$ identified cases 10, 15, 65, and 160, whereas $\text{CD}_{i,2}$ identified case 131 (Figure 2.2 a, b, c). These findings confirmed the suspected cases reported in Table 2.1. The QD_i, $\text{QD}_{i,1}$, and $\text{QD}_{i,2}$ gave similar results (not presented here). The standardized conditional residuals, $\text{SCR}^{(1)}$, detected cases 10, 15, 65, and 160 as influential observations (Figure 2.2 d) and $\text{SCR}^{(2)}$

gave similar results (not presented).

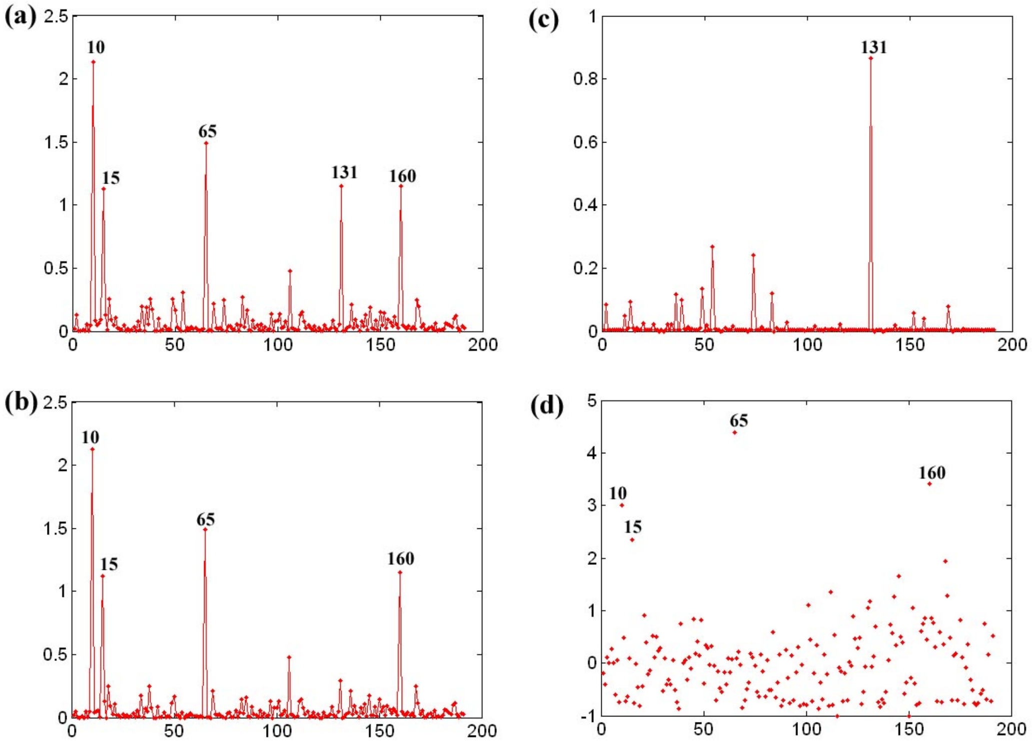

Figure 2.2: Liver cancer data: index plots of diagnostic measures: (a) CD_i, (b) $\mathrm{CD}_{i,1}$, (c) $\mathrm{CD}_{i,2}$, (d) $\mathrm{SCR}_i^{(1)}$.

The $p-$values of the goodness of fit test using CM_1 based on $\mathrm{CR}_i^{(1)}$ and $\mathrm{CR}_i^{(2)}$ were 0.56 and 0.48 respectively, whereas the $p-$value of the goodness of fit test using CM_2 was 0.06. These indicated that either $E(y_i|\mathbf{v}_i) \neq \exp(\mathbf{v}_i'\beta)$ or the missing data mechanism depended on the response variable. So we considered the following MAR mechanism,

$$p(\mathbf{r}_i|\mathbf{x}_i, y_i) = p(r_{i3}|r_{i1}, r_{i2}, \mathbf{x}_i, y_i, \xi_2)p(r_{i2}|r_{i1}, \mathbf{x}_i, y_i, \xi_2)p(r_{i1}|\mathbf{x}_i, y_i, \xi_1),$$

where $p(r_{i2}|r_{i1}, \mathbf{x}_i, y_i, \xi_2)$ and $p(r_{i1}|\mathbf{x}_i, y_i, \xi_1)$ are

$$p(r_{i1}|y_i, \mathbf{x}_i, \xi_1) = \frac{\exp(f_{i1})}{1+\exp(f_{i1})},$$

$$p(r_{i2}|r_{i1}, y_i, \mathbf{x}_i, \xi_2) = \frac{\exp(f_{i2})}{1+\exp(f_{i2})},$$

$$p(r_{i3}|r_{i1}, r_{i2}, y_i, \mathbf{x}_i, \xi_3) = \frac{\exp(f_{i3})}{1 + \exp(f_{i3})},$$

in which $f_{i1} = \xi_{10} + \xi_{11}x_{i1} + \xi_{12}x_{i2} + \xi_{13}x_{i3} + \xi_{14}y_i$, $f_{i2} = \xi_{20} + \xi_{21}x_{i1} + \xi_{22}x_{i2} + \xi_{23}x_{i3} + \xi_{24}y_i + \xi_{25}r_{i1}$, and $f_{i3} = \xi_{30} + \xi_{31}x_{i1} + \xi_{32}x_{i2} + \xi_{33}x_{i3} + \xi_{34}y_i + \xi_{35}r_{i1} + \xi_{36}r_{i2}$.

We found that the missing data mechanism of z_{i1} depended on the response variable, so we should use CM$_3$ for the goodness of fit test. The goodness of fit test using CM$_3$ was not significant ($p-$value $= 0.56$), indicating that the model fit well.

2.6 Discussion

We have derived goodness of fit statistics in the presence of missing data based on novel definitions of case-deletion and residual diagnostics. The asymptotic properties of the goodness of fit measures based on conditional residuals were also derived, as well as MCMC algorithms for carrying out the EM algorithm. The simulation studies and liver cancer dataset showed very promising results for the proposed methods. Future work in this area includes extending the methodologies to the Cox proportional hazards model with right censored survival data and missing covariates, as well as to parametric and semiparametric models for longitudinal data with MAR or NMAR response and/or covariate data.

We also note several limitations of our proposed tests. The first limitation is that we assume parametric distributions throughout the paper, whereas the goodness of fit tests focus on testing the regression function. It is very interesting to extend the definitions of conditional residuals and associated test statistics to semiparametric models. In addition, our preliminary results have shown that the misspecified distributions can have profound effects on the finite sample performance of our proposed test statistics. The second limitation is that it is difficult to pinpoint the cause of the rejection in case of rejection and subsequently to suggest an alternative model. This limitation is inherent in all omnibus tests based on integrated regressions (Stute, 1997). All these issues merit further research, and we will study them in our future work.

2.7 Appendix

The following assumptions are needed to facilitate development of our methods, although they may not be the weakest possible conditions.

(C1) η_* is unique and an interior point of Υ, where Υ is a compact set in $R^{\text{dim}(\eta)}$.

(C2) $\hat{\eta} \to \eta_*$ in probability as $n \to \infty$.

(C3) For each i, $\ell(\mathbf{d}_i; \eta) = \log p(\mathbf{d}_i; \eta)$ is three-times continuously differentiable on Υ and $|\partial_j \ell(\mathbf{d}_i; \eta)|^2$ and $|\partial_j \partial_k \ell(\mathbf{d}_i; \eta)|$ are dominated by an integrable function $B_i(\mathbf{d}_i)$ for all $j, k = 1, \cdots, d$, where $\partial_j = \partial / \partial \eta_j$.

(C4) For each $\epsilon > 0$, there exists a finite K such that

$$\sup_{n \geq 1} n^{-1} \sum_{i=1}^{n} E[B_i(\mathbf{d}_i)^2 \mathbf{1}\{B_i(\mathbf{d}_i) > K\}] < \epsilon$$

for all n, where $\mathbf{1}\{B_i(\mathbf{d}_i) > K\}$ is the indicator function of $B_i(\mathbf{d}_i) > K$.

(C5) $\lim_{n\to\infty} n^{-1}\{-\sum_{i=1}^{n} \partial_\eta^2 \ell(\mathbf{d}_{o,i}; \eta_*)\} = A(\eta_*)$ and

$$\lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} \{\partial_\eta \ell(\mathbf{d}_{o,i}; \eta_*) \partial_\eta \ell(\mathbf{d}_{o,i}; \eta_*)'\} = B(\eta_*),$$

where $A(\eta_*)$ is nonsingular and $B(\eta_*)$ is positive definite.

(C6) Let $\rho((\varphi, t), (\varphi_*, t_*))$ be the limit of $\rho_n((\varphi_n, t_n), (\varphi_{*n}, t_{*n}))$, where

$$\rho_n((\varphi_n, t_n), (\varphi_{*n}, t_{*n})) = \left(n^{-1} \sum_{i=1}^{n} E[|\text{CR}_i(\eta_*)|^2 |\mathbf{1}(\varphi_n' \mathbf{x} \leq t_n) - \mathbf{1}(\varphi_{*n}' \mathbf{x} \leq t_{*n})|^2 \right)^{1/2}.$$

For any sequences $\{(\varphi_n, t_n)\}$ and $\{(\varphi_{*n}, t_{*n})\}$, $\rho_n((\varphi_n, t_n), (\varphi_{*n}, t_{*n}))$ converges to zero when $\rho((\varphi_n, t_n), (\varphi_{*n}, t_{*n})) \to 0$ as $n \to \infty$. A similar condition also holds for $I_3((\tilde{\varphi}, t); \eta_*)$.

(C7) $\Delta_1(\varphi, t)$ and $F_\varphi(dt)d\varphi$ are absolutely continuous with respect to Lebesgue measure on Π.

(C8) For any small $a_0 > 0$, we assume that

$$\sup_{(\alpha,\varphi,t)\in\mathcal{A}\times\Pi} \mathrm{P}\left\{-\delta < [\mathbf{c}_i(\alpha)'\varphi - t]/V_i < \delta\right\} \leq C_0\delta^{c_1},$$

where C_0 and c_1 are two positive scalars, $\mathcal{A} = \{\alpha : ||\alpha - \alpha_*||_2 \leq a_0\}$, and $\sup_{\alpha\in\mathcal{A}} ||\partial_\alpha[\mathbf{c}_i(\alpha)]||_2^2 + \sup_{\alpha\in\mathcal{A}} ||\mathbf{c}_i(\alpha)||_2^2 + 1 = V_i(\mathbf{x}_i, \mathbf{z}_{o,i})^2$.

Comments. Condition (C1) is a standard identifiability condition. Some sufficient conditions for Condition (C2) have been widely presented in the literature; see van der Vaart and Wellner (1996) and Andrews (1999). Conditions (C3)-(C5) are required to ensure the asymptotic normality of $\hat{\eta}$. Conditions (C6) is required to invoke the central limit theory for the sums of independent but not identically distributed stochastic processes (van der Vaart and Wellner, 1996; Kosorok, 2007). Condition (C7) is required to ensure the asymptotic distributions of the Cramer-von Mises test statistics. (C8) are required to invoke Ossiander's entropy conditions (Ossiander, 1987; Andrews, 1994).

Proof of Proposition 1. For ease of exposition, we omit η_* in some notation, such as $p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i)$.

(i) For brevity, we only consider $\mathrm{CR}_i^{(1)}(\eta_*)$. It can be shown that

$$E[E[g((\mathbf{x}_i', \mathbf{z}_i')\beta)|\mathbf{x}_i, \mathbf{z}_{o,i}]|\mathbf{x}_i] = E[g((\mathbf{x}_i', \mathbf{z}_i')\beta)|\mathbf{x}_i],$$

$$E[y_i|\mathbf{x}_i] = \int y_i p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{z}_i|\mathbf{x}_i)p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i)dy_i d\mathbf{z}_i d\mathbf{r}_i = \int g((\mathbf{x}_i', \mathbf{z}_i')\beta_*)p(\mathbf{z}_i|\mathbf{x}_i)d\mathbf{z}_i,$$

which yields $E[\mathrm{CR}_i^{(1)}(\eta_*)|\mathbf{x}_i] = 0$. Furthermore, $E[\mathrm{CR}_i^{(1)}(\eta_*)] = E\{E[\mathrm{CR}_i^{(1)}(\eta_*)|\mathbf{x}_i]\} = 0$. However, it can be shown that

$$E[y_i|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i] = \frac{\int y_i p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i)d\mathbf{z}_{m,i}dy_i}{\int p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i)d\mathbf{z}_{m,i}dy_i} \neq E[g((\mathbf{x}_i', \mathbf{z}_i')\beta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i].$$

(ii) For MAR covariates, we have $p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i) = p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_{o,i}, y_i)$. It can be shown that

$$E[g((\mathbf{x}_i', \mathbf{z}_i')\beta)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i] = \frac{\int g((\mathbf{x}_i', \mathbf{z}_i')\beta)p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_{o,i})d\mathbf{z}_{m,i}}{\int p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_{o,i})d\mathbf{z}_{m,i}}$$
$$= \frac{\int g((\mathbf{x}_i', \mathbf{z}_i')\beta)p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)d\mathbf{z}_{m,i}}{\int p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)d\mathbf{z}_{m,i}} = E[g((\mathbf{x}_i', \mathbf{z}_i')\beta)|\mathbf{x}_i, \mathbf{z}_{o,i}, y_i].$$

Thus, we have

$$\mathrm{CR}_i^{(2)}(\eta) = y_i - E[g((\mathbf{x}_i', \mathbf{z}_i')\beta)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i, y_i] = y_i - E[g((\mathbf{x}_i', \mathbf{z}_i')\beta)|\mathbf{x}_i, \mathbf{z}_{o,i}, y_i].$$

Furthermore, it can be shown that

$$E\left[\frac{\mathrm{CR}_i^{(1)}}{p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_{o,i}, y_i)}\middle| \mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i\right] = \frac{\int \mathrm{CR}_i^{(1)} p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)d\mathbf{z}_{m,i}dy_i}{\int p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_{o,i})d\mathbf{z}_{m,i}dy_i} = 0.$$

(iii) Using $p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i) = p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i)$, we obtain

$$E[y_i|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i] = \frac{\int y_i p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i)d\mathbf{z}_{m,i}dy_i}{\int p(y_i|\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i)d\mathbf{z}_{m,i}dy_i}$$
$$= \frac{\int g((\mathbf{x}_i', \mathbf{z}_i')\beta_*)p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i)d\mathbf{z}_{m,i}}{\int p(\mathbf{x}_i, \mathbf{z}_i)p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i)d\mathbf{z}_{m,i}} = E[g((\mathbf{x}_i', \mathbf{z}_i')\beta_*)|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i].$$

Thus, $E[\mathrm{CR}_i^{(2)}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i] = 0$ and $E[\mathrm{CR}_i^{(1)}|\mathbf{x}_i, \mathbf{z}_{o,i}, \mathbf{r}_i] \neq 0$.

(iv) and (v) Using first-order Taylor's series expansions yields the desired results.

Proof of Theorem 1. (i) Conditions (C1)-(C5) are sufficient for establishing (i) (Andrews, 1994; van der Vaart and Wellner, 1996).

(ii) First, we can prove weak convergence of $I_1(\cdot;\eta_*)$ using a standard argument of empirical process theory. The finite-dimensional marginals of $I_1(\cdot;\eta_*)$ converge weakly to the corresponding marginals of the zero-mean Gaussian process $G_1(\cdot)$. This can be proved by using Assumptions C3 and C4. Because $\mathcal{F} = \{f(\varphi, t) = \mathrm{CR}(\eta_*)\mathbf{1}(\varphi'\mathbf{x} \leq t) : (\varphi, t) \in \Pi\}$ is a VC (Vapnik and Cervonenkis) class, which satisfies the universal entropy condition (van der Vaart and Wellner, 1996; Sections 2.5 and 2.6), the tightness of $I_1(\cdot;\eta_*)$ follows from the Donsker Theorem (van der Vaart and Wellner, 1996; Section 2.11). Second, the convergence of $\sqrt{n}(\hat{\eta}-\eta_*)$ follows from the standard Lindeberg-Feller theorem. Third, we can prove the joint convergence of $I_1(\cdot;\eta_*)$ and $\sqrt{n}(\hat{\eta}-\eta_*)$ using the Cramer-Wold device and empirical process theory.

(iii) It can be shown from a Taylor's series expansion that

$$\begin{aligned} I_1((\varphi,t);\hat{\eta}) &= I_1((\varphi,t);\eta_*) + n^{1/2}(\hat{\eta}-\eta_*)n^{-1}\sum_{i=1}^{n}\partial_\eta[\mathrm{CR}_i(\eta_*)]\mathbf{1}(\varphi^{'}\mathbf{x}_i \le t) \quad (2.30) \\ &+ n^{1/2}(\hat{\eta}-\eta_*)n^{-1}\sum_{i=1}^{n}\{\partial_\eta[\mathrm{CR}_i(\tilde{\eta})] - \partial_\eta[\mathrm{CR}_i(\eta_*)]\}\mathbf{1}(\varphi^{'}\mathbf{x}_i \le t), \end{aligned}$$

where $||\tilde{\eta}-\eta_*|| \le ||\hat{\eta}-\eta_*|| \to 0$. It follows from the law of large numbers and Assumptions C3 and C4 that $n^{-1}\sum_{i=1}^{n}\{\partial_\eta[\mathrm{CR}_i(\tilde{\eta})] - \partial_\eta[\mathrm{CR}_i(\eta_*)]\}\mathbf{1}(\varphi^{'}\mathbf{x}_i \le t)$ converges to zero uniformly in (φ,t) in probability (van der Vaart and Wellner, 1996). Similarly, $n^{-1}\sum_{i=1}^{n}\partial_\eta[\mathrm{CR}_i(\eta_*)]\mathbf{1}(\varphi^{'}\mathbf{x}_i \le t)$ converges to $\Delta_1(\varphi,t)$ uniformly in (φ,t) in probability. Because $n^{1/2}(\hat{\eta}-\eta_*)$ is asymptotically normal and $\Delta_1(\varphi,t)$ is uniformly continuous, the second term of (2.30) on the right hand side is asymptotically tight. Since we have already established weak convergence of $I_1((\varphi,t);\eta_*)$, we can use a standard argument of the empirical process theory to establish that $I_1(\cdot;\hat{\eta})$ converges weakly to $G_1(\cdot)+\Delta_1(\cdot)'\nu_1$ as $n\to\infty$. Applying the continuous mapping theorem ensures that CK_1 converges in distribution to $\sup_{(\varphi,t)}|G_1(\varphi,t)+\Delta_1(\varphi,t)'\nu_1|$. To prove weak convergence of CM_1, we use Proposition 7.27 of Kosorok (2007) to prove that $CM_1 = \int_\Pi |I_1((\varphi,t);\hat{\eta})|^2 F_{n,\varphi}(dt)d\varphi$ converges weakly to $\int_\Pi |G_1(\varphi,t)+\Delta_1(\varphi,t)'\nu_1|^2 F_\varphi(dt)d\varphi$ as $n\to\infty$, since $I_1((\varphi,t);\hat{\eta})$ converges weakly to $G_1(\varphi,t)+\Delta_1(\varphi,t)'\nu_1 \in \ell^\infty(\Pi)$ and $\sup_{t\in[-\infty,\infty]}|F_{n,\varphi}(t)-F_\varphi(t)|\to 0$ as $n\to\infty$.

Proof of Theorem 2. (i) We define $\ell_n(t)=\log p(D_o;t)$, where

$$p(D_o;t)=\prod_{i=1}^{n}\int p(y_i|\mathbf{x}_i,\mathbf{z}_i,t)p(\mathbf{x}_i,\mathbf{z}_i)p(\mathbf{r}_i|\mathbf{x}_i,\mathbf{z}_i,y_i)d\mathbf{z}_{m,i}.$$

The true density function of $\mathbf{d}_{o,i}$ under local alternatives equals $p(D_o; n^{-1/2})$. Using a Taylor's series expansion, we get

$$\ell_n(n^{-1/2}) = \ell_n(0) + n^{-1/2}\partial_t \ell_n(0) + 0.5n^{-1}\partial_t^2 \ell_n(0) + o_p(1),$$

where $\partial_t = d/dt$ and $\partial_t^2 = d^2/dt^2$. In particular, we have

$$\partial_t \ell_n(0) = \sum_{i=1}^n E[a_i(\tau_*)^{-1}\partial_t \theta_i(0)(y_i - \mu_i)|\mathbf{d}_{o,i}, t = 0],$$

where the conditional expectation is taken with respect to $\mathbf{z}_{m,i}$ given $\mathbf{d}_{o,i}$ under $t = 0$. Under $p(D_o; t = 0)$, $(\sqrt{n}(\hat{\eta} - \eta_*), \ell_n(n^{-1/2}) - \ell_n(0))$ can be approximated by

$$(n^{-1/2}\sum_{i=1}^n \psi_{n,i}, \partial_t \ell_n(0) n^{-1/2}) + (0, -0.5n^{-1}[-\partial_t^2 \ell_n(0)]) + o_p(1).$$

Following the arguments in Example 12.3.8 of Lehmann and Romano (2006), we can show that under local alternative hypotheses, $\sqrt{n}(\hat{\eta} - \eta_*)$ converges in distribution to $A_1 + \nu_1$.

(ii) For simplicity, we only consider $\mathrm{CR}_i^{(1)}$. The process $I_1((\varphi, t); \eta_*)$ can be represented as

$$\begin{aligned} I_1((\varphi, t); \eta_*) &= n^{-1/2}\sum_{i=1}^n \mathbf{1}(\varphi'\mathbf{x}_i \le t)(y_i - E[g((\mathbf{x}_i', \mathbf{z}_i')\beta_*) + n^{-1/2} g_0(\mathbf{x}_i, \mathbf{z}_i)|\mathbf{x}_i, \mathbf{z}_{o,i}]) \\ &+ n^{-1}\sum_{i=1}^n \mathbf{1}(\varphi'\mathbf{x}_i \le t) E[g_0(\mathbf{x}_i, \mathbf{z}_i)|\mathbf{x}_i, \mathbf{z}_{o,i}], \end{aligned}$$

in which the first term on the right side converges weakly to $G_1(\cdot)$ by using similar arguments as in Theorem 1 (ii). In addition, it follows from the law of large number that $n^{-1}\sum_{i=1}^n \mathbf{1}(\varphi'\mathbf{x}_i \le t) E[g_0(\mathbf{x}_i, \mathbf{z}_i)|\mathbf{x}_i, \mathbf{z}_{o,i}]$ converges to $A_2(\varphi, t)$ uniformly in probability.

(iii) Following similar arguments as in Theorem 1 (iii), we use a Taylor's series expansion to show that

$$\begin{aligned} I_1((\varphi,t);\hat{\eta}) = \; & I_1((\varphi,t);\eta_*) + n^{1/2}(\hat{\eta}-\eta_*)n^{-1}\textstyle\sum_{i=1}^n \partial_\eta[\mathrm{CR}_i(\eta_*)]\mathbf{1}(\varphi'\mathbf{x}_i \leq t) \\ & + n^{1/2}(\hat{\eta}-\eta_*)n^{-1}\textstyle\sum_{i=1}^n \{\partial_\eta[\mathrm{CR}_i(\tilde{\eta})] - \partial_\eta[\mathrm{CR}_i(\eta_*)]\}\mathbf{1}(\varphi'\mathbf{x}_i \leq t), \end{aligned}$$

where $||\tilde{\eta}-\eta_*||_2 \leq ||\hat{\eta}-\eta_*||_2$. Similar to the arguments in Theorem 1 (iii), we can use standard arguments of empirical processes and the results in Theorem 2 (i) and (ii) to complete the proof of (iii).

Proof of Theorem 3. The proof of Theorem 3 consists of two steps as follows. In Step 1, we need to prove that $I_2((\tilde{\varphi},t);\eta_*)$ can be represented as

$$n^{-1/2}\sum_{i=1}^n \mathbf{1}(\mathbf{c}_{i,*}'\tilde{\varphi} \leq t)\mathrm{CR}_i^{(2)}(\eta_*) + n^{-1/2}\sum_{i=1}^n [\mathbf{1}(\hat{\mathbf{c}}_i'\tilde{\varphi} \leq t) - \mathbf{1}(\mathbf{c}_{i,*}'\tilde{\varphi} \leq t)]\mathrm{CR}_i^{(2)}(\eta_*), \quad (2.31)$$

where $\mathbf{c}_{i,*} = \mathbf{c}_i(\alpha_*)$. We first show that the second term of (2.31) converges to zero uniformly in probability and a sufficient condition is that $\{n^{-1/2}\sum_{i=1}^m \mathbf{1}(\mathbf{c}_i(\alpha)'\tilde{\varphi} \leq t)\mathrm{CR}_i^{(2)}(\eta_*) : \kappa = (\alpha,\tilde{\varphi},t) \in \mathcal{A}\times\Pi\}$ is stochastically equicontinuous, where $\mathcal{A} = \{\alpha : ||\alpha-\alpha_*||_2 \leq a_0\}$ for a sufficient small $a_0 > 0$. We invoke Ossiander's entropy condition to show that $\mathcal{M} = \{\mathbf{1}(\mathbf{c}(\alpha)'\tilde{\varphi} \leq t)\mathrm{CR}^{(2)}(\eta_*) : \kappa = (\alpha,\tilde{\varphi},t) \in \mathcal{A}\times\Pi\}$ is a type IV class (Ossiander, 1987; Andrews, 1994). We need to check the following condition:

$$\sup_i E\{[\mathrm{CR}_i^{(2)}(\eta_*)]^2 \sup_{\kappa_1:||\kappa_1-\kappa||_2<\delta} |\mathbf{1}(\mathbf{c}_i(\alpha_1)'\tilde{\varphi}_1 \leq t_1) - \mathbf{1}(\mathbf{c}_i(\alpha)'\tilde{\varphi} \leq t)|^2\} \leq C\delta^{c_1}, \quad (2.32)$$

where $\kappa_1 = (\alpha_1,\tilde{\varphi}_1,t_1)$ and C and c_1 are some finite positive constants. The left-hand side of (2.32) can be bounded above by

$$
\begin{aligned}
& \sup_i E\{E[[\mathrm{CR}_i^{(2)}(\eta_*)]^2|\mathbf{d}_{o,i}] \sup_{\kappa_1:||\kappa_1-\kappa||_2<\delta} |\mathbf{1}(\mathbf{c}_i(\alpha_1)'\tilde{\varphi}_1 \leq t_1) - \mathbf{1}(\mathbf{c}_i(\alpha)'\tilde{\varphi} \leq t)|\} \\
= & \sup_i E\{[\mathrm{CR}_i^{(2)}(\eta_*)]^2 \sup_{\kappa_1:||\kappa_1-\kappa||_2<\delta} |\mathbf{1}(\mathbf{c}_i(\alpha_1)'\tilde{\varphi}_1 \leq t_1) - \mathbf{1}(\mathbf{c}_i(\alpha)'\tilde{\varphi} \leq t)|\} \\
\leq & \sup_i E\{[\mathrm{CR}_i^{(2)}(\eta_*)]^4\}^{1/2} \sup_i E[\sup_{\kappa_1:||\kappa_1-\kappa||_2<\delta} |\mathbf{1}(\mathbf{c}_i(\alpha_1)'\tilde{\varphi}_1 \leq t_1) - \mathbf{1}(\mathbf{c}_i(\alpha)'\tilde{\varphi} \leq t)|]^{1/2},
\end{aligned}
$$

in which we have used the Cauchy-Schwartz inequality twice and $|\mathbf{1}(S_1) - \mathbf{1}(S_2)|^2 = |\mathbf{1}(S_1) - \mathbf{1}(S_2)|$ for any two sets S_1 and S_2. Since $E[\mathrm{CR}_i^{(2)}(\eta_*)]^4 = E[\mathrm{CR}_1^{(2)}(\eta_*)]^4$ for all i, it follows from Condition (C2) that $\sup_i E[\mathrm{CR}_i^{(2)}(\eta_*)]^4 = E[\mathrm{CR}_1^{(2)}(\eta_*)]^4 < \infty$. Let $h_i(\kappa) = \mathbf{c}_i(\alpha)'\varphi - t$. It follows from a Taylor's series expansion that $h_i(\kappa) = h_i(\kappa_1) + \partial_\kappa h_i(\tilde{\kappa})'(\kappa - \kappa_1)$, where $||\tilde{\kappa} - \kappa_1||_2 \leq ||\kappa - \kappa_1||_2$. Thus, we have $|h_i(\kappa) - h_i(\kappa_1)| \leq ||\partial_\kappa h_i(\tilde{\kappa})||_2 ||\kappa - \kappa_1||_2$, where $\partial_\kappa h_i(\tilde{\kappa}) = (\partial_\alpha[\mathbf{c}_i(\alpha)'\varphi]', \mathbf{c}_i(\alpha)', 1)'$. Then, we have $||\partial_\kappa h_i(\tilde{\kappa})||_2^2 \leq ||\partial_\alpha[\mathbf{c}_i(\alpha)]||_2^2 + ||\mathbf{c}_i(\alpha)||_2^2 + 1 \leq V_i$. Using $|\mathbf{1}(S_1) - \mathbf{1}(S_2)| \leq \mathbf{1}(S_1 \cap S_2^c) + \mathbf{1}(S_2 \cap S_1^c)$ and Condition (C7), we can further show that

$$
\begin{aligned}
& E[\sup_{\kappa_1:||\kappa_1-\kappa||_2<\delta} |\mathbf{1}(h_i(\kappa) \leq h_i(\kappa) - h_i(\kappa_1)) - \mathbf{1}(h_i(\kappa) \leq 0)|] \\
& \leq E[\mathbf{1}(-\sqrt{V_i}\delta \leq h_i(\kappa) \leq \sqrt{V_i}\delta)] \leq C_0 \delta^{c_1}.
\end{aligned}
$$

In Step 2, we follow the arguments of Theorem 1 (ii) to prove that $n^{-1/2}\sum_{i=1}^n \mathbf{1}(\mathbf{c}_{i,*}'\tilde{\varphi} \leq t)\mathrm{CR}_i^{(2)}(\eta_*)$ converges to $G_2(\cdot)$ in distribution.

Chapter 3

Local Influence for GLMs with Missing Covariates

3.1 Introduction

Methods for handling missing data strongly depend on the mechanism that generated the missing values as well as the distributional and model assumptions at various stages. Therefore, the resulting estimates and tests may be sensitive to these assumptions. For this reason, sensitivity analyses are commonly used to check the sensitivity of the parameter estimates of interest with respect to the model assumptions. Sensitivity analyses are often carried out in two consecutive steps: selection of perturbation schemes to various model assumptions and use of influence measures to quantify the effects of those perturbations. Some literature on sensitivity analysis for missing data problems includes Copas and Li (1997), Copas and Eguchi (2005), Troxel (1998), Jansen et al. (2003), Van Steen, Molenberghs, and Thijs (2001), Verbeke et al. (2001), Hens et al. (2005), Jansen et al. (2006), and Troxel, Ma, and Heitjan (2004). For instance, Copas and Eguchi (2005) proposed a general formulation for assessing the bias of maximum likelihood estimates due to incomplete data in the presence of small model uncertainty. Verbeke et al. (2001), Hens et al. (2005), and Jansen et al. (2006) developed local influence methods for assessing nonrandom dropout in incomplete longitudinal data.

The goal of this paper is to systematically investigate Cook's (1986) local influence methods for GLMs with MAR covariates as well as NMAR covariates. Our local influence method provides a general framework for carrying out sensitivity analyses for missing data problems, compared to the existing literature (Copas and Eguchi, 2005; Van Steen, Molenberghs, and Thijs, 2001; Troxel, Ma, and Heitjan, 2004; Hens et al., 2005; Jansen et al., 2006). We examine two types of perturbation schemes for perturbing various model assumptions and individual observations. We also develop a methodology for selecting appropriate perturbation schemes. We examine two objective functions, including the maximum likelihood estimate and the likelihood ratio statistic, and then we develop influence measures based on these functions to assess appropriate perturbation schemes.

To motivate the proposed methodology, we consider a quality of life dataset and a liver cancer dataset. The quality of life study of the International Breast Cancer Study Group (IBCSG) compares several chemotherapies in premenapausal women with breast cancer. These women were randomly assigned in a 2 × 2 factorial design to receive tamoxifen either alone or with oral cyclophosphamide, intravenous methotrexate and flourouracil (CMF) in three early cycles, three delayed cycles, or both early and delayed cycles. For ease of exposition, the four treatment arms are labeled A, B, C, and D. The response variable is the logarithm of the survival time. The dataset has 404 observations and the covariates are: physical ability; mood; indicator for treatment A (yes, no); indicator for treatment B (yes, no); indicator for treatment C (yes, no); age (in years); and language (English, otherwise). Among these seven covariates, physical ability and mood have 13% and 31% missingness percentages respectively and the remaining covariates are fully observed. The liver cancer dataset has 191 patients from two Eastern Cooperative Oncology Group clinical trials (Ibrahim, Chen, and Lipsitz, 1999). Previous analyses of these data focused on characterizing how the number of cancerous liver nodes (response) when entering the trials was predicted by six other baseline characteristics: time since diagnosis of the disease (in weeks); two biochemical markers (alpha fetoprotein and anti-hepatitis antigen, each classified as normal or abnormal); associated jaundice (yes, no); body mass index (weight in kilograms divided by the square of height in meters); and age (in years). Among these six covariates, three have missing data and the remaining covariates are completely observed. The three with missing data, which are time since diagnosis of the disease, alpha fetoprotein, and anti-hepatitis antigen, have 8.9%, 5.8%, and 18.3% missingness percentages, yielding a total missingness percentage of 29%. Here, it is of interest to carry out local influence methods to possibly detect influential cases and to carry out sensitivity analyses on the model assumptions. For instance, using our new methodology, we detected that cases 10, 15, 65, and 160 in the liver cancer data have abnormally large response values, and case 131 has an extreme covariate value in time since diagnosis compared to the rest of the cases (Table 2.1). More details regarding these two real datasets are given in Section 3.5.

The paper is organized as follows. In Section 3.2, we review the model development for GLMs with missing covariates. In Section 3.3, we systematically develop local influence measures for assessing small perturbations to model assumptions in GLMs with missing covariates. We present several simulation studies in Section 3.4, and analyze two real datasets in Section 3.5. We conclude the paper with some final remarks in Section 3.6.

3.2 Model and notation

Suppose that we have the complete data $D_c = \{\mathbf{d}_i = (\mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i, y_i) : i = 1, \cdots, n\}$, where y_i is the univariate response, $\mathbf{x}_i$ is a $p_1 \times 1$ vector of completely observed covariates, and $\mathbf{z}_i$ is a $p_2 \times 1$ vector of partially observed covariates. We use $\mathbf{r}_i$, a $p_2 \times 1$ random vector, to indicate the missingness of $\mathbf{z}_i$: $r_{ik} = 1$ if z_{ik} is observed, and $r_{ik} = 0$ if z_{ik} is missing, where r_{ik} and z_{ik} are the k-th component of $\mathbf{r}_i$ and $\mathbf{z}_i$, respectively.

We use $p(D_c|\boldsymbol{\eta})$ to denote the complete-data density function with $\boldsymbol{\eta}$ being the vector of all unknown parameters. One way of modeling the complete-data density is to use three layers of conditional densities as follows:

$$p(D_c|\boldsymbol{\eta}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}, \tau)p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\alpha})p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}), \tag{3.1}$$

where $(\boldsymbol{\beta}, \tau)$ are the parameters for the conditional distribution of y_i given $(\mathbf{x}_i, \mathbf{z}_i)$, $\boldsymbol{\alpha}$ is the parameter vector for the covariate distribution $p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\alpha})$, and $\boldsymbol{\xi}$ is the parameter vector for modeling the missing data mechanism $p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi})$. The three sets of parameters are assumed distinct from one another, and $\boldsymbol{\eta} = (\boldsymbol{\beta}^{'}, \tau, \boldsymbol{\alpha}^{'}, \boldsymbol{\xi}^{'})^{'}$.

We need to specify each of the three components in (3.1). Under the GLM, y_i given $(\mathbf{x}_i, \mathbf{z}_i)$ has a density in the exponential family

$$p(y_i \mid \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}, \tau) = \exp\left\{a_i^{-1}(\tau)[y_i\theta_i(\boldsymbol{\beta}) - b(\theta_i(\boldsymbol{\beta}))] + c(y_i, \tau)\right\}, \tag{3.2}$$

$i = 1, \ldots, n$, indexed by the canonical parameter θ_i and the scale parameter τ, where the functions $b(\cdot)$ and $c(\cdot, \cdot)$ determine a particular distribution in the class. The functions $a_i(\tau)$ are commonly of the form $a_i(\tau) = \tau^{-1}k_i^{-1}$, where the k_i's are known weights. Further, the θ_i's satisfy the equations $\theta_i = \theta(\mu_i)$, and $\mu_i = g((\mathbf{x}_i', \mathbf{z}_i')\boldsymbol{\beta})$ are the components of $\mu = E(\mathbf{y} \mid \mathbf{x}, \mathbf{z}, \boldsymbol{\beta}, \tau)$, where $g(\cdot)$ is a known link function and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is a $(p+1) \times 1$ vector of regression coefficients, in which $p = p_1 + p_2$.

Next, we need to specify a distribution for $\mathbf{z}_i$ given $\mathbf{x}_i$. We suggest specifying the covariate distribution via a sequence of one-dimensional conditional distributions:

$$p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\alpha}) = p(z_{ip_2}|z_{i(p_2-1)}, \cdots, z_{i1}, \mathbf{x}_i, \boldsymbol{\alpha}) \times \cdots p(z_{i2}|z_{i1}, \mathbf{x}_i, \boldsymbol{\alpha}) \times p(z_{i1}|\mathbf{x}_i, \boldsymbol{\alpha}). \tag{3.3}$$

We typically assume specific parametric forms for these one-dimensional conditional distributions. This strategy allows much flexibility in the specification of the joint covariate distribution and has the potential of reducing the number of nuisance parameters (Lipsitz and Ibrahim, 1996; Ibrahim, Lipsitz, and Chen, 1999). Furthermore, we model the missing data mechanism using a sequence of one-dimensional conditional distributions as

$$\begin{aligned} p(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\xi}) &= p(r_{ip_2}|r_{i(p_2-1)}, \cdots, r_{i1}, \mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}) \times \\ & \cdots \times p(r_{i2}|r_{i1}, \mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}) p(r_{i1}|\mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}). \qquad (3.4) \end{aligned}$$

Since r_{ij} is binary, a sequence of logistic regressions is commonly used.

3.3 Local influence

We will develop a local influence method for carrying out sensitivity analyses of various assumptions of a GLM with missing covariates. Specifically, we will address three important issues related to local influence methods: perturbation schemes for perturbing the distributions for each component in (3.1), the appropriate choice of a perturbation vector, and the development of influence measures.

3.3.1 A simple example

Throughout this section, we examine a linear regression model with one missing covariate to illustrate our methodological development. We consider the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i, \qquad (3.5)$$

where $\epsilon_i \sim N(0, \tau)$. We assume that y_i and x_i are completely observed for $i = 1, \ldots, n$, but the covariate z_i may be missing for some cases. We also assume that $(z_i|x_i, \boldsymbol{\alpha}) \sim N(\alpha_0 + \alpha_1 x_i, \alpha_2)$, where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)$. We let $r_i = 1$ if z_i is missing and $r_i = 0$ if z_i is observed. Furthermore, we assume that the z_i's are MAR with missing data mechanism

$$p(r_i = 1|y_i, x_i, z_i) = \frac{\exp(\xi_0 + \xi_1 y_i + \xi_2 x_i)}{1 + \exp(\xi_0 + \xi_1 y_i + \xi_2 x_i)}. \tag{3.6}$$

We introduce various perturbations to perturb $p(D_c|\boldsymbol{\eta})$ and then we assess the sensitivity of each perturbation scheme to the proposed model and associated statistical inference. As an illustration, we consider four perturbations as follows. These perturbations illustrate two different types of perturbation schemes, which we discuss in the next subsection. The first is to perturb the variances of ϵ_i such that

$$\mathrm{Var}(\epsilon_1, \cdots, \epsilon_n) = \tau \mathrm{diag}(1/\omega_1, \cdots, 1/\omega_n). \tag{3.7}$$

Throughout, we let $\boldsymbol{\omega}^0$ denote no perturbation. In this case, $\boldsymbol{\omega}^0 = \mathbf{1}_n$ is an $n \times 1$ vector with all 1's. This perturbation is designed to assess the homogeneous variance assumption of the ϵ_i's. The second is to introduce a perturbation to z_i to assess the linear relationship between y_i and z_i such that

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (z_i + \omega_i) + \epsilon_i, \tag{3.8}$$

for $i = 1, \cdots, n$. In this case, $\boldsymbol{\omega}^0 = \mathbf{0}_n$, which is an $n \times 1$ vector with all 0's.

The third is to extend the MAR assumption such that

$$p(r_i = 1|y_i, x_i, z_i, \omega) = \frac{\exp(\xi_0 + \xi_1 y_i + \xi_2 x_i + \omega z_i)}{1 + \exp(\xi_0 + \xi_1 y_i + \xi_2 x_i + \omega z_i)}. \tag{3.9}$$

If $\omega \neq 0$, then the missing data mechanism is NMAR. This strategy for checking NMAR is similar to that of Verbeke et al. (2001) in the context of longitudinal data. Thus, (3.9) explores the influence of perturbing the MAR assumption ($\omega^0 = 0$) in the direction of NMAR. We emphasize here that formal tests for MAR or NMAR missingness should be approached with great caution, although they might be possible. Our main goal here and throughout this paper is to use local influence methods to carry out sensitivity analyses in order to assess the effect of perturbing the given GLM with MAR covariates in the direction of NMAR. An alternative to (3.9) is the individual-specific infinitesimal perturbation as used in Verbeke et al. (2001), Hens et al. (2005), and Jansen et al. (2006), which is given by

$$p(r_i = 1|y_i, x_i, z_i, \omega) = \frac{\exp(\xi_0 + \xi_1 y_i + \xi_2 x_i + \omega_i z_i)}{1 + \exp(\xi_0 + \xi_1 y_i + \xi_2 x_i + \omega_i z_i)}. \tag{3.10}$$

This can provide insight into which case may have large influence.

The fourth perturbation extends the linear relationship between z_i and x_i such that $(z_i|x_i, \boldsymbol{\alpha}) \sim N(\alpha_0 + \alpha_1 x_i + g(x_i), \alpha_2)$ for $i = 1, \cdots, n$, where $g(\cdot)$ is an unknown function. For instance, we may approximate $g(x)$ using a set of m basis functions (e.g., Fourier series, B-splines) $B_1(x)$, $\cdots$,, $B_m(x)$ such that $g(x) \approx \sum_{j=1}^{m} \omega_j B_j(x)$. Thus, we obtain

$$(z_i|x_i, \boldsymbol{\alpha}) \sim N(\alpha_0 + \alpha_1 x_i + \sum_{j=1}^{m} \omega_j B_j(x_i), \alpha_2) \tag{3.11}$$

for $i = 1, \cdots, n$. In (3.11), we are interested in assessing whether there is a nonlinear relationship between the covariate z_i and x_i. In this case, $\boldsymbol{\omega}^0 = \mathbf{0}_m$.

3.3.2 Perturbation schemes

We formally define two classes of perturbation schemes: the single-case and the global perturbation scheme. Let $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_m) \in R^m$ be a perturbation vector for the complete-data density $p(D_c|\boldsymbol{\eta})$. We use $p(D_c|\boldsymbol{\eta}, \boldsymbol{\omega})$ to denote the perturbed complete-data density such that $\int p(D_c|\boldsymbol{\eta}, \boldsymbol{\omega}) dD_c = 1$ and $p(D_c|\boldsymbol{\eta}, \boldsymbol{\omega}^0) = p(D_c|\boldsymbol{\eta})$. To assess the local influence of a model perturbation, we are primarily interested in the behavior of $p(D_c|\boldsymbol{\eta}, \boldsymbol{\omega})$ as a function of $\boldsymbol{\omega}$ around $\boldsymbol{\omega}^0$. We set $\boldsymbol{\eta}$ at a given value (e.g., the maximum likelihood estimate).

The single-case perturbation scheme refers to any scheme that independently perturbs individual observations (Verbeke et al., 2001). The single-case perturbation is mainly for identifying influential observations. Specifically, the perturbed complete-data density is

$$p(D_c|\boldsymbol{\eta}, \boldsymbol{\omega}) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i, y_i|\boldsymbol{\eta}, \omega_i), \tag{3.12}$$

where $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_n)$ and ω_i denotes the perturbation to the $i-$th observation. Such perturbation schemes, for example, include case weights for each of the three components of (3.1), perturbing individual components of $(\mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i)$ and perturbing individual components (or multiple components) of $\mathbf{r}_i$. Perturbation (3.7), (3.8), and (3.10) of the previous subsection belong to such a class.

The global perturbation scheme refers to any scheme that perturbs all observations simultaneously (Copas and Eguchi, 2005; Troxel, Ma, and Heitjan, 2004). The global perturbation is mainly for assessing the robustness of model assumptions to small perturbations. Specifically, the perturbed complete-data density is

$$p(D_c|\boldsymbol{\eta}, \boldsymbol{\omega}) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{z}_i, \mathbf{r}_i, y_i|\boldsymbol{\eta}, \boldsymbol{\omega}), \tag{3.13}$$

where $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_m)$ is shared by all the observations. Such a perturbation scheme includes the perturbation of each of the three components of (3.1) and simultaneous perturbations of the three components of (3.1), among many others. The number of components in $\boldsymbol{\omega}$ can be as small as one, such as perturbation (3.9) and other examples (Gustafson, 2001; Copas and Eguchi, 2005; Troxel, Ma, and Heitjan, 2004; Zhu et al., 2007a). Perturbation (3.11) is also a global perturbation scheme, in which m in the perturbation can increase with n.

3.3.3 Appropriate perturbation

We develop a new geometric framework to address the issue of selecting an appropriate perturbation scheme for (3.1). This issue is central to the development of the local influence approach, because arbitrarily perturbing a model may lead to inappropriate inference about the cause (e.g., influential observations) of a large effect.

The perturbed model $p(D_c|\boldsymbol{\eta}, \boldsymbol{\omega})$ has a natural geometrical structure. The perturbed model $M = \{p(D_c|\boldsymbol{\eta}, \boldsymbol{\omega}) : \boldsymbol{\omega} \in R^m\}$ can be regarded as an $m-$dimensional manifold. At each $\boldsymbol{\omega} \in R^m$, there is a tangent space $T_{\boldsymbol{\omega}}$ of M spanned by m functions $\partial_{\omega_j}\ell_c(\boldsymbol{\omega})$, where $\ell_c(\boldsymbol{\omega}) = \log p(D_c|\boldsymbol{\eta}, \boldsymbol{\omega})$. The m^2 quantities $g_{jk}(\boldsymbol{\omega}) = \mathrm{E}_{\boldsymbol{\omega}}[\partial_{\omega_j}\ell_c(\boldsymbol{\omega})\partial_{\omega_k}\ell_c(\boldsymbol{\omega})]$, $j, k = 1, \cdots, m$ form the *metric* tensor of M, in which $\mathrm{E}_{\boldsymbol{\omega}}$ denotes the expectation taken with respect to $p(D_c|\boldsymbol{\eta}, \boldsymbol{\omega})$, and the metric matrix $G(\boldsymbol{\omega}) = (g_{ij}(\boldsymbol{\omega}))$ is the Fisher information matrix with respect to the perturbation vector $\boldsymbol{\omega}$ (Figure 3.1).

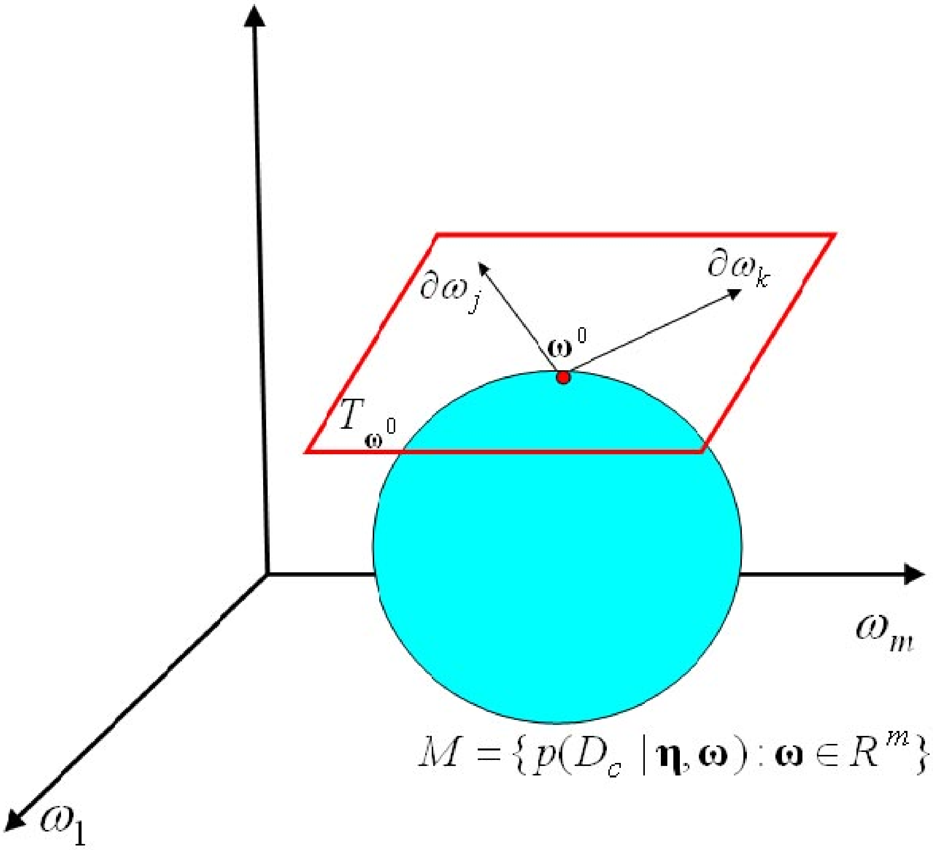
$\partial\omega_j$
$\partial\omega_k$
$\boldsymbol{\omega}^0$
$T_{\boldsymbol{\omega}^0}$
ω_m
$M = \{p(D_c \mid \boldsymbol{\eta}, \boldsymbol{\omega}) : \boldsymbol{\omega} \in R^m\}$
ω_1

Figure 3.1: A graphical representation of the perturbation manifold.

An *appropriate perturbation* to (3.1) requires that $G(\boldsymbol{\omega}^0) = \mathrm{diag}(g_{11}(\boldsymbol{\omega}^0), \cdots, g_{mm}(\boldsymbol{\omega}^0))$. The elements of $G(\boldsymbol{\omega})$ measure the amount of perturbation introduced by all components of the perturbation vector $\boldsymbol{\omega}$. The $g_{ii}(\boldsymbol{\omega})$ can be interpreted as the amount of perturbation introduced by ω_i, whereas $r_{ij}(\boldsymbol{\omega}) = g_{ij}(\boldsymbol{\omega})/\sqrt{g_{ii}(\boldsymbol{\omega})g_{jj}(\boldsymbol{\omega})}$ indicates an association between ω_i and ω_j. For a diagonal matrix $G(\boldsymbol{\omega})$, all components of $\boldsymbol{\omega}$ may be regarded as being orthogonal to each other in the perturbed model (Cox and Reid, 1987), and therefore it becomes easy to pinpoint the cause of a large effect. In applications, although $G(\boldsymbol{\omega}^0)$ may not be diagonal, we can always choose a new perturbation vector $\tilde{\boldsymbol{\omega}}$, defined by

$$\tilde{\boldsymbol{\omega}}(\boldsymbol{\omega}) = \boldsymbol{\omega}^0 + c^{-1/2}G(\boldsymbol{\omega}^0)^{1/2}(\boldsymbol{\omega} - \boldsymbol{\omega}^0), \tag{3.14}$$

such that $G(\tilde{\boldsymbol{\omega}})$ evaluated at $\boldsymbol{\omega}^0$ equals $c\mathbf{I}_m$, where $c > 0$.

For the single-case perturbation scheme (3.12), we have $g_{jk}(\boldsymbol{\omega}) = \delta_{jk}\mathrm{E}_{\boldsymbol{\omega}}[\partial_{\omega_j}\ell_{c,j}(\boldsymbol{\omega})]^2$, for $j, k = 1, \cdots, n$, where δ_{jk} is the Kronecker delta and $\ell_{c,j}(\boldsymbol{\omega}) = \log p(\mathbf{d}_j|\boldsymbol{\eta}, \omega_j)$. The diagonal structure of $G(\boldsymbol{\omega}) = (g_{jk}(\boldsymbol{\omega}))$ indicates that all components of $\boldsymbol{\omega}$ are orthogonal to each other. Furthermore, if $p(\mathbf{d}_i|\boldsymbol{\eta}, \omega_i)$ is invariant across all i, then $G(\boldsymbol{\omega}) = g_{11}(\boldsymbol{\omega})\mathbf{I}_n$, which indicates that different components of $\boldsymbol{\omega}$ have the same influence on the corresponding distributions.

For the global perturbation scheme, we have $g_{jk}(\boldsymbol{\omega}) = -\sum_{i=1}^n E_{\boldsymbol{\omega}}[\partial^2_{\omega_j\omega_k}\ell_{c,i}(\boldsymbol{\omega})]$. Although $\boldsymbol{\omega}$ may not be appropriate, we can choose a new perturbation $\tilde{\boldsymbol{\omega}} = \boldsymbol{\omega}^0 + G(\boldsymbol{\omega}^0)^{1/2}(\boldsymbol{\omega} - \boldsymbol{\omega}^0)$ such that $G(\tilde{\boldsymbol{\omega}}^0) = \mathbf{I}_n$. Thus, $\tilde{\boldsymbol{\omega}}$ is an appropriate perturbation at least at $\tilde{\boldsymbol{\omega}}^0 = \boldsymbol{\omega}^0$. For instance, we consider the perturbation (3.11) to the model in Section 3.1. It can be shown that

$$-\partial^2_{\omega_j\omega_k}\ell_c(\boldsymbol{\omega}) = \alpha_2^{-1}B_j(x_i)B_k(x_i) \text{ and } g_{jk}(\boldsymbol{\omega}) = \alpha_2^{-1}\sum_{i=1}^n \int B_j(x_i)B_k(x_i)p(x_i|\boldsymbol{\alpha})dx_i,$$

where $p(x_i|\boldsymbol{\alpha})$ is the distribution of x_i. If $\{B_j(x) : j = 1, \cdots, m\}$ forms an orthonormal basis with respect to $p(x|\boldsymbol{\alpha})$, then $G(\boldsymbol{\omega})$ is just an $m \times m$ identity matrix. However, since the x_i's are always observed, we can always treat x_i as fixed and approximate $g_{jk}(\boldsymbol{\omega})$ using $g_{jk}(\boldsymbol{\omega}) = \alpha_2^{-1}\sum_{i=1}^n B_j(x_i)B_k(x_i)$.

3.3.4 Influence measures

(i) First-order influence measures

We consider a $b \times 1$ objective function $f(\boldsymbol{\omega}) : M \rightarrow R^b$ such as the maximum likelihood estimate of $\boldsymbol{\eta}$ (Copas and Eguchi, 2001, 2005; Troxel, Ma, and Heitjan, 2004; Gustafson, 2001). The objective function $f(\boldsymbol{\omega})$ defines the aspect of inference of interest for sensitivity analysis. Let $\boldsymbol{\omega}(t)$ be a geodesic on M with $\boldsymbol{\omega}(0) = \boldsymbol{\omega}^0$ and $\partial_t \boldsymbol{\omega}(t)|_{t=0} = \mathbf{h} \in R^m$. It follows from a Taylor's series expansion that $f(\boldsymbol{\omega}(t)) = f(\boldsymbol{\omega}(0)) + \dot{f}_{\mathbf{h}}(0)t + O(t^2)$, where $\dot{f}_{\mathbf{h}}(0) = \sum_j \partial_{\omega_j} f(\boldsymbol{\omega}^0) h_j = \nabla'_f \mathbf{h}$. If $\nabla_f \neq 0$, then the first-order term $\dot{f}_{\mathbf{h}}(0)$ mainly characterizes the local influence of a perturbation vector $\boldsymbol{\omega}$ to a model.

We introduce a first-order influence measure to assess the local influence of minor perturbations when $\nabla_f \neq 0$. The *first-order influence measure* (FI) in the direction $\mathbf{h} \in R^m$ is $\mathrm{FI}_{f,\mathbf{h}} = \mathrm{FI}_{f(\boldsymbol{\omega}^0),\mathbf{h}} = \frac{\mathbf{h}' \nabla_f W_f \nabla'_f \mathbf{h}}{\mathbf{h}' G \mathbf{h}}$, where $G = G(\boldsymbol{\omega}^0)$ and W_f is a non-negative symmetric matrix.

Although $\boldsymbol{\omega}$ may not be an appropriate perturbation, we can always use the appropriate perturbation $\tilde{\boldsymbol{\omega}}(\boldsymbol{\omega})$ in (3.14), which yields

$$\mathrm{FI}_{f(\tilde{\boldsymbol{\omega}}),\mathbf{h}}\Big|_{\tilde{\boldsymbol{\omega}}=\boldsymbol{\omega}^0} = \frac{\mathbf{h}' G^{-1/2} \nabla_f W_f \nabla'_f G^{-1/2} \mathbf{h}}{\mathbf{h}'\mathbf{h}}. \tag{3.15}$$

The maximum value of $\mathrm{FI}_{f,\mathbf{h}}$ equals the principal eigenvalue of $G^{-1/2} \nabla_f W_f \nabla'_f G^{-1/2}$, which quantifies the largest degree of local influence of $\tilde{\boldsymbol{\omega}}$ to a statistical model, while the corresponding eigenvector of $G^{-1/2} \nabla_f W_f \nabla'_f G^{-1/2}$, denoted by $\mathbf{h}_{\max}$, can be used either for identifying influential observations for single-case perturbations or for identifying influential directions for global perturbations (Copas and Eguchi, 2005). The $\mathbf{h}_{\max}$ is the worst perturbation direction for $f(\tilde{\boldsymbol{\omega}})$.

(ii) Maximum likelihood estimate as the objective function

Let D_o denote the observed data. We consider $\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}) = (\hat{\boldsymbol{\beta}}_o(\boldsymbol{\omega}), \hat{\boldsymbol{\alpha}}_o(\boldsymbol{\omega}), \hat{\boldsymbol{\xi}}_o(\boldsymbol{\omega}))'$, which is the maximum likelihood estimate of $\boldsymbol{\eta}$ based on the perturbed observed-data density. The perturbed observed-data density, denoted by $p(D_o|\boldsymbol{\eta}, \boldsymbol{\omega})$, is associated with the perturbed complete-data density through $p(D_o|\boldsymbol{\eta}, \boldsymbol{\omega}) = \int p(D_c|\boldsymbol{\eta}, \boldsymbol{\omega}) dD_m$. It can be shown that

$$\partial_{\boldsymbol{\omega}} \hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0) = I^{-1}_{\boldsymbol{\eta},o} \Delta_o(\boldsymbol{\eta}, \boldsymbol{\omega}) \Big|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}, \boldsymbol{\omega}=\boldsymbol{\omega}^0}, \tag{3.16}$$

where $I_{\boldsymbol{\eta},o} = -\partial^2_{\boldsymbol{\eta}} \log p(D_o|\boldsymbol{\eta})$ and $\Delta_o(\boldsymbol{\eta}, \boldsymbol{\omega}) = \partial^2_{\boldsymbol{\eta}\boldsymbol{\omega}} \log p(D_o|\boldsymbol{\eta}, \boldsymbol{\omega})$. Then, the asymptotic bias in the estimate of $\boldsymbol{\eta}$ is $\partial_{\boldsymbol{\omega}} \hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0)(\boldsymbol{\omega} - \boldsymbol{\omega}^0)$ under $p(D_o|\boldsymbol{\eta}, \boldsymbol{\omega})$.

We choose $\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega})$ as the object of interest and set $W_f = I_{\hat{\boldsymbol{\eta}},o}$. We can show that

$$\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\tilde{\boldsymbol{\omega}}),\mathbf{h}} = \mathbf{h}' G^{-1/2} \Delta_o(\hat{\boldsymbol{\eta}}, \boldsymbol{\omega}^0)' I^{-1}_{\hat{\boldsymbol{\eta}},o} \Delta_o(\hat{\boldsymbol{\eta}}, \boldsymbol{\omega}^0) G^{-1/2} \mathbf{h}, \tag{3.17}$$

where $\mathbf{h}'\mathbf{h} = 1$. If $G = \mathbf{I}_m$, then it can be shown that $\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\tilde{\boldsymbol{\omega}}),\mathbf{h}}$ is the same as Cook's (1986) local influence measure based on the likelihood displacement. Finally, for most GLMs with missing covariates, computing the matrix $G^{-1/2} \Delta_o(\hat{\boldsymbol{\eta}}, \boldsymbol{\omega}^0)' I^{-1}_{\hat{\eta},o} \Delta_o(\hat{\boldsymbol{\eta}}, \boldsymbol{\omega}^0) G^{-1/2}$ involves the computation of G, $\Delta_o(\hat{\boldsymbol{\eta}}, \boldsymbol{\omega}^0)$, and $I_{\boldsymbol{\eta},o}$, which can be obtained using MCMC methods.

For the single-case perturbation in (3.12), we obtain $G(\boldsymbol{\omega}^0) = g_{11}(\boldsymbol{\omega}^0)\mathbf{I}_n$ and the i-th column of $\Delta_o(\boldsymbol{\eta}, \boldsymbol{\omega}^0)$, denoted by $\delta_{\boldsymbol{\eta},i}$, is given by $\partial^2_{\boldsymbol{\eta}\omega_i}\{\log \int p(\mathbf{d}_i|\boldsymbol{\eta}, \omega_i) d\mathbf{z}_{m,i}\}$. Thus,

$$\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\tilde{\boldsymbol{\omega}}),\mathbf{h}} = g_{11}(\boldsymbol{\omega}^0)^{-1} \mathbf{h}' \Delta_o(\hat{\boldsymbol{\eta}}, \boldsymbol{\omega}^0)' I^{-1}_{\hat{\boldsymbol{\eta}},o} \Delta_o(\hat{\boldsymbol{\eta}}, \boldsymbol{\omega}^0) \mathbf{h}. \tag{3.18}$$

In particular, for the i-th observation, $\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\tilde{\boldsymbol{\omega}}),\mathbf{e}_i} = g_{11}(\boldsymbol{\omega}^0)^{-1} \delta'_{\hat{\boldsymbol{\eta}},i} I^{-1}_{\hat{\boldsymbol{\eta}},o} \delta_{\hat{\boldsymbol{\eta}},i}$, and $\sum_{i=1}^n \mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\tilde{\boldsymbol{\omega}}),\mathbf{e}_i} = g_{11}(\boldsymbol{\omega}^0)^{-1} \mathrm{tr}\{\sum_{i=1}^n \delta_{\hat{\boldsymbol{\eta}},i} \delta'_{\hat{\boldsymbol{\eta}},i} I^{-1}_{\hat{\boldsymbol{\eta}},o}\}$. Under mild conditions, $\sum_{i=1}^n \delta_{\hat{\boldsymbol{\eta}},i} \delta'_{\hat{\boldsymbol{\eta}},i}/n$ and $I_{\hat{\boldsymbol{\eta}},o}/n$ converge in probability to J_o and I_o, respectively. Therefore, $\sum_{i=1}^n \mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\tilde{\boldsymbol{\omega}}),\mathbf{e}_i}$ is a direct estimate of $\lambda_0 = \mathrm{tr}(J_o I_o^{-1}) g_{11}(\boldsymbol{\omega}^0)^{-1}$. Under exchangeability of the observations, each $\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\tilde{\boldsymbol{\omega}}),\mathbf{e}_i}$ should be around its mean λ_0. However, in real applications, if a particular $\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\tilde{\boldsymbol{\omega}}),\mathbf{e}_i}$ is much larger than λ_0, then this observation may be regarded as an influential case.

(iii) Likelihood ratio as the objective function

We consider $f_{lr}(\boldsymbol{\omega}) = \log p(D_o|\hat{\boldsymbol{\eta}}, \boldsymbol{\omega}) - \log p(D_o|\hat{\boldsymbol{\eta}})$ as our objective function. For the single-case perturbation in (3.12), we can obtain that $G(\boldsymbol{\omega}^0) = g_{11}(\boldsymbol{\omega}^0)\mathbf{I}_n$ and

$$\partial_{\omega_i} f_{lr}(\boldsymbol{\omega}) = \partial_{\omega_i} \log p(\mathbf{d}_{o,i}|\hat{\boldsymbol{\eta}}, \omega_i) = E[\partial_{\omega_i} \log p(\mathbf{d}_i|\hat{\boldsymbol{\eta}}, \omega_i)|\mathbf{d}_{o,i}, \hat{\boldsymbol{\eta}}]$$

for $i = 1, \cdots, n$, where the expectation is taken with respect to the conditional distribution of $\mathbf{z}_{m,i}$ given $\mathbf{d}_{o,i}$. Thus, by setting $W_{f_{lr}} = 1$, we get $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}),\mathbf{h}} = g_{11}(\boldsymbol{\omega}^0)^{-1}\mathbf{h}'\nabla_{f_{lr}}\nabla'_{f_{lr}}\mathbf{h}$. For the i-th observation, we have $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}),\mathbf{e}_i} = g_{11}(\boldsymbol{\omega}^0)^{-1}\{\partial_{\omega_i} f_{lr}(\boldsymbol{\omega}^0)\}^2$. If a particular $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}),\mathbf{e}_i}$ is much larger than the mean of all $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}),\mathbf{e}_i}$'s, then the $i-$th observation can be regarded as influential.

For the global-case perturbation in (3.13), $\log p(D_o|\hat{\boldsymbol{\eta}}, \boldsymbol{\omega}) = \sum_{i=1}^n \log p(\mathbf{d}_{o,i}, \hat{\boldsymbol{\eta}}, \boldsymbol{\omega})$. Direct calculation leads to

$$\nabla_{f_{lr}} = \sum_{i=1}^{n} \partial_{\boldsymbol{\omega}} \log p(\mathbf{d}_{o,i}, \hat{\boldsymbol{\eta}}, \boldsymbol{\omega}^0) = \sum_{i=1}^{n} E[\partial_{\boldsymbol{\omega}} \log p(\mathbf{d}_i, \hat{\boldsymbol{\eta}}, \boldsymbol{\omega}^0)|\mathbf{d}_{o,i}, \hat{\boldsymbol{\eta}}]. \tag{3.19}$$

Setting $W_{f_{lr}} = 1$ and choosing $\tilde{\boldsymbol{\omega}}$ in (3.14), we have $\mathrm{FI}_{f_{lr}(\tilde{\boldsymbol{\omega}}),\mathbf{h}} = \mathbf{h}'G^{-1/2}\nabla_{f_{lr}}\nabla'_{f_{lr}}G^{-1/2}\mathbf{h}$, where $\mathbf{h}'\mathbf{h} = 1$. The maximum value is the principal eigenvalue $\mathrm{FI}_{f_{lr}(\tilde{\boldsymbol{\omega}}),\mathbf{h}_{\max}} = \nabla'_{f_{lr}}G^{-1}\nabla_{f_{lr}}$ and its corresponding $\mathbf{h}_{\max}$ is $G^{-1/2}\nabla_{f_{lr}}/||G^{-1/2}\nabla_{f_{lr}}||$. Moreover, under some mild conditions $\nabla'_{f_{lr}}G^{-1}\nabla_{f_{lr}}$ can be used as a test statistic for testing $H_0 : \boldsymbol{\omega} = 0$. Under $H_0 : \boldsymbol{\omega} = 0$, it can be shown that $\nabla_{f_{lr}}/\sqrt{n}$ converges in distribution to a Gaussian distribution with zero mean and covariance matrix $\Sigma_{f_{lr}}$ as $n \to \infty$. Thus, $\nabla'_{f_{lr}}\Sigma_{f_{lr}}^{-1/2}\Sigma_{f_{lr}}^{1/2}G^{-1}\Sigma_{f_{lr}}^{1/2}\Sigma_{f_{lr}}^{-1/2}\nabla_{f_{lr}}$ converges in distribution to a weighted chi-squared distribution as $n \to \infty$. Therefore, we may use the asymptotic distribution of $\nabla'_{f_{lr}}G^{-1}\nabla_{f_{lr}}$ to characterize the asymptotic behavior of the influence measures $\mathrm{FI}_{f_{lr}(\tilde{\boldsymbol{\omega}}),\mathbf{h}}$.

3.4 Simulation studies

We applied the proposed local influence measures to several simulated datasets in which various assumptions were misspecified in order to examine their performance. First, we applied two single-case perturbation schemes to simulated datasets in each of which an 'outlier' was added. We expected that both schemes could detect the 'outlier' both in the response and in the covariates. Secondly, we used several perturbation schemes to examine the functional form of the missing data mechanism and to assess the relationship between the response and covariates.

We generated 500 simulated datasets from model (3.5) with $n = 100$, $\beta_0 = \beta_1 = \beta_2 = 1$ and $\tau = 1$. Moreover, (x_i, z_i) were generated from a $N_2(\mathbf{0}, \mathbf{I}_2)$ distribution. We also assumed an MAR missing data mechanism for z_i given by

$$p(r_i = 1|x_i, z_i, y_i) = \frac{\exp(\xi_0 + \xi_1 x_i)}{1 + \exp(\xi_0 + \xi_1 x_i)}, \tag{3.20}$$

with $\xi_0 = -0.5$ and $\xi_1 = 1.0$, resulting in an average missingness fraction of 40%. Then, we fit $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ with MAR z_i, and changed y_{100} to $y_{100} + \delta$ with $\delta = 1.0$, 2.0, 3.0, 4.0, and 5.0 in order to add an 'outlier'. We applied two single-case perturbation schemes. The first was to perturb the variance of ϵ_i such that $\text{Var}(\epsilon_1, \cdots, \epsilon_n) = \tau \text{diag}(1/\omega_1, \cdots, 1/\omega_n)$, where $\boldsymbol{\omega}^0 = \mathbf{1}_n$ is an $n \times 1$ vector with all 1's. The second perturbation was to perturb the missing covariate z_i such that $y_i = \beta_0 + \beta_1 x_i + \beta_2(z_i + \omega_i) + \epsilon_i$ for $i = 1, \cdots, n$, where $\boldsymbol{\omega}^0 = \mathbf{0}_n$ is an $n \times 1$ vector with all 0's. We calculated $\text{FI}_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0), \mathbf{e}_i}$ and $\text{FI}_{f_{lr}(\boldsymbol{\omega}^0), \mathbf{e}_i}$ for both perturbations, and their values for the last case were larger than those for the rest of the cases, especially when δ is large. The first half of Table 3.1 summarizes the percentages of detecting the 'outlier' using either nonrobust methods with the sample mean and standard deviation (>mean+2×SD or >mean+3×SD) or robust methods with the sample median and median absolute deviation (>median+2×MAD or >median+3×MAD) for different δ values. As expected, the percentage of detecting the 'outlier' increases with δ, and the results based on the robust methods are better compared to the nonrobust methods. The threshold based on 3 deviations (SD or MAD) is not very different from using a threshold based on 2 deviations. Based on a simulated dataset, in which $\delta = 4$, the index plots of the two influence measures (Figure 3.2) can effectively detect the 'outlier'.

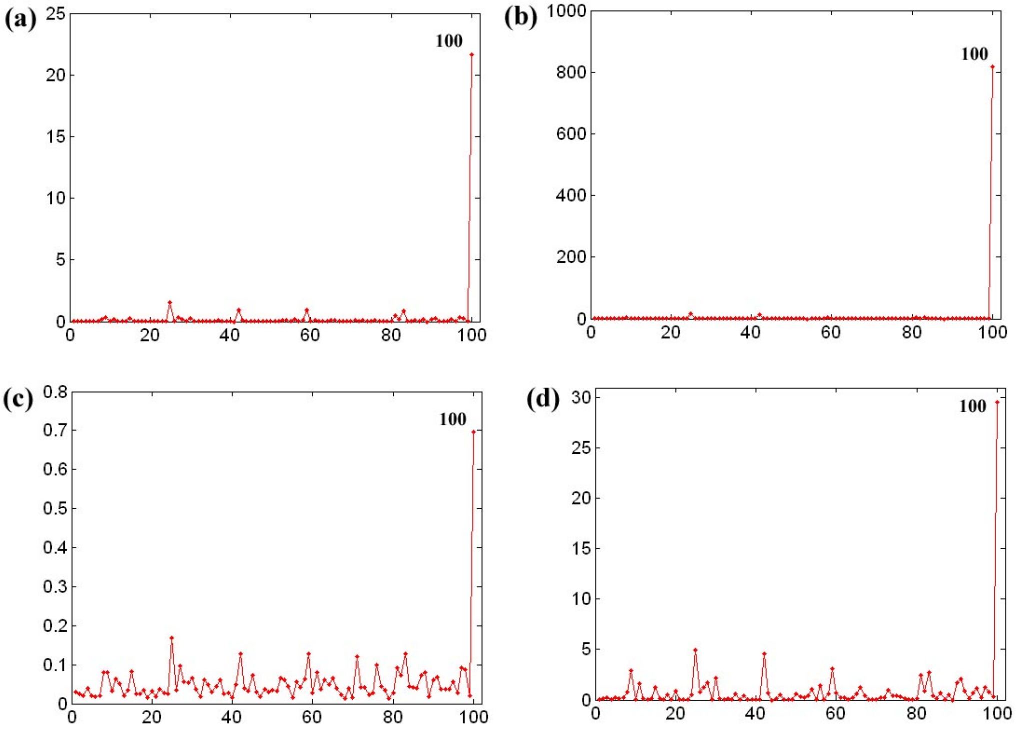

Figure 3.2: Index plots of influence measures from a simulated dataset with y_{100} as an influential case: (a) $\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0),\mathbf{e}_i}$ and (b) $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}^0),\mathbf{e}_i}$ for the variance perturbation; (c) $\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0),\mathbf{e}_i}$ and (d) $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}^0),\mathbf{e}_i}$ for the missing covariate perturbation.

Instead of having an 'outlier' in the response, we examined a scenario with the presence of the 'outlier' in the covariates. We changed z_{100} to $z_{100} + \delta$ with $\delta =$ 1.0, 2.0, 3.0, 4.0, and 5.0, and applied the same two single-case perturbation schemes. The values of $\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0),\mathbf{e}_i}$ and $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}^0),\mathbf{e}_i}$ for both perturbations for the last case were again larger than those for the rest of the cases, especially when δ is large. The second half of Table 3.2 lists the percentages of detecting the 'outlier' using either nonrobust or robust methods mentioned previously. It shows similar findings as when the 'outlier'

Table 3.1: Percentages of detecting the ‘outlier’ using nonrobust methods and robust methods for different δ values and thresholds. 500 simulated datasets were used for each case.

The ‘outlier’ is in the response											
		$> mean + 2 \times SD$					$> mean + 3 \times SD$				
		δ					δ				
Pertubation	Stat	1	2	3	4	5	1	2	3	4	5
1	FI$_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0)}$	.108	.420	.718	.888	.966	.078	.354	.618	.840	.944
	FI$_{f_{lr}(\boldsymbol{\omega}^0)}$	.112	.404	.678	.850	.932	.084	.350	.624	.814	.916
2	FI$_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0)}$	.112	.426	.720	.888	.974	.060	.310	.576	.808	.926
	FI$_{f_{lr}(\boldsymbol{\omega}^0)}$	.140	.466	.740	.884	.950	.084	.350	.636	.812	.908
		$> median + 2 \times MAD$					$> median + 3 \times MAD$				
		δ					δ				
Pertubation	Stat	1	2	3	4	5	1	2	3	4	5
1	FI$_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0)}$	.484	.772	.924	.982	.998	.434	.746	.906	.974	.998
	FI$_{f_{lr}(\boldsymbol{\omega}^0)}$	.274	.642	.858	.960	.996	.254	.642	.844	.958	.994
2	FI$_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0)}$	.336	.642	.854	.958	.994	.228	.526	.792	.936	.986
	FI$_{f_{lr}(\boldsymbol{\omega}^0)}$	.448	.756	.920	.974	.996	.358	.696	.872	.966	.996

The ‘outlier’ is in the covariates											
		$> mean + 2 \times SD$					$> mean + 3 \times SD$				
		δ					δ				
Pertubation	Stat	1	2	3	4	5	1	2	3	4	5
1	FI$_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0)}$	.150	.464	.762	.806	.830	.100	.366	.712	.804	.830
	FI$_{f_{lr}(\boldsymbol{\omega}^0)}$	.102	.320	.580	.710	.800	.080	.264	.516	.680	.784
2	FI$_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0)}$	.192	.534	.784	.816	.832	.090	.388	.724	.814	.830
	FI$_{f_{lr}(\boldsymbol{\omega}^0)}$	.136	.376	.644	.750	.820	.082	.270	.532	.688	.788
		$> median + 2 \times MAD$					$> median + 3 \times MAD$				
		δ					δ				
Pertubation	Stat	1	2	3	4	5	1	2	3	4	5
1	FI$_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0)}$	.460	.734	.828	.844	.858	.422	.704	.820	.840	.852
	FI$_{f_{lr}(\boldsymbol{\omega}^0)}$	.246	.498	.750	.804	.834	.232	.482	.738	.798	.834
2	FI$_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0)}$	.382	.706	.820	.828	.840	.282	.620	.806	.822	.836
	FI$_{f_{lr}(\boldsymbol{\omega}^0)}$	.398	.644	.810	.828	.848	.338	.586	.780	.814	.838

is in the response. Thus our local influence can detect the ‘outlier’ in the covariates effectively when δ is reasonably large.

Next, we explored the potential deviations of the MAR missing data mechnism in the direction of NMAR. We generated data from model (3.5) with $n = 200$, $\beta_0 = \beta_1 = \beta_2 = 1$, $\tau = 1$, (x_i, z_i) were generated from a $N_2(\mathbf{0}, \mathbf{I}_2)$ distribution, and the following missing data mechanism for z_i was assumed,

$$p(r_i = 1|x_i, z_i, y_i) = \frac{\exp(\xi_0 + \xi_1 y_i + \xi_2 x_i + a z_i)}{1 + \exp(\xi_0 + \xi_1 y_i + \xi_2 x_i + a z_i)}, \tag{3.21}$$

with $\xi_0 = -1.8, \quad \xi_1 = 1.0$, and $\xi_2 = 1.0$ being chosen to make the missing data fraction approximately 40% for various values of a. If $a \neq 0$, then the missing data mechanism is nonignorable. We fit $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ with the MAR mechanism given by

$$p(r_i = 1|x_i, z_i, y_i) = \frac{\exp(\xi_0 + \xi_1 y_i + \xi_2 x_i)}{1 + \exp(\xi_0 + \xi_1 y_i + \xi_2 x_i)}. \tag{3.22}$$

Then, we applied a global perturbation given by

$$p(r_i = 1|y_i, x_i, z_i, \omega) = \frac{\exp(\xi_0 + \xi_1 y_i + \xi_2 x_i + \omega z_i)}{1 + \exp(\xi_0 + \xi_1 y_i + \xi_2 x_i + \omega z_i)}. \tag{3.23}$$

The $\mathrm{FI}_{f_{lr}(\omega^0)}$ were $0.084, 1.448$, and 4.795 for $a = 0, \quad 0.5$, and 1.0 respectively. From these results, we see that as a increases, the influence measure of $\mathrm{FI}_{f_{lr}(\omega^0)}$ also increases, which may suggest that an NMAR model is tenable for large a. We also used the corresponding single-case perturbation given by

$$p(r_i = 1|y_i, x_i, z_i, \omega) = \frac{\exp(\xi_0 + \xi_1 y_i + \xi_2 x_i + \omega_i z_i)}{1 + \exp(\xi_0 + \xi_1 y_i + \xi_2 x_i + \omega_i z_i)}. \tag{3.24}$$

No large $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}^0),e_i}$ was observed for any i even when a is large. This result might suggest that this type of NMAR mechanism is not detectable using only $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}^0),e_i}$, the diagonal entries of $G^{-1/2}\nabla_{f_{lr}} W_{f_{lr}} \nabla'_{f_{lr}} G^{-1/2}$, confirming the analyses in Jansen et al. (2006).

However, we observed increases in the off-diagonal entries of $G^{-1/2}\nabla_{f_{lr}} W_{f_{lr}} \nabla'_{f_{lr}} G^{-1/2}$ as a increases, indicating influence through combinations of cases.

As noted in Hens et al. (2005) and Jansen et al. (2006), a local influence tool for the missing data mechanism is able to pick up anomalous features of cases that are not necessarily related to the missing data mechanism. To study this notion, we generated an original dataset from model (3.5) with $n = 200$, $\beta_0 = \beta_1 = \beta_2 = 1$, $\tau = 1$, (x_i, z_i) were generated from a $N_2(\mathbf{0}, \mathbf{I}_2)$ distribution, and MAR was assumed. Then we generated a perturbed dataset in which we added 20 to the responses of the last five cases. We fit $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ with the MAR missing data mechanism given by (3.22). The perturbation (3.24) identified the last five cases as influential. Thus single-case perturbation for the missing data mechanism is able to pick up some deviations in the data even though the deviations are different from the functional form of the missing data mechanism. The global perturbation (3.23) resulted in $\text{FI}_{f_{lr}(\omega^0)} = 1.61$, a big qualitative change compared to $\text{FI}_{f_{lr}(\omega^0)} = .011$ for the original dataset. These results may thus raise some concerns about the MAR assumption, and/or about the model as a whole.

We also examined whether our influence measures can assess the relationship between the response and the covariates of interest. We generated data from $y_i = 1 + x_i + z_i + c * z_i^2 + \epsilon_i$ for $i = 1, \cdots, 100$, where $\epsilon_i \sim N(0, 1)$ and (x_i, z_i) were independently generated from a $N_2(\mathbf{0}, \mathbf{I}_2)$ distribution. The missing data mechanism was assumed MAR as in (3.20) with a 40% missingness fraction. We fit $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ assuming MAR z_i's, and thus the fitted model would be misspecified if $c \neq 0$. We considered a global perturbation scheme $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \sum_{j=1}^{m+3} \omega_j B_j(z_i) + \epsilon_i$, where the $B_j(z)$ are truncated polynomials of order $2-4$, given by z^2, z^3, z^4, $(z-k_1)_+^4, \cdots, (z-k_m)_+^4$, where $k_1, \cdots, k_m$ are the $m = 3$ prefixed knots. The principal eigenvalue of $\text{FI}_{f_{lr}(\boldsymbol{\omega}^0)}$ was 0.582, 13.675, 24.535, and 33.233 for $c = 0,\ 0.4,\ 0.8$, and 1.2 respectively. The principal eigenvalue of $\text{FI}_{f_{lr}(\boldsymbol{\omega}^0)}$ was statistically significant at the 5% significance level ($p-$value $= 0.002$) for $c = .8$, but not for $c < .8$. Thus, the local influence measures are useful for detecting model misspecification in this example.

3.5 Real data analysis

3.5.1 Quality of life data

As mentioned in Section 3.1, the response variable for these data is the logarithm of the survival time. The dataset has 404 observations and the covariates are: physical ability (z_1); mood (z_2); indicator for treatment A (yes, no) (x_1); indicator for treatment B (yes, no) (x_2); indicator for treatment C (yes, no) (x_3); age (x_4); and language (English, otherwise)(x_5). Among these seven covariates, z_1 has 13% missingness and z_2 has 31% missingness, and the remaining covariates are fully observed.

We fit a regression model $y_i = \mathbf{v}_i'\boldsymbol{\beta}+\epsilon_i$, where $\epsilon_i \sim N(0,\tau)$, $\mathbf{v}_i' = (1, z_{i1}, z_{i2}, x_{i1}, \cdots, x_{i5})$ is the 1×8 vector of covariates and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_7)^T$ are the corresponding regression coefficients. Since only the continuous covariates z_1 and z_2 have missing values, we assumed $(z_{i1}, z_{i2}) \sim N_2(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$, for $i = 1, \cdots, n$. We assumed that the missing covariates are MAR and calculated the maximum likelihood estimates of $(\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ using the EM algorithm.

To detect the influential cases, we employed two single-case perturbation schemes. The first was to perturb the variance of ϵ_i such that $\text{Var}(\epsilon_1, \cdots, \epsilon_n) = \tau\text{diag}(1/\omega_1, \cdots, 1/\omega_n)$. We calculated $\text{FI}_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0),\mathbf{e}_i}$ and $\text{FI}_{f_{lr}(\boldsymbol{\omega}^0),\mathbf{e}_i}$, and both indicated that cases 132 and 404 were very influential (Figure 3.3 a, b). The second perturbation was to simultaneously perturb the missing covariates z_{i1} and z_{i2} such that $y_i = \beta_0 + \beta_1(z_{i1} + \omega_i) + \beta_2(z_{i2} + \omega_i) + \beta_3 x_{i1} + \cdots + \beta_7 x_{i5} + \epsilon_i$. Again, cases 132 and 404 were very influential (Figure 3.3 c, d). The response values of cases 132 and 404 are very small, compared to the rest of the cases.

Next, we were interested in the sensitivity analyses on the MAR assumption in the

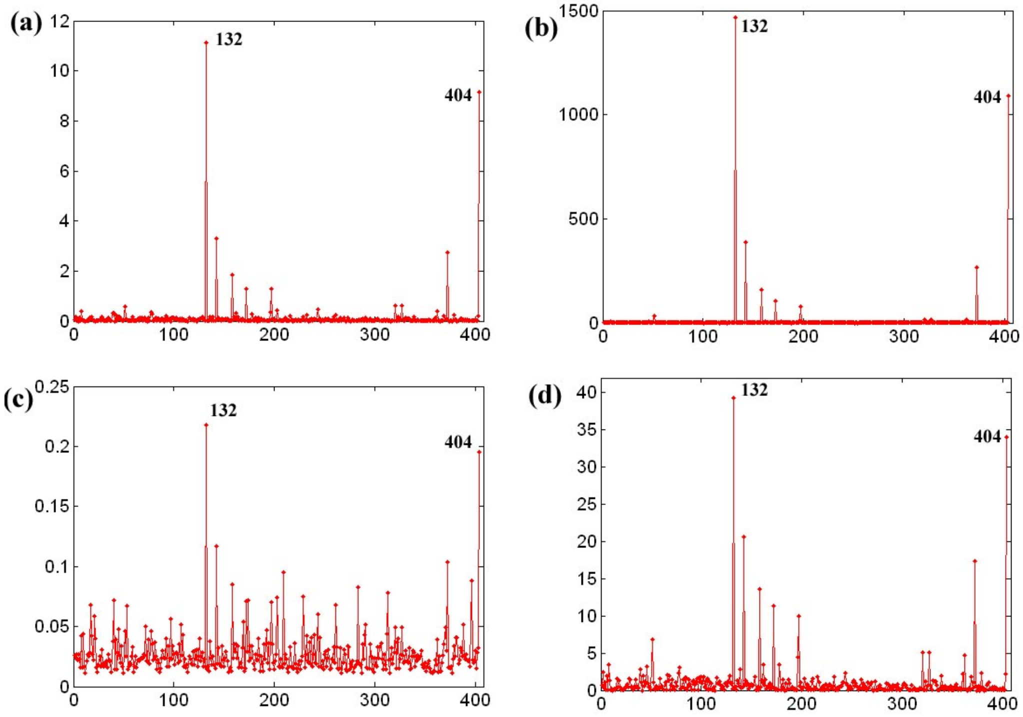

Figure 3.3: Index plots of influence measures for quality of life data: (a) $\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0),\mathbf{e}_i}$ and (b) $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}^0),\mathbf{e}_i}$ for the variance perturbation; (c) $\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0),\mathbf{e}_i}$ and (d) $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}^0),\mathbf{e}_i}$ for the missing covariate perturbation.

direction of NMAR. First, we fit the model with a MAR missing data mechanism

$$p(\mathbf{r}_i|\mathbf{x}_i, y_i, \boldsymbol{\xi}) = p(r_{i2}|r_{i1}, \mathbf{x}_i, y_i, \boldsymbol{\xi}_2)p(r_{i1}|\mathbf{x}_i, y_i, \boldsymbol{\xi}_1), \tag{3.25}$$

where $p(r_{i1}|\mathbf{x}_i, y_i, \boldsymbol{\xi}_1) = \frac{\exp(r_{i1}f_{i1})}{1+\exp(f_{i1})}$ and $p(r_{i2}|r_{i1}, \mathbf{x}_i, y_i, \boldsymbol{\xi}_2) = \frac{\exp(r_{i2}f_{i2})}{1+\exp(f_{i2})}$, $f_{i1} = \xi_{10} + \xi_{11}x_{i1} + \cdots + \xi_{15}x_{i5} + \xi_{16}y_i$ and $f_{i2} = \xi_{20} + \xi_{21}x_{i1} + \cdots + \xi_{25}x_{i5} + \xi_{26}y_i + \xi_{27}r_{i1}$. Then, we considered a global perturbation

$$p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}, \boldsymbol{\omega}) = p(r_{i2}|r_{i1}, \mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}_2, \boldsymbol{\omega})p(r_{i1}|\mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}_1, \boldsymbol{\omega}) \tag{3.26}$$

$$p(r_{i1}|\mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}_1, \boldsymbol{\omega}) = \frac{\exp[r_{i1}(f_{i1} + \omega_1 z_{i1} + \omega_2 z_{i2})]}{1 + \exp(f_{i1} + \omega_1 z_{i1} + \omega_2 z_{i2})},$$

$$p(r_{i2}|r_{i1}, \mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}_2, \boldsymbol{\omega}) = \frac{\exp[r_{i2}(f_{i2} + \omega_3 z_{i1} + \omega_4 z_{i2})]}{1 + \exp(f_{i2} + \omega_3 z_{i1} + \omega_4 z_{i2})}.$$

The principal eigenvalue of $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}^0)}$ was 0.11, far smaller than the weighted chi-squared .05 cut-off point. This may suggest that the missing data mechanism is likely to be MAR.

In fitting the model using (3.25), the large value of the estimate for ξ_{26} indicated that the missingness of x_2 might depend on the response, whereas the estimates for all other ξ's were nonsignificant. Thus, we dropped the y_i term in f_{i2} of (3.25), leading to $f_{i2} = \xi_{20} + \xi_{21}x_{i1} + \cdots + \xi_{25}x_{i5} + \xi_{27}r_{i1}$. Then we used the global perturbation in (3.26) with

$$p(r_{i1}|\mathbf{x}_i, y_i, \boldsymbol{\xi}_1, \boldsymbol{\omega}) = \frac{\exp(r_{i1}f_{i1})}{1+\exp(f_{i1})} \text{ and } p(r_{i2}|r_{i1}, \mathbf{x}_i, y_i, \boldsymbol{\xi}_2, \boldsymbol{\omega}) = \frac{\exp[r_{i2}(f_{i2}+\omega y_i)]}{1+\exp(f_{i2}+\omega y_i)}.$$

It turned out that $\mathrm{FI}_{f_{lr}(\omega^0)}$ was 4.51, which is larger than the chi-squared .05 cut-off point. This suggests that the missingness of x_2 may depend on the response.

Furthermore, to assess the linear relationship between the response and the covariates (z_1, z_2), we employed a global perturbation scheme as follows:

$$y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 x_{i1} + \cdots + \beta_7 x_{i5} + \sum_{j=1}^{m+3} \omega_j B_j(z_{i1}) + \sum_{j=1}^{m+3} \omega_{j+m+3} B_j(z_{i2}) + \epsilon_i,$$

where the $B_j(z)$ are truncated polynomials of order $2-4$ given by z^2, z^3, z^4, $(z-k_1)_+^4, \cdots, (z-k_m)_+^4$, where $k_1, \cdots, k_m$ are the $m = 3$ prefixed knots. The principal eigenvalue of $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega})}$ was 3.44, which was not statistically significant at the 5% significance level ($p-$value $= 0.65$). Thus, the fitted model appears to be robust to this global perturbation scheme.

3.5.2 Liver dancer data

To further illustrate our proposed methods, we revisit the liver cancer data as introduced in Section 3.1 (Ibrahim, Chen, and Lipsitz, 1999). We are interested in how the number of cancerous liver nodes (y) when entering the trials is predicted by six other baseline characteristics: time since diagnosis of the disease (in weeks) (z_1); two biochemical markers (each classified as normal or abnormal), alpha fetoprotein (z_2) and anti-hepatitis antigen (z_3); associated jaundice (yes, no) (x_1); body mass index (weight in kilograms divided by the square of height in meters) (x_2); and age (in years) (x_3).

We used a Poisson regression model, $p(y_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}) \propto \exp[y_i(\mathbf{v}_i^T\boldsymbol{\beta}) - \exp(\mathbf{v}_i^T\boldsymbol{\beta})]$, where $\mathbf{v}_i^T = (1, x_{i1}, x_{i2}, x_{i3}, z_{i1}, z_{i2}, z_{i3})$ is the 1×7 vector of covariates including an intercept, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_6)^T$ are the corresponding regression coefficients. Logarithm of the time since diagnosis was used to achieve approximate normality. Since only $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3})$ has missing values, we need to consider a joint distribution only for these covariates. Because z_{i2} and z_{i3} were both dichotomous, we used logistic regressions. Thus,

$$p(z_{i1}, z_{i2}, z_{i3}|\mathbf{x}_i, \boldsymbol{\alpha}) = p(z_{i3}|z_{i1}, z_{i2}, \mathbf{x}_i, \boldsymbol{\alpha}_3) \times p(z_{i2}|z_{i1}, \mathbf{x}_i, \boldsymbol{\alpha}_2) \times p(z_{i1}|\mathbf{x}_i, \boldsymbol{\alpha}_1),$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3)$ and $(z_{i3}|z_{i1}, z_{i2}, \mathbf{x}_i)$ is a logistic regression with probability of success

$$p(z_{i3} = 1|z_{i1}, z_{i2}, \mathbf{x}_i, \boldsymbol{\alpha}_3) = \frac{\exp(\alpha_{30} + \alpha_{31}z_{i1} + \alpha_{32}z_{i2} + \boldsymbol{\alpha}_{3x}^T\mathbf{x}_i)}{1 + \exp(\alpha_{30} + \alpha_{31}z_{i1} + \alpha_{32}z_{i2} + \boldsymbol{\alpha}_{3x}^T\mathbf{x}_i)},$$

and $\boldsymbol{\alpha}_{3x}^T = (\alpha_{33}, \alpha_{34}, \alpha_{35})$. Similarly,

$$p(z_{i2} = 1|z_{i1}, \mathbf{x}_i, \boldsymbol{\alpha}_2) = \frac{\exp(\alpha_{20} + \alpha_{21}z_{i1} + \boldsymbol{\alpha}_{2x}^T\mathbf{x}_i)}{1 + \exp(\alpha_{20} + \alpha_{21}z_{i1} + \boldsymbol{\alpha}_{2x}^T\mathbf{x}_i)},$$

and $\boldsymbol{\alpha}_{2x}^T = (\alpha_{22}, \alpha_{23}, \alpha_{24})$. In addition, we took a normal distribution for the missing covariate z_1, specifically, $z_{i1} \sim N(\alpha_{11}, \alpha_{12})$ and $\boldsymbol{\alpha}_1^T = (\alpha_{11}, \alpha_{12})$. We assumed that the missing covariates are MAR and estimated $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ using the EM algorithm.

To detect the influential cases, we employed a perturbation to simultaneously perturb the missing covariates z_{i1}, z_{i2} and z_{i3} such that $y_i = \beta_0 + \beta_1(z_{i1} + \omega_i) + \beta_2(z_{i2} + \omega_i) + \beta_3(z_{i3} + \omega_i) + \cdots + \beta_6 x_{i3} + \epsilon_i$. Both $\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0), \mathbf{e}_i}$ and $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}^0), \mathbf{e}_i}$ indicated that cases 10, 15, 65, and 160 were very influential for this perturbation (Figure 3.4 a, b). Then we employed a perturbation to the distribution of z_{i1} such that $z_{i1} \sim N(\alpha_{11} + \omega_i, \alpha_{12}), i = 1, \cdots n$, and both influence measures detected case 131 to be influential for the distributional assumption of z_{i1} (Figure 3.4 c, d). These findings confirmed the suspected cases reported in Table 2.1.

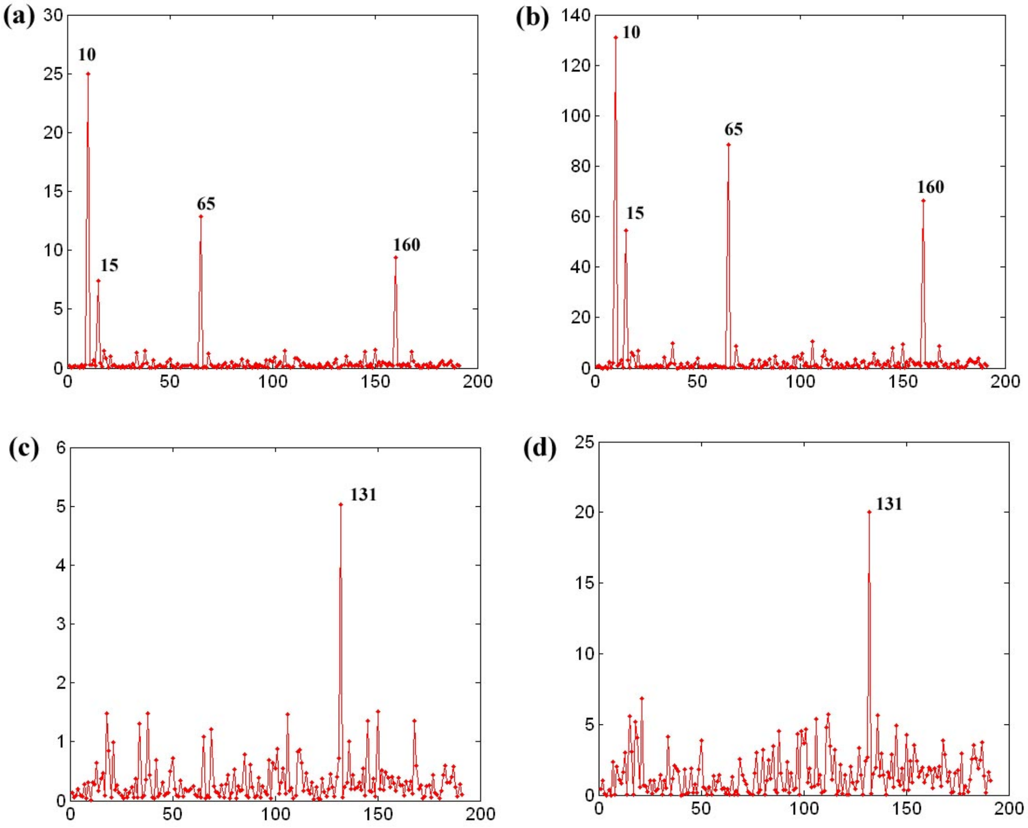

Figure 3.4: Index plots of influence measures for liver cancer data: (a) $\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0),\mathbf{e}_i}$ and (b) $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}^0),\mathbf{e}_i}$ for the missing covariate perturbation; (c) $\mathrm{FI}_{\hat{\boldsymbol{\eta}}_o(\boldsymbol{\omega}^0),\mathbf{e}_i}$ and (d) $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}^0),\mathbf{e}_i}$ for the perturbation to the distribution of z_{i1}.

Next, we examined the functional form of the missing data mechanism given by

$$p(\mathbf{r}_i|\mathbf{x}_i, y_i, \boldsymbol{\xi}) = p(r_{i3}|r_{i1}, r_{i2}, \mathbf{x}_i, y_i, \boldsymbol{\xi}_2) \times p(r_{i2}|r_{i1}, \mathbf{x}_i, y_i, \boldsymbol{\xi}_2) \times p(r_{i1}|\mathbf{x}_i, y_i, \boldsymbol{\xi}_1),$$

$$p(r_{i1}|y_i, \mathbf{x}_i, \boldsymbol{\xi}_1) = \frac{\exp(r_{i1} f_{i1})}{1 + \exp(f_{i1})}, \tag{3.27}$$

$$p(r_{i2}|r_{i1}, y_i, \mathbf{x}_i, \boldsymbol{\xi}_2) = \frac{\exp(r_{i2}f_{i2})}{1+\exp(f_{i2})}, \tag{3.28}$$

$$p(r_{i3}|r_{i1}, r_{i2}, y_i, \mathbf{x}_i, \boldsymbol{\xi}_3) = \frac{\exp(r_{i3}f_{i3})}{1+\exp(f_{i3})}, \tag{3.29}$$

in which $f_{i1} = \xi_{10} + \xi_{11}x_{i1} + \xi_{12}x_{i2} + \xi_{13}x_{i3} + \xi_{14}y_i$, $f_{i2} = \xi_{20} + \xi_{21}x_{i1} + \xi_{22}x_{i2} + \xi_{23}x_{i3} + \xi_{24}y_i + \xi_{25}r_{i1}$, and $f_{i3} = \xi_{30} + \xi_{31}x_{i1} + \xi_{32}x_{i2} + \xi_{33}x_{i3} + \xi_{34}y_i + \xi_{35}r_{i1} + \xi_{36}r_{i2}$. Then, we considered a global perturbation for the missing mechanism:

$$p(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}, \boldsymbol{\omega}) = p(r_{i3}|r_{i1}, r_{i2}, \mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}_2, \boldsymbol{\omega})p(r_{i2}|r_{i1}, \mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}_2, \boldsymbol{\omega})p(r_{i1}|\mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}_1, \boldsymbol{\omega})$$

$$p(r_{i1}|\mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}_1, \boldsymbol{\omega}) = \frac{\exp[r_{i1}(f_{i1} + \omega_1 z_{i1} + \omega_2 z_{i2} + \omega_3 z_{i3})]}{1+\exp(f_{i1} + \omega_1 x_{i1} + \omega_2 x_{i2} + \omega_3 z_{i3})},$$

$$p(r_{i2}|r_{i1}, \mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}_2, \boldsymbol{\omega}) = \frac{\exp[r_{i2}(f_{i2} + \omega_4 z_{i1} + \omega_5 z_{i2} + \omega_6 z_{i3})]}{1+\exp(f_{i2} + \omega_4 z_{i1} + \omega_5 z_{i2} + \omega_6 z_{i3})},$$

$$p(r_{i3}|r_{i2}, r_{i1}, \mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}_2, \boldsymbol{\omega}) = \frac{\exp[r_{i3}(f_{i3} + \omega_7 z_{i1} + \omega_8 z_{i2} + \omega_9 z_{i3})]}{1+\exp(f_{i3} + \omega_7 z_{i1} + \omega_8 z_{i2} + \omega_9 z_{i3})}.$$

The principal eigenvalue of $\mathrm{FI}_{f_{lr}(\boldsymbol{\omega}^0)}$ (0.24) was quite small, which suggests that the missing data mechanism is likely to be MAR. Following the arguments in Zhu et al. (2007a), we considered a single-case perturbation for the missing mechanism as follows:

$$p(r_{i1}|\mathbf{x}_i, \mathbf{z}_i, y_i, \boldsymbol{\xi}_1, \boldsymbol{\omega}) = \frac{\exp[r_{i1}(f_{i1} + \omega_i(z_{i1}/s_{z1} + z_{i2}/s_{z2} + z_{i3}/s_{z3}))]}{1+\exp(f_{i1} + \omega_i(z_{i1}/s_{z1} + z_{i2}/s_{z2} + z_{i3}/s_{z3}))},$$

where s_{z1}, s_{z2}, and s_{z3} are the sample standard deviations for z_1, z_2, and z_3, respectively. Then, a similar perturbation was introduced for r_{i2} and r_{i3}. All perturbations revealed case 131 to be influential. However, the perturbation for r_{i3} revealed only case 65 as an influential case. The reason that cases 10, 15, and 160 did not stand out under the single-case perturbation for all missing covariates and case 65 did not stand out under the single-case perturbation for z_1 or z_2, is that: i) they all have very large values in the response, ii) large response values y_i tend to yield large values of $p(r_i = 1|\mathbf{x}_i, y_i, \boldsymbol{\xi})$ for all z_1, z_2, and z_3, iii) cases 10, 15, and 160 have no missing values in z_1, z_2, and z_3 so they fit (3.27), (3.28), and (3.29) well, whereas case 65 has no missing values in z_1 and z_2 so it fits (3.27) and (3.28) well.

3.6 Discussion

We have developed a general local influence methodology for carrying out sensitivity analyses in GLMs with MAR or NMAR covariate data. We have also proposed a novel methodology for choosing an appropriate perturbation scheme and examined several influence measures within this context. The simulation studies and the real dataset showed very promising results for the proposed methods. We emphasize again that in missing data problems, there is typically little information in the data regarding the form of the missing data mechanism, and the parametric assumption of the missing data mechanism itself is not 'testable' from the data. Thus, NMAR modeling should be viewed as a sensitivity analysis concerning a more complicated model. In this sense, it is not advisable to carry out formal tests directly to assess and compare MAR and NMAR models. Future work in this area includes extending these methodologies to the Cox proportional hazards model with right censored survival data and missing covariates, as well as to parametric and semiparametric models for longitudinal data with MAR or NMAR response and/or covariate data.

Chapter 4

Adjusted Exponetially Tilted Empirical Likelihood with Applications to Neuroimaging Data

4.1 Introduction

In this paper, we will develop three statistical methods for the analysis of neuroimaging data from cross-sectional and longitudinal studies. The first methodology is to develop an adjusted exponentially tilted empirical likelihood (ETEL) procedure and investigate its associated estimators. The adjusted ETEL is a nonparametric method that is built on a set of estimating equations, and thus it avoids the parametric assumptions in the linear mixed model. This feature is desirable for the analysis of real neuroimaging data, including brain morphological measures, because the distribution of the univariate (or multivariate) neuroimaging measurements often deviates from the Gaussian distribution (Ashburner and Friston, 2000; Salmond et al., 2002; Luo and Nichols, 2003).

Statistically, the adjusted ETEL improves several alternative methods for estimating equations, including empirical likelihood (EL), generalized estimating equations (GEE), and generalized method of moments (GMM), both empirically and theoretically. The maximum adjusted ETEL estimator has similar asymptotic properties as the maximum EL estimator under the correctly specified estimating equations, whereas the maximum EL may cease to have regular $\sqrt{n}$ convergence when estimating functions are misspecified (Schennach, 2007; Qin and Lawless, 1994). The adjusted ETEL can produce efficient estimators by incorporating more estimating equations than the number of parameters, whereas the GEE cannot (Lai and Small, 2007; Liang and Zeger, 1986). This advantage becomes very obvious when handling longitudinal data with various types of time-dependent covariates (Lai and Small, 2007). The GMM requires an estimator of the weighting matrix, which often causes problems in finite samples, whereas the adjusted ETEL avoids such problems (Newey and Smith, 2004; Kitamura, 2006; Schennach, 2007).

Secondly, we propose an adjusted ETEL ratio statistic to test linear hypotheses of unknown parameters, such as associations between brain structure and function, and covariates of interest, such as diagnostic groups and severity of disease. Theoretically, the adjusted ETEL ratio statistic, denoted by LR_{Aetel}, is approximately χ^2 distributed, while empirically it has better finite coverage based on the χ^2 distribution, so that it provides a partial solution to the under-coverage problem for the unadjusted ETEL ratio statistic (Tsao, 2004; Owen, 2001; Chen et al., 2007). Although bootstrap calibration and Bartlett corrections have been proposed, applying these methods is essentially infeasible for large neuroimaging datasets. An accurate χ^2 approximation is also important for controlling the family-wise error rate and false discovery rate (FDR) across the entire brain region, because all multiple testing methods require an accurate estimate of the $p-$value of the test statistic at each voxel of the brain region (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). Finally, LR_{Aetel} is guaranteed to have a solution and thus it avoids the issue of $LR_{etel} = +\infty$ (Tsao, 2004; Chen et al., 2007). This issue has a serious drawback in neuroimaging analysis. The significant voxels, which have a high probability of having $LR_{etel} = +\infty$ are the most important areas of interest in most neuroimaging studies, whereas the value $+\infty$ does not provide any qualitative measure of the significance level in these voxels. In particular, when two voxels have $LR_{etel} = +\infty$, it is impossible to judge whether one voxel is more significant than the other.

Thirdly, we develop goodness of fit statistics for testing potential misspecification of the estimating equations. We first use the adjusted ETEL ratio statistic, LR_{GF}, to test the validity of overidentifying restrictions, while controlling for the type I error (Qin and Lawless, 1994). This test statistic follows a χ^2 distribution asymptotically, and thus the use of this statistic provides a computationally simple specification test. If we reject the notion that all estimating equations are correctly specified, it is then interesting to find out which estimating equations are misspecified; however, the LR_{GF} cannot identify which estimating equations are misspecified. We develop a new marked empirical process of the estimating equations (EPEE), based on projections, which is simple to compute. Although the marked process of residuals has been developed under the GEE framework, the residual process suffers from the issue of the curse of dimensionality and the subjective choice of bandwidths (Escanciano, 2006; Lin, Wei, and Ying, 2002). In addition, the residual process cannot reveal the specific type of time-dependent covariates, which may be useful for increasing efficiency of parameter estimates (Lai and Small, 2007). Our marked EPEE can be used to detect which estimating equations are misspecified. Moreover, we can use the maximum of the standardized marked EPEEs of all estimating equations to test the overall validity of the restrictions, and this statistic can be used even when the number of estimating functions equals the number of parameters in situations where the adjusted ETEL ratio statistic is not applicable. Although the asymptotic null distribution of the new marked EPEE depends on the data generating process, we develop a computationally efficient procedure to approximate the $p-$value.

Section 4.2 of this paper presents the three new statistical methods just described. In Section 4.3, we conduct simulation studies to examine the finite sample performance of the maximum adjusted ETEL estimator, the adjusted ETEL ratio statistic, and the marked EPEE. Section 4.4 illustrates an application of the proposed methods in a neuroimaging dataset. We present concluding remarks in Section 4.5.

4.2 Adjusted exponetially tilted empirical likelihood

4.2.1 Data and estimating equations

We consider a longitudinal study of imaging data with n subjects, where a $q \times 1$ covariate $\mathbf{x}_{ij}$ (e.g., age, gender, height, and brain volume) is obtained for the i-th subject at the j-th time point t_{ij} for $i = 1, \cdots, n$ and $j = 1, \cdots, m_i$. Thus, there are at least $\sum_{i=1}^{n} m_i = N$ images in the study. Based on each image, we observe or compute neuroimaging measures, denoted by $\mathbf{Y}_i = \{\mathbf{y}_{ij}(d) : d \in \mathcal{D}, j = 1, \cdots, m_i\}$, across all m_i time points from the i-th subject, where d represents a voxel on $\mathcal{D}$, a specific brain region. The imaging measure $\mathbf{y}_{ij}(d)$ at each voxel d can be either univariate or multivariate. For example, gray matter density and signed Euclidean distance of cortical/subcortical surfaces are univariate measures, whereas the spherical harmonic shape description (SPHARM) of subcortical surfaces is a three dimensional MRI measure at each point (Styner and Gerig, 2003). For notational simplicity, we assume that the $\mathbf{y}_{ij}(d)$ are univariate measures.

We temporarily drop voxel d from our notation. At a specific voxel d in the brain region, $\mathbf{z}_i = \{(\mathbf{y}_{ij}, \mathbf{x}_{ij}) : j = 1, \cdots, m_i\}$ is independent and satisfies a moment condition

$$E\{\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})\} = 0, \qquad \text{for} \;\; i = 1, \cdots, n, \tag{4.1}$$

where $\boldsymbol{\theta}$ is a $p \times 1$ vector, $\mathbf{g}(\cdot, \cdot)$ is an $r \times 1$ vector of known functions with $r \geq p$ and E denotes the expectation with respect to the true distribution of all $\mathbf{z}_i$'s. Equation (4.1) is often referred to as a set of unbiased estimating equations or moments model (Qin and Lawless, 1994; Hansen, 1982).

The moments model (4.1) is more general than the standard linear regression model

(LRM), which has been a standard method in the analysis of neuroimaging data from cross-sectional studies (Worsley et al., 2004). Specifically, the LRM assumes that $\mathbf{y}_{i1} = \mathbf{x}_{i1}^{'}\boldsymbol{\beta} + \epsilon_{i1}$ and $m_i = 1$ for all $i = 1, \cdots, n$, where $\epsilon_{i1} \sim N(0, \sigma^2)$. In this case, $\boldsymbol{\theta} = (\boldsymbol{\beta}^{'}, \sigma^2)^{'}$, and $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$ can be set to $(\mathbf{x}_{i1}^{'}(\mathbf{y}_{i1} - \mathbf{x}_{i1}^{'}\boldsymbol{\beta}), (\mathbf{y}_{i1} - \mathbf{x}_{i1}^{'}\boldsymbol{\beta})^2 - \sigma^2)^{'}$. If it is assumed that ϵ_{i1} has zero mean and heterogeneous variance σ_i^2, then we may set $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$ to $\mathbf{x}_{i1}^{'}(\mathbf{y}_{i1} - \mathbf{x}_{i1}^{'}\boldsymbol{\beta})$, which also leads to a consistent estimate of $\boldsymbol{\beta}$.

For longitudinal data, although the measurements from different subjects are independent, those within the same subject may be highly correlated. The generalized estimating equations (GEE) assume a working covariance matrix for $\mathbf{y}_i = (\mathbf{y}_{i1}, \cdots, \mathbf{y}_{im_i})^{'}$ given by V_i. Let $E(\mathbf{y}_i) = \mu_i(\boldsymbol{\beta}) = (\mu_{i1}(\boldsymbol{\beta}), \cdots, \mu_{im_i}(\boldsymbol{\beta}))^{'}$ and $D_i(\boldsymbol{\beta}) = \partial\mu_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. Under the assumption that $E\{D_i(\boldsymbol{\beta})^{'}V_i^{-1}[\mathbf{y}_i - \mu_i(\boldsymbol{\beta})]\} = 0$, Liang and Zeger (1986) proposed to use an estimator, denoted by $\hat{\boldsymbol{\beta}}_{gee}$, which solves a set of GEEs as follows:

$$G(\boldsymbol{\beta}) = \sum_{i=1}^{n} D_i(\boldsymbol{\beta})^{'}V_i^{-1}[\mathbf{y}_i - \mu_i(\boldsymbol{\beta})] = \mathbf{0}. \tag{4.2}$$

Following Lai and Small (2007), we classify the time-dependent covariates into three types. The l-th covariate, $\mathbf{x}_{ij,l}$, is a type I time-dependent covariate if

$$E\{\partial_{\beta_l}\mu_{is}(\boldsymbol{\beta})[\mathbf{y}_{ij} - \mu_{ij}(\boldsymbol{\beta})]\} = 0 \ \text{ for all } \ s, j = 1, \cdots, m_i, \tag{4.3}$$

where $\partial_{\beta_l} = \partial/\partial\beta_l$. A sufficient condition for type I covariates is that $E[y_{ij}|\mathbf{x}_{ij}] = E[y_{ij}|\mathbf{x}_{i1}, \cdots, \mathbf{x}_{im_i}]$ (Lai and Small, 2007). If all time-dependent covariates are of type I, we can set $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = D_i(\boldsymbol{\beta})^{'}V_i^{-1}[\mathbf{y}_i - \mu_i(\boldsymbol{\beta})]$ and show that $E[\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})] = 0$. Particularly, if V_i is the true covariance matrix of $\mathbf{y}_i$, then the estimator $\hat{\boldsymbol{\beta}}_{gee}$ is an efficient estimator. However, $\hat{\boldsymbol{\beta}}_{gee}$ is inefficient under a misspecified V_i. To increase the efficiency, we may choose several candidate working covariance matrices $M_i^{(1)}, \cdots, M_i^{(s_0)}$ and assume $V_i^{-1} = \sum_{k=1}^{s_0} \alpha_k M_i^{(k)}$ for some unknown constants α_k (Qu et al., 2000). Then, following Qu et al. (2000), we consider a set of estimating equations given by

$$\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \begin{pmatrix} D_i(\boldsymbol{\beta})' M_i^{(1)}[\mathbf{y}_i - \mu_i(\boldsymbol{\beta})] \\ \vdots \\ D_i(\boldsymbol{\beta})' M_i^{(s_0)}[\mathbf{y}_i - \mu_i(\boldsymbol{\beta})] \end{pmatrix} \quad \text{for} \quad i = 1, \cdots, n. \tag{4.4}$$

In this case, the number of functions in $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$ is $s_0 q > q$, when $s_0 > 1$.

The time-dependent covariate $\mathbf{x}_{ij,l}$ is of type II if

$$E\{\partial_{\beta_l} \mu_{is}(\boldsymbol{\beta})[\mathbf{y}_{ij} - \mu_{ij}(\boldsymbol{\beta})]\} = 0 \ \text{ for all } \ s \geq j, \ j = 1, \cdots, m_i. \tag{4.5}$$

A sufficient condition for all time-dependent covariates to be of type II is given by

$$p(\mathbf{x}_{i(t+1)}, \cdots, \mathbf{x}_{im_i} | \mathbf{y}_{it}, \mathbf{x}_{it}) = p(\mathbf{x}_{i(t+1)}, \cdots, \mathbf{x}_{im_i} | \mathbf{x}_{it}) \ \text{ for } \ t = 1, \cdots, m_i. \tag{4.6}$$

If all covariates are time-dependent and of type II, we can set $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = D_i(\boldsymbol{\beta})'[\mathbf{y}_i - \mu_i(\boldsymbol{\beta})]$, in which an independent working covariance matrix is used. However, the estimator $\hat{\boldsymbol{\beta}}_{gee}$ based on the independent working correlation matrix is inefficient, since we do not use the information contained in $E\{\partial_{\boldsymbol{\beta}} \mu_{is}(\boldsymbol{\beta})[\mathbf{y}_{ij} - \mu_{ij}(\boldsymbol{\beta})]\} = 0$ for all $s > j$. To increase the efficiency of the estimate, we choose a set of lower triangular matrices $L_i^{(1)}, \cdots, L_i^{(s_0)}$, and then we consider estimating equations given by

$$\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \begin{pmatrix} D_i(\boldsymbol{\beta})' L_i^{(1)}[\mathbf{y}_i - \mu_i(\boldsymbol{\beta})] \\ \vdots \\ D_i(\boldsymbol{\beta})' L_i^{(s_0)}[\mathbf{y}_i - \mu_i(\boldsymbol{\beta})] \end{pmatrix} \quad \text{for} \quad i = 1, \cdots, n. \tag{4.7}$$

In this case, the number of functions in $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$ is $s_0 q > q$, when $s_0 > 1$. Suppose that $m_1 = \cdots = m_n$, we can set $s_0 = m_1(m_1 + 1)/2$ and $L_i^{(b)} = \mathbf{e}_s \mathbf{e}_j$, where $\mathbf{e}_s$ is a $q \times 1$ vector with sth component 1 and 0 otherwise. Thus, similar to Lai and Small (2007), we are able to pick $\partial_{\beta_l} \mu_{is}(\boldsymbol{\beta})[\mathbf{y}_{ij} - \mu_{ij}(\boldsymbol{\beta})]$ for all $s \geq j$.

The time-dependent covariate $\mathbf{x}_{ij,l}$ is of type III if

$$E\{\partial_{\beta_l}\mu_{is}(\boldsymbol{\beta})[\mathbf{y}_{ij} - \mu_{ij}(\boldsymbol{\beta})]\} \neq 0 \text{ for some } s > j. \tag{4.8}$$

If all covariates are time-dependent and of type III, we must choose V_i as a diagonal matrix. For instance, if $V_i = \mathbf{I}_i$, where $\mathbf{I}_i$ is an identity matrix, then $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = D_i(\boldsymbol{\beta})^{'}[\mathbf{y}_i - \mu_i(\boldsymbol{\beta})]$. If we further assume the specific form for the variances of all y_{ij}, then we may set $V_i = \text{diag}(\text{Cov}(\mathbf{y}_i))$.

An overall strategy to analyze models with time-dependent covariates is first to assume that the time-dependent covariates are of type III. Then we test whether the time-dependent covariates are of type II, and if the test is not rejected, we can go on to test if they are of type I. Once the type of all the time-dependent covariates is decided, we use the corresponding estimating equations. See Section 4.4 for more details.

4.2.2 Adjusted exponetially tilted empirical likelihood

We consider a nonparametric method, called adjusted ETEL, to carry out statistical inference about $\boldsymbol{\theta}$ based on a set of estimating equations $\{\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) : i = 1, \cdots, n\}$ (Schennach, 2007); see for example $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$ in equation (4.7). The adjusted ETEL is a combination of the empirical likelihood and the exponetially tilted method. We introduce

$$\mathbf{g}_{n+1}(\boldsymbol{\theta}) = -\frac{a_n}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}), \tag{4.9}$$

where $a_n = \max(1, \log(n)/2)$. Then, the maximum adjusted ETEL estimator, denoted by $\hat{\boldsymbol{\theta}}_{Aetel}$, minimizes a criterion given by

$$\min_{\boldsymbol{\theta}}\{-(n+1)^{-1}\sum_{i=1}^{n+1}\log(n+1)\hat{p}_i(\boldsymbol{\theta}))\},$$

where $\hat{p}_i(\boldsymbol{\theta})$ is the solution to

$$\min_{p_1,\cdots,p_{n+1}} (n+1)^{-1} \sum_{i=1}^{n+1} [(n+1)p_i] \log[(n+1)p_i]$$

subject to

$$\sum_{i=1}^{n+1} p_i = 1, \quad p_i \geq 0, \quad \text{and} \quad \sum_{i=1}^{n} p_i \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) + p_{n+1}\mathbf{g}_{n+1}(\boldsymbol{\theta}) = 0.$$

According to a duality theorem in convex analysis (Newey and Smith, 2004), $\hat{\boldsymbol{\theta}}_{Aetel}$ is also the solution to a saddle point problem

$$\hat{\boldsymbol{\theta}}_{Aetel} = \text{argmax}_{\boldsymbol{\theta}} \ell_{Aetel}(\boldsymbol{\theta}), \tag{4.10}$$

where $\ell_{Aetel}(\boldsymbol{\theta}) = (n+1)^{-1} \sum_{i=1}^{n+1} \log((n+1)\hat{p}_i(\boldsymbol{\theta}))$ and

$$\hat{p}_{n+1}(\boldsymbol{\theta}) = \frac{\exp(\hat{\mathbf{t}}(\boldsymbol{\theta})' \mathbf{g}_{n+1}(\boldsymbol{\theta}))}{T_{\mathbf{g}}(\boldsymbol{\theta})} \ \text{ and } \ \hat{p}_i(\boldsymbol{\theta}) = \frac{\exp(\hat{\mathbf{t}}(\boldsymbol{\theta})' \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}))}{T_{\mathbf{g}}(\boldsymbol{\theta})} \text{ for } i = 1, \cdots, n, \tag{4.11}$$

in which $T_{\mathbf{g}}(\boldsymbol{\theta}) = \sum_{j=1}^{n} \exp(\hat{\mathbf{t}}(\boldsymbol{\theta})' \mathbf{g}(\mathbf{z}_j, \boldsymbol{\theta})) + \exp(\hat{\mathbf{t}}(\boldsymbol{\theta})' \mathbf{g}_{n+1}(\boldsymbol{\theta}))$. In addition,

$$\hat{\mathbf{t}}(\boldsymbol{\theta}) = \text{argmax}_{\mathbf{t}} \{-\sum_{i=1}^{n} \exp(\mathbf{t}' \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})) - \exp(\mathbf{t}' \mathbf{g}_{n+1}(\boldsymbol{\theta}))\}.$$

We consider testing the linear hypotheses:

$$H_0 : R\boldsymbol{\theta} = \mathbf{b}_0 \quad \text{vs.} \quad H_1 : R\boldsymbol{\theta} \neq \mathbf{b}_0, \tag{4.12}$$

where R is a $c_0 \times p$ matrix of full row rank and $\mathbf{b}_0$ is a $c_0 \times 1$ specified vector. Most scientific questions in neuroimaging studies can be formulated into linear hypotheses, such as a comparison of brain regions across diagnostic groups and a detection of changes in brain regions across time. The adjusted ETEL ratio statistic for testing $R\boldsymbol{\theta} = \mathbf{b}_0$ can be constructed as follows:

$$LR_{Aetel} = -2(n+1)\{\sup_{\boldsymbol{\theta}:R\boldsymbol{\theta}=\mathbf{b}_0} \ell_{Aetel}(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta}} \ell_{Aetel}(\boldsymbol{\theta})\}. \tag{4.13}$$

Thus, to compute LR_{Aetel}, we also need to compute the maximum adjusted ETEL estimator, denoted by $\hat{\boldsymbol{\theta}}_{Aetel,0}$, subject to an additional constraint $R\boldsymbol{\theta} = \mathbf{b}_0$.

Under some conditions on $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$, we have the following theorem, whose detailed proof can be found in the appendix.

Theorem 1

If assumptions A-D in the Appendix are true, then we have

(i) $\sqrt{n}(\hat{\boldsymbol{\theta}}_{Aetel} - \boldsymbol{\theta}_0)$ *converges to* $\nu_0 = N(0, \Sigma)$ *in distribution, where* $\boldsymbol{\theta}_0$ *denotes the true value of* $\boldsymbol{\theta}$, *and* $\Sigma = (DV^{-1}D')^{-1}$,

$$D = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \partial_{\boldsymbol{\theta}} \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) \text{ and } V = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})^{\otimes 2};$$

(ii) under the null hypothesis H_0, LR_{Aetel} *converges to a* $\chi^2(c_0)$ *distribution;*

(iii) if $E[\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})] = 0$ *for all* i, *and* $r > p$, *then* $LR_{GF} = 2(n+1)\sup_{\boldsymbol{\theta}} \ell_{Aetel}(\boldsymbol{\theta})$ *is asymptotically* $\chi^2(r-p)$.

Theorem 1 establishes the asymptotic consistency and asymptotically normality of $\hat{\boldsymbol{\theta}}_{Aetel}$ and the asymptotic χ^2 distribution of LR_{Aetel}. Theorem 1 also shows that the adjusted ETEL have the same first-order asymptotic properties as ETEL (Schennach, 2007). It will be shown that the χ^2 approximation of the adjusted ETEL likelihood ratio statistics is quite precise, compared to the existing ETEL (Tsao, 2004; Owen, 2001). Providing a reliable p-value at each voxel is crucial for controlling the family-wise error rate and false discovery rate (FDR) across the entire brain region (Benjamini and Hochberg, 1995; Worsley et al., 2004).

4.2.3 Goodness of fit statistics

We are interested in testing the following hypotheses:

$$H_0 : (4.1) \text{ is true} \quad \text{versus} \quad H_1 : (4.1) \text{ is not true.} \tag{4.14}$$

We can use LR_{GF} to test the validity of overidentifying restrictions in order to control for the type I error (Qin and Lawless, 1994). As shown in Theorem 1, LR_{GF} follows a χ^2 distribution asymptotically and provides a computationally simple specification test. If we reject the null hypothesis H_0, it is then of interest to find out which estimating equations are misspecified. However, the LR_{GF} cannot pinpoint those misspecified estimating equations. This then motivates us to pursue some goodness of fit statistics to test the misspecification of each estimating equation.

To ease the notational complexity, we use the estimating equations for the cross-sectional studies to develop goodness of fit statistics to test the possible misspecification of each of the estimating equations. Specifically, we develop new marked empirical processes of the estimating equations (EPEE) based on projections. Moreover, we use the maximum of the standardized marked EPEEs of all estimating equations to test the overall validity of restrictions, and such statistics are applicable even when the number of estimating functions equals the number of parameters. Finally, we discuss the extension of the marked EPEEs to the estimating equations for longitudinal studies as discussed above.

For cross-sectional studies, note that $\mathbf{z}_i = (\mathbf{y}_{i1}, \mathbf{x}_{i1})$. For the $k-$th component of $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$, we are interested in testing the null hypothesis

$$H_{0k} : E[g_k(\mathbf{z}_i, \boldsymbol{\theta})|\mathbf{x}_{i1}] = 0 \ \text{ for some } \boldsymbol{\theta} \in \Theta \subset R^p, \tag{4.15}$$

against the alternative

$$H_{Ak} : P(E[g_k(\mathbf{z}_i, \boldsymbol{\theta})|\mathbf{x}_{i1}] \neq 0) > 0 \ \text{ for all } \boldsymbol{\theta} \in \Theta \subset R^p. \tag{4.16}$$

Note that equation (4.15) is only a sufficient condition for $E[g_k(\mathbf{z}_i, \boldsymbol{\theta})] = 0$; however, assumption (4.15) arises naturally within the framework of regression (Lin et al., 2002; Escanciano, 2006). Actually, H_{0k} is equivalent to an infinite number of unconditional moment restrictions

$$E[g_k(\mathbf{z}_i, \boldsymbol{\theta})\omega(\mathbf{x}_{i1}, \mathbf{x})] = 0 \ \text{ for any } \ \mathbf{x}, \tag{4.17}$$

where $\omega(\mathbf{x}_{i1}, \mathbf{x})$ is a parametric family such that the above equivalence holds. Examples of such families include $\omega(\mathbf{x}_{i1}, \mathbf{x}) = 1(\mathbf{x}_{i1} \leq \mathbf{x})$ (the indicator function) and $\omega(\mathbf{x}_{i1}, \mathbf{x}) = \sin(\mathbf{x}_{i1}^{'}\mathbf{x})$, among many others (Escanciano, 2006; Bierens and Ploberger, 1997). We are now led to the following lemma:

Lemma 1

A neccessary and sufficient condition for (4.15) to hold is that for any vector $\boldsymbol{\eta} \in R^q$,

$$E\{\mathbf{g}_k(\mathbf{z}_i, \boldsymbol{\theta}_0)|\boldsymbol{\eta}^{'}\mathbf{x}_{i1}\} = 0.$$

An important implication of Lemma 1 is that consistent tests for H_{0k} can be based on one-dimensional projections. Specifically, H_{0k} is equivalent to

$$E\{g_k(\mathbf{z}_i, \boldsymbol{\theta})\mathbf{1}(\boldsymbol{\eta}^{'}\mathbf{x}_{i1} \leq u)\} = 0$$

almost everywhere for $(\boldsymbol{\eta}, u) \in \Pi$ and some $\boldsymbol{\theta} \in \Theta \in R^p$, where $\Pi = R^q \times [-\infty, \infty]$ is the nuisance parameter space. Without loss of generality, we can constrain $\boldsymbol{\eta}$ to be within $B_q = \{\boldsymbol{\eta} \in R^q : \|\eta\| = 1\}$. The test that we consider here rejects the null hypothesis for large values of the stardardized sample analogue of $E\{g_k(\mathbf{z}_i, \boldsymbol{\theta}_0)\mathbf{1}(\boldsymbol{\eta}^{'}\mathbf{x}_{i1} \leq u)\}$. This motivates us to construct the marked EPEE given by

$$\mathrm{GF}_k(\eta, u; \hat{\boldsymbol{\theta}}_{Aetel}) = n^{-1/2} \sum_{i=1}^{n} g_k(\mathbf{z}_i, \hat{\boldsymbol{\theta}}_{Aetel}) \mathbf{1}(\boldsymbol{\eta}^{'} \mathbf{x}_{i1} \leq u). \tag{4.18}$$

Then, we construct a Cramér-von Mises (CvM) statistic to formally test H_{0k}. The CvM test for the k-th estimating equation is

$$\mathrm{PCvM}_{n,k} = \int_{\Pi} [\mathrm{GF}_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}}_{Aetel})]^2 F_{n,\boldsymbol{\eta}}(du) d\boldsymbol{\eta}, \tag{4.19}$$

where $F_{n,\boldsymbol{\eta}}(u)$ is the empirical distribution function of the projected regressors $\{\boldsymbol{\eta}^{'} \mathbf{x}_{i1}\}_{i=1}^{n}$ and $d\boldsymbol{\eta}$ the uniform density on the unit sphere, since we treat all directions as equally important. We reject the null hypothesis for large values of $\mathrm{PCvM}_{n,k}$. Computationally, it is straightforward to compute $\mathrm{PCvM}_{n,k}$; see the appendix for details.

Theorem 2

If assumptions A-E in the Appendix and the null hypothesis H_{0k} in (4.15) are true, then for each $k = 1, \cdots, r$, we have

(i) $GF_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}}_{Aetel})$ converges weakly to $GF_k(\boldsymbol{\eta}, u)$, a Gaussian process with zero mean and covariance function $\Sigma_k((\boldsymbol{\eta}_1, u_1), (\boldsymbol{\eta}_2, u_2))$, which is the limit of

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} [\{g_k(\mathbf{z}_i)\mathbf{1}(\boldsymbol{\eta}_1^{'}\mathbf{x}_{i1} \leq u_1) - D_k(\boldsymbol{\eta}_1, u_1)\mathbf{g}(\mathbf{z}_i)\}\{g_k(\mathbf{z}_i)\mathbf{1}(\boldsymbol{\eta}_2^{'}\mathbf{x}_{i1} \leq u_2) - D_k(\boldsymbol{\eta}_2, u_2)\mathbf{g}(\mathbf{z}_i)\}]$$

for any $(\boldsymbol{\eta}_1, u_1)$ and $(\boldsymbol{\eta}_2, u_2)$, where $g_k(\mathbf{z}_i) = g_k(\mathbf{z}_i, \boldsymbol{\theta}_0)$, $\mathbf{g}(\mathbf{z}_i) = \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}_0)$, and

$$D_k(\boldsymbol{\eta}, u) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \{\mathbf{1}(\boldsymbol{\eta}^{'}\mathbf{x}_{i1} \leq u) \partial_{\boldsymbol{\theta}} g_k(\mathbf{z}, \boldsymbol{\theta}_0)^{'}\} \Sigma D V^{-1}.$$

(ii) $PCvM_{n,k}$ converges weakly to $PCvM_{\infty,k} = \int_{\Pi} [GF_k(\boldsymbol{\eta}, u; \boldsymbol{\theta}_0)]^2 F_{\boldsymbol{\eta}}(du) d\boldsymbol{\eta}$ for $k = 1, \cdots, r$, where $F_{\boldsymbol{\eta}}(u)$ is the true cumulative distribution function of u.

Theorem 2 formally characterizes the asymptotic distributions of the stochastic processes of interest $\{\mathrm{GF}_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}})\}_1^r$, which form the foundation for using $\{\mathrm{PCvM}_{n,k}\}_1^r$ as test statistics, and $\mathrm{PCvM}_{n,k}$. Then, we can use $\mathrm{PCvM}_{n,k}$ to test potential misspecification of each estimating equation. By combining multiple estimating equations, we can further test the specific type of time-dependent covariates.

Now we study the asymptotic distribution of $\mathrm{PCvM}_{n,k}$ under a sequence of local alternatives converging to the null hypothesis H_{0k} at a parametric rate $n^{-1/2}$. We consider the local alternatives as follows:

$$H_{Ak,n} : E[g_k(\mathbf{z}_i, \boldsymbol{\theta}_0)|\mathbf{x}_{i1}] = n^{-1/2}a(\mathbf{x}_{i1}), \quad \text{a.s.,} \quad 1 \leq i \leq n, \tag{4.20}$$

where $a(\mathbf{x}_{i1})$ is an integrable function such that $Pr(a(\mathbf{x}_{i1}) = 0) < 1$.

Theorem 3

Under the hypothesis $H_{Ak,n}$ and assumptions A-F in the Appendix, we have the following results:

(i) $\sqrt{n}(\hat{\boldsymbol{\theta}}_{Aetel} - \boldsymbol{\theta}_0)$ converges in distribution to $\nu_0 + A_{1k}$, where ν_0 is the same normal distribution in Theorem 1, and A_{1k} is a constant depending on $a(\mathbf{x})$;

(ii) $GF_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}}_{Aetel})$ converges in distribution to $GF_k(\boldsymbol{\eta}, u)+A_{2k}(\boldsymbol{\eta}, u)$, where $A_{2k}(\boldsymbol{\eta}, u)$ is a deterministic function defined in the appendix.

(iii) $PCvM_{n,k}$ converges in distribution to $\int_\Pi |GF_k(\boldsymbol{\eta}, u) + A_{2k}(\boldsymbol{\eta}, u)|^2 F_{\boldsymbol{\eta}}(du)d\boldsymbol{\eta}$.

Since the null hypothesis H_0 in (4.14) states that all estimating equations are correctly specified, we can combine all statistics $\mathrm{PCvM}_{n,k}$ and test whether or not any of the $\mathrm{PCvM}_{n,k}$ show any patterns beyond random fluctuation. Because $\{g_k(\mathbf{z}, \hat{\boldsymbol{\theta}}) : k = 1, \ldots, r\}$ may be quite different, we must adjust such differences between the variances of $g_k(x, \hat{\boldsymbol{\theta}})$ before we combine the statistics $\mathrm{PCvM}_{n,k}$. Thus, we construct another maximum statistic as follows:

$$\mathrm{PCvM}_n = \max_{1 \leq k \leq r} \{\mathrm{PCvM}_{n,k}/(S_{n,11}(\hat{\boldsymbol{\theta}}_{Aetel}))_{k,k}\}, \tag{4.21}$$

where $(S_{n,11}(\hat{\boldsymbol{\theta}}_{Aetel}))_{k,k}$ is the $k-$th diagonal term of $S_{n,11}(\hat{\boldsymbol{\theta}}_{Aetel})$ as defined in the appendix. The continuous mapping theorem yields that PCvM_n converges weakly to $\max_{1\leq k\leq r}[\text{PCvM}_{\infty,k}\{(S_{11})_{k,k}\}^{-1/2}]$, where $(S_{11})_{k,k}$ is the $k-$th diagonal element of S_{11}.

We can devise a resampling method to approximate the $p-$values of $\{\text{PCvM}_{n,k}\}_1^r$ while correcting for multiple comparisons and accounting for the correlations among the $\{\text{PCvM}_{n,k}\}_1^r$.

4.2.4 Extensions to longitudinal data

We discuss how to extend $\text{GF}_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}}_{Aetel})$ and $\text{PCvM}_{n,k}$ to longitudinal data with time-dependent covariates. We start with the simplest case, in which all covariates are time-dependent and of type III. Since $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \sum_{j=1}^{m_i} \partial_{\boldsymbol{\beta}} \mu_{ij}(\boldsymbol{\beta})[\mathbf{y}_{ij} - \mu_{ij}(\boldsymbol{\beta})]$, the $k-$th component of $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$ is given by $g_k(\mathbf{z}_i, \boldsymbol{\theta}) = \sum_{j=1}^{m_i} \partial_{\boldsymbol{\beta}_k} \mu_{ij}(\boldsymbol{\beta})[\mathbf{y}_{ij} - \mu_{ij}(\boldsymbol{\beta})]$. A sufficient condition for $E[g_k(\mathbf{z}, \boldsymbol{\theta})] = 0$ is that

$$E\{\mathbf{y}_{ij} - \mu_{ij}(\boldsymbol{\beta})|\mathbf{x}_{ij}\} = 0 \quad \text{for all} \;\; j = 1, \cdots, m_i; i = 1, \cdots, n. \tag{4.22}$$

Similar to Lemma 1, a necessary and sufficient condition for the expectation to hold in (4.22) is that for any vector $\boldsymbol{\eta} \in R^q$,

$$E\{\mathbf{y}_{ij} - \mu_{ij}(\boldsymbol{\beta})|\boldsymbol{\eta}'\mathbf{x}_{ij}\} = 0 \quad \text{for all} \;\; j = 1, \cdots, m_i; i = 1, \cdots, n.$$

Thus, we can define the marked EPEE $\text{GF}_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}}_{Aetel})$ for the k-th estimating equation as follows:

$$\text{GF}_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}}_{Aetel}) = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \partial_{\boldsymbol{\beta}_k} \mu_{ij}(\boldsymbol{\beta})[\mathbf{y}_{ij} - \mu_{ij}(\boldsymbol{\beta})]\mathbf{1}(\boldsymbol{\eta}'\mathbf{x}_{ij} \leq u). \tag{4.23}$$

Similar to (4.19), we can define $\text{PCvM}_{n,k}$.

For type II time-dependent covariates, the $k-$th component of $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$ takes the form

$$g_k(\mathbf{z}_i, \boldsymbol{\theta}) = \sum_{j=1}^{m_i}[\sum_{s=j}^{m_i} \partial_{\boldsymbol{\beta}_{\tilde{k}}} \mu_{is}(\boldsymbol{\beta})\ell_{i,sj}^{(b)}][\mathbf{y}_{ij} - \mu_{ij}(\boldsymbol{\beta})],$$

where $\ell_{i,sj}^{(b)}$ is the (s, j)th component of $L_i^{(b)}$ and $\tilde{k}$ can be determined according to equation (4.7). A sufficient condition of $E[g_k(\mathbf{z}_i, \boldsymbol{\theta})] = 0$ is (4.6). Following the argument for the type III time-dependent covariates, we can define the marked EPEE $\mathrm{GF}_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}}_{Aetel})$ for the k-th estimating equation as follows:

$$\mathrm{GF}_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}}_{Aetel}) = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m_i}[\sum_{s=j}^{m_i} \partial_{\boldsymbol{\beta}_{\tilde{k}}} \mu_{is}(\boldsymbol{\beta})\ell_{i,sj}^{(b)}][\mathbf{y}_{ij} - \mu_{ij}(\boldsymbol{\beta})]\mathbf{1}(\boldsymbol{\eta}'\mathbf{x}_{ij} \le u). \quad (4.24)$$

Similar to type II and III time-dependent covariates, we can define the marked EPEE $\mathrm{GF}_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}}_{Aetel})$ for type I time-dependent covariates accordingly.

4.3 Simulation studies

Simulation studies were conducted to examine the performance of our adjusted ETEL ratio statistic and the goodness of fit statistics. We considered both cross-sectional and longitudinal studies. In each case, we compared the rejection rates of LR_{Aetel} with those of the unadjusted ETEL ratio statistic under various situations (different sample sizes or different distributions) and evaluated the finite sample performance of the goodness of fit statistics proposed. Moreover, we followed a simulation study in Lai and Small (2007) to demonstrate the necessity of choosing the proper estimating equations.

4.3.1 Study I: cross-sectional study

We simulated data from:

$$\mathbf{y}_{i1} = \mathbf{x}_{i1}'\boldsymbol{\beta} + \epsilon_i \quad \text{for} \quad i = 1, \cdots, n,$$

where $\mathbf{x}_{i1} = (1, x_{i1,1}, x_{i1,2})'$, ϵ_i was a random error with zero mean and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ was a 3×1 unknown parameter vector. We set $n = 20, 40, 60$ and the true value of $\boldsymbol{\beta}$ was set to $(0, 0, 0)'$. We used the estimating equations $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{x}_{i1}(\mathbf{y}_{i1} - \mathbf{x}_{i1}'\boldsymbol{\beta})$.

Suppose the null hypothesis is $H_0 : \beta_1 = 0$. To compare the type I and type II errors of LR_{Aetel} with those of the unadjusted ETEL ratio statistic, we considered three different error distributions: $\epsilon_i \sim N(0, 1)$, $\epsilon_i \sim t(3)$, and $\epsilon_i \sim \chi^2(3) - 3$, where $\chi^2(3)$ represents a Chi-square distribution with 3 degrees of freedom. The $\chi^2(3) - 3$ is a skewed distribution, whereas $t(3)$ is a distribution with heavy tails. Furthermore, we also examined the effect of using different covariate distributions. We first generated $(x_{i1,1}, x_{i1,2})$ from a $N_2(0, I_2)$ distribution. Then, we generated $x_{i1,1}$ independently from a Bernoulli distribution with success probability of 0.5, and $x_{i1,2}$ independently from a uniform distribution on $[0, 1]$. For each case, we set the significance level to $\alpha = 5\%$, and used 5000 replications to estimate the rejection rates.

As seen from Table 4.1, for all sample sizes considered, all type I errors ($\beta_1 = 0$) for the unadjusted ETEL ratio statistic are larger than 0.05, confirming the under-coverage problem in constructing confidence regions. For the adjusted ETEL method, the type I errors are much closer to the nominal value of 0.05. Different error distributions and covariate distributions influence the finite sample performance of both the adjusted and unajusted ETEL methods. For instance, the under-coverage problem gets worse when $\epsilon_i \sim \chi^2(3) - 3$, and is the worst when $\epsilon_i \sim t(3)$. Moreover, there are about $0.5\% - 2\%$ of replications for which the unadjusted ETEL doesn't have a solution for the parameter estimates. Although the power of rejecting the null when the null is false ($\beta_1 \neq 0$) is larger for the unadjusted ETEL ratio statistic than the adjusted ETEL ratio statistic, it is mostly due to the fact that the type I error is much more inflated for the unadjusted ETEL ratio statistic.

We also simulated data from: $\mathbf{y}_{i1} = \beta_0 + \beta_1 x_{i1,1} + \beta_2 x_{i1,2} + c x_{i1,1}^2 + \epsilon_i$, for $i = 1, \cdots, n$, where the true value of $\boldsymbol{\beta}$ was set to $(0, 0, 0)'$, $(x_{i1,1}, x_{i1,2}) \sim N_2(0, I_2)$, and we considered

Table 4.1: Comparison of the rejection rates of unadjusted ETEL and adjusted ETEL ratio tests at the 5% significance level

$(x_{i1,1}, x_{i1,2}) \sim N(0, I_2)$										
		$\epsilon_i \sim N(0,1)$			$\epsilon_i \sim \chi^2(3) - 3$			$\epsilon_i \sim t(3)$		
β_1	Methods	n=20	n=40	n=60	n=20	n=40	n=60	n=20	n=40	n=60
0	unadj ETEL	0.133	0.092	0.070	0.140	0.098	0.088	0.156	0.115	0.100
	adj ETEL	0.094	0.073	0.060	0.101	0.077	0.074	0.115	0.096	0.086
0.5	unadj ETEL	0.662	0.876	0.964	0.298	0.339	0.439	0.464	0.603	0.697
	adj ETEL	0.588	0.854	0.958	0.241	0.306	0.413	0.393	0.575	0.675
1.0	unadj ETEL	0.976	0.999	1.000	0.579	0.761	0.860	0.806	0.927	0.963
	adj ETEL	0.963	0.999	1.000	0.513	0.732	0.846	0.767	0.919	0.960
1.5	unadj ETEL	0.999	1.000	1.000	0.805	0.938	0.983	0.928	0.982	0.990
	adj ETEL	0.995	0.999	1.000	0.760	0.926	0.980	0.913	0.978	0.989
2.0	unadj ETEL	1.000	1.000	1.000	0.909	0.985	0.997	0.967	0.991	0.996
	adj ETEL	0.996	0.999	1.000	0.881	0.983	0.996	0.961	0.990	0.995

$x_{i1,1} \sim$ Bernoulli (0.5), $x_{i1,2} \sim U(0,1)$										
		$\epsilon_i \sim N(0,1)$			$\epsilon_i \sim \chi^2(3) - 3$			$\epsilon_i \sim t(3)$		
β_1	Methods	n=20	n=40	n=60	n=20	n=40	n=60	n=20	n=40	n=60
0	unadj ETEL	0.097	0.069	0.064	0.115	0.078	0.067	0.125	0.099	0.081
	adj ETEL	0.075	0.058	0.054	0.095	0.065	0.059	0.096	0.079	0.069
0.5	unadj ETEL	0.267	0.393	0.527	0.139	0.146	0.163	0.213	0.241	0.279
	adj ETEL	0.227	0.357	0.497	0.117	0.126	0.143	0.183	0.218	0.257
1.0	unadj ETEL	0.605	0.874	0.975	0.245	0.312	0.387	0.408	0.559	0.682
	adj ETEL	0.594	0.858	0.970	0.211	0.285	0.365	0.394	0.533	0.661
1.5	unadj ETEL	0.874	0.995	1.000	0.387	0.537	0.660	0.598	0.795	0.883
	adj ETEL	0.883	0.995	1.000	0.346	0.492	0.636	0.632	0.787	0.877
2.0	unadj ETEL	0.979	1.000	1.000	0.536	0.737	0.854	0.710	0.912	0.967
	adj ETEL	0.974	1.000	1.000	0.501	0.708	0.842	0.779	0.919	0.967

Table 4.2: Rejection rates for the goodness of fit statistics for various values of c in a cross-sectional study at the 5% significance level

ϵ_i	c	$\text{PCvM}_{n,1}$	$\text{PCvM}_{n,2}$	$\text{PCvM}_{n,3}$	PCvM_n
$N(0,1)$	0	0.062	0.052	0.056	0.048
	0.1	0.220	0.422	0.056	0.336
	0.2	0.656	0.888	0.066	0.828
	0.3	0.932	0.992	0.102	0.986
	0.4	0.996	1.000	0.128	1.000
$t(3)$	0	0.032	0.030	0.036	0.034
	0.1	0.066	0.154	0.046	0.118
	0.2	0.198	0.546	0.048	0.436
	0.3	0.478	0.808	0.070	0.748
	0.4	0.756	0.958	0.096	0.932
$\chi^2(3)-3$	0	0.030	0.058	0.060	0.062
	0.1	0.046	0.098	0.060	0.068
	0.2	0.114	0.234	0.062	0.156
	0.3	0.250	0.504	0.086	0.404
	0.4	0.404	0.754	0.066	0.678

three different error distributions: $\epsilon_i \sim N(0,1)$, $\epsilon_i \sim t(3)$, and $\epsilon_i \sim \chi^2(3)-3$. We used $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}) = \mathbf{x}_{i1}(\mathbf{y}_{i1} - \mathbf{x}_{i1}^{'}\boldsymbol{\beta})$, where $\mathbf{x}_{i1} = (1, x_{i1,1}, x_{i1,2})^{'}$. The $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$ is correctly specified if $c = 0$, whereas $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta})$ is misspecified if $c \neq 0$. Table 4.2 shows the finite sample performance of $\{\text{PCvM}_{n,k}\}_{k=1}^{3}$ and PCvM_n at the 5% significance level. Consistent with our expectations, the power for detecting misspecified equations increases with the value of $|c|$. The performance is better when $\epsilon_i \sim N(0,1)$ compared to $\epsilon_i \sim t(3)$, and $\epsilon_i \sim \chi^2(3)-3$ is the worst. We note that the power of $\text{PCvM}_{n,3}$, which is to test the estimating equation $g_3(\mathbf{z}_i, \boldsymbol{\theta}) = x_{i1,2}(\mathbf{y}_{i1} - \mathbf{x}_{i1}^{'}\boldsymbol{\beta})$, does not increase much with the value of $|c|$ since $x_{i1,1}$ and $x_{i1,2}$ were independently generated.

4.3.2 Study II: longitudinal data

We considered the following model:

$$\mathbf{y}_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_i + \beta_3 t_{ij} x_i + b_i + \epsilon_{ij}, \tag{4.25}$$

Table 4.3: Rejection rates for the goodness of fit statistics for various values of c in a longitudinal study at the 5% significance level

c	PCvM$_{n,1}$	PCvM$_{n,2}$	PCvM$_{n,3}$	PCvM$_{n,4}$	PCvM$_n$
0	0.080	0.070	0.070	0.070	0.050
0.2	0.210	0.260	0.260	0.170	0.250
0.4	0.810	0.880	0.880	0.670	0.700
0.6	1.000	1.000	1.000	0.980	1.000

for $i = 1, \cdots, n$, where t_{ij} denotes time taking values in $(1, 2, 3, 4, 5)$, x_i was independently generated from a $N(0, 1)$ distribution, b_i was independently generated from a $N(0, 1)$ distribution, and ϵ_{ij} was independently generated from a $N(0, 1)$ distribution. The true value of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ was set at $(1, 1, 1, 1)'$. We used the GEE with an independent working correlation matrix.

We tested the null hypothesis $H_0 : \beta_3 = 1$ and used 5000 replications to estimate the type I errors. We considered $n = 40, 60$ and 80. At significance level $\alpha = 5\%$, the type I errors of LR_{Aetel} were $0.064, 0.060, 0.056$ respectively, whereas those of the unadjusted ETEL ratio statistic were $0.079, 0.070, 0.066$ respectively. Again our LR_{Aetel} was more accurate.

We added an extra term cx_i^2 in the model (4.25), but didn't include it in $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$. Table 4.3 shows the performance of $\{\text{PCvM}_{n,k}\}_{k=1}^4$ and PCvM$_n$ for various values of c. As the value of $|c|$ increases, the power for detecting misspecified equations goes up.

4.3.3 Study III: testing the type of time-dependent covariates

We used the simulation study for a type II time-dependent covariate in Section 4.1 of Lai and Small (2007) to show the performance of our adjusted ETEL method and goodness of fit statistics. The data were simulated under the mechanism

$$y_{it} = \gamma_0 + \gamma_1 x_{it} + \gamma_2 x_{i,t-1} + b_i + e_{it} \quad \text{and} \quad x_{i,t} = \rho x_{i,t-1} + \epsilon_{it}$$

Table 4.4: Results of adjusted ETEL with various estimating equations for a type II time-dependent covariate

Estimating equations	Bias	RMSE	Efficiency
Type II	0.00	0.040	1.82
Type III	0.00	0.053	1.04
GEE independence	0.00	0.054	1.00
GEE exchangeable	-0.12	0.090	$-$
GEE AR-1	-0.79	0.037	$-$

where b_i, e_{it} and ϵ_{it} are mutually independent and normally distributed with mean 0 and variances 4, 1 and 1 respectively; the x_{it}-process is stationary, i.e. $x_{i0} \sim N(0, \sigma_\epsilon^2/(1 - \rho^2))$. We refer the reader to Lai and Small (2007) for more details. Note that x_{it} is a type II covariate. We used our adjusted ETEL method with the following estimating equations: (a) the type II estimating equations according to (4.5), labelled type II; (b) the type III estimating equations according to (4.8), labelled type III; (c) GEE using the independent working correlation, labelled GEE independence; (d) GEE using the exchangeable working correlation, labelled GEE exchangeable; (e) GEE using the AR-1 working correlation, labelled GEE AR-1. We compared the bias, root-mean-square error and the efficiency of each case for the parameter β_1 with the GEE independence case (the efficiency is the ratio of the mean-square error of the GEE independence case to that of the competing case). As we can see from Table 4.4, the GEE independence and GEE AR-1 are biased, because they use some invalid estimating equations. The other three are all unbiased, with type II being more efficient than the other two. Combining all available valid estimating equations does improve efficiency.

With the same type II estimating equations, our method has slightly less RMSE (0.0345 vs 0.0407) than Lai and Small (2007)'s method. Furthermore, our goodness of fit test for the nominal 0.05-level test of the null hypothesis that x_{it} is a type II time-dependent covariate has more reliable type I error (0.055 vs 0.066) than Lai and Small (2007)'s method.

4.4 Longitudinal Schizophrenia study of hippocampus

We consider a neuroimaging dataset about the shape of the hippocampus structure in the left and right brain hemispheres in schizophrenia (SC) patients and healthy controls, collected at 14 academic medical centers in North America and western Europe, with partial funding from Lilly Research Laboratories (Lieberman et al., 2005; Styner et al., 2004). The hippocampus, a gray matter structure in the limbic system, is involved in processes of motivation and emotions and has a central role in the formation of memory.

This is a longitudinal, randomized, controlled, multisite, double-blind study. In this study, 238 first-episode schizophrenia patients were enrolled meeting the following criteria: age 16 to 40 years; onset of psychiatric symptoms before age 35; diagnosis of schizophrenia, schizophreniform, or schizoaffective disorder according to DSM-IV criteria; and various treatment and substance dependence conditions. After random allocation at baseline, 123 patients were selected to receive a conventional antipsychotic, haloperidol (2-20 mg/d), and 115 were selected to receive an atypical antipsychotic olanzapine (5-20 mg/d). Patients were treated and followed up to 47 months. 56 healthy control subjects matched to the patient's demographic characteristics were also enrolled. Neurocognitive and MRI assessments were performed at months 0 (baseline), 3, 6, 13, 24, 36 and 47 approximately, with different subjects having different visiting times, and some subjects dropped out during the course of the study. Covariates of interest were WBV (whole brain volume), race (Caucasian, African American and others), age (in years), gender, group (the two schizophrenia groups and the healthy control group) and time (visiting times in months). Among all the covariates, WBV is the only time-dependent covariate.

The aim of our study was to investigate the difference of the hippocampus medial representation (m-rep) thickness across groups (the two schizophrenia groups and the healthy control group) while controlling for the other covariates of interest. The response of interest was the hippocampus m-rep thickness at the 24 medial atoms of the left and the right brain. The m-rep is a linked set of medial primitives named medial atoms, which are formed from two equal length vectors and are composed of a position, a radius, a frame implying the tangent plane to the medial manifold, and an object angle (Styner et al., 2004). The medial atoms are grouped by intra-figural links into figures that are connected by inter-figural links. Via interpolation, a fully connected boundary is implied by the m-rep. The m-rep thickness is the radius of each medial atom. The procedure for generating a m-rep model was detailed in Styner et al. (2004)

First, we considered the baseline analysis. We used the moment model based on the estimating equations $\mathbf{x}_{i1}(\mathbf{y}_{i1} - \mathbf{x}_{i1}^{'}\boldsymbol{\beta})$, where $\mathbf{y}_{i1}$ is the m-rep thickness measured at baseline for the i-th subject at each medial atom of the left and right hippocampi; $\mathbf{x}_{i1}$ is an 8×1 vector given by $\mathbf{x}_{i1} = (1, \text{gender}_i, \text{age}_i, \text{SC1}_i, \text{SC2}_i, race1_i, race2_i, WBV_{i1})^{'}$, where SC1 and SC2 were, respectively, dummy variables for haloperidol-treated SC patients and olanzapine-treated SC patients versus healthy controls, and $race1$ and $race2$ were, respectively, dummy variables for Caucasian and African American versus other race; $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_7)^{'}$. Many existing statistical methods for image data require that the error distribution is Gaussian and the variance is constant. The Shapiro-Wilk normality test was applied to check this parametric assumption of the general linear model at each atom for the left hippocampus and right hippocampus using the residuals. Figure 4.1 shows that the Shapiro-Wilk test rejects the normality assumption at many atoms of the both left and right hippocampus structures, therefore our nonparametric adjusted ETEL method is prefered for the analysis of this dataset.

Since our goal is to detect the difference in the m-rep thickness across the three groups, we set up the null hypotheses $H_0 : \beta_4 = \beta_5 = 0$ at all 24 atoms for both the left

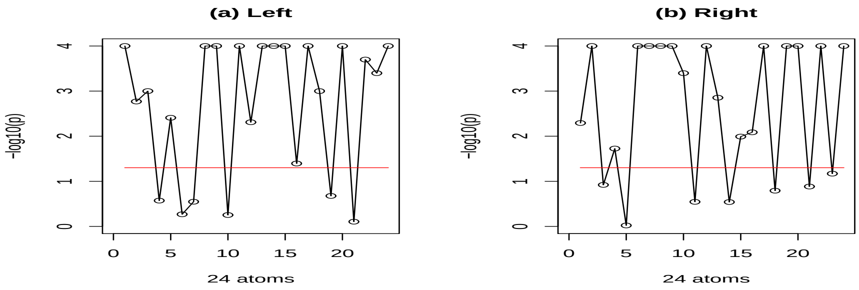

Figure 4.1: Results from the longitudinal schizophrenia study. (a) shows the $-\log_{10}(p)$-values for the Shapiro-Wilk test for the residuals at each atom on the left hippocampus; (b) shows the $-\log_{10}(p)$-values for the Shapiro-Wilk test for the residuals at each atom on the right hippocampus. The red horizontal line is the 0.05 cut-off line.

and right hippocampi. Accordingly, we have

$$R = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix},$$

and $\mathbf{b}_0 = (0,0)'$. We used LR_{Aetel} to carry out the test. The color-coded p-values of the LR_{Aetel} across the atoms of both the left and right reference hippocampi are shown in Figure 4.3 a and b. The false discovery rate approach was used to correct for multiple comparisons, and the resulting adjusted p-values were shown in Figure 4.3 c and d. Before correcting for multiple comparisons, there was a significant group difference in the m-rep thickness at the central atoms near the tail in the left hippocampus and some area in the right hippocampus. However, there is not a significant group effect at any atoms after correcting for multiple comparisons.

Secondly, we did a longitudinal data analysis. The advantage of a longitudinal study over a baseline study is that it allows us to determine (i) whether the change pattern of the response is similar across the three groups; (ii) whether there is difference in the response across the three groups on average over time,. We considered the moment model with $\mathbf{x}_{ij} = (1, \text{gender}_i, \text{age}_i, \text{SC1}_i, \text{SC2}_i, race1_i, race2_i, WBV_{ij}, time_{ij}, \text{SC1}_i{*}time_{ij}, \text{SC2}_i{*}$

$time_{ij})'$.

Since WBV is a time-dependent covariate, we needed to verify its appropriate type. Moreover, from a neuroscience point of view, the m-rep thickness at each atom serves as a local volumetric measure and covaries with WBV. We started with type III and used the GEE estimating equations in (4.2) with $V_i = I_i$. Then we used the type II equations specified in (4.5) and tested whether WBV is type II against type III. The LR_{GF} did not reject for almost all 24 atoms, suggesting WBV is a type II covariate for most atoms. Furthermore, we used the type I equations specified in (4.3) and tested whether WBV is type I against type II. The LR_{GF} rejected that WBV was of type I for most atoms (Figure 4.2). This indicates the invalidity of some type I equations. Our goodness of fit statistics revealed that some of the extra equations added for type I, such as

$$E\{\partial_{\beta_8}\mu_{is}(\boldsymbol{\beta})[\mathbf{y}_{ij} - \mu_{ij}(\boldsymbol{\beta})]\} = 0 \ \text{ for all } \ s < j, j = 1, \cdots, m_i.$$

were not valid. For instance, for the 3-nd atom on the left hippocampus, the $p-$value of the goodness of fit test for the newly added equation $E\{\partial_{\beta_8}\mu_{i2}(\boldsymbol{\beta})[\mathbf{y}_{i3} - \mu_{i3}(\boldsymbol{\beta})]\} = 0$ was smaller than 0.001 (Figure 4.2 e); for the 14-th atom on the right hippocampus, the $p-$value of the goodness of fit test for the newly added equation $E\{\partial_{\beta_8}\mu_{i2}(\boldsymbol{\beta})[\mathbf{y}_{i3} - \mu_{i3}(\boldsymbol{\beta})]\} = 0$ was smaller than 0.001 (Figure 4.2 f). Therefore, we treated WBV as a type II time-dependent covariate and used the corresponding estimating equations for the longitudinal analysis.

To determine whether the change pattern of the thickness of the hippocampus is similar or not across the three groups over time, we tested the null hypotheses $H_0 : \beta_9 = \beta_{10} = 0$ (β_9 and β_{10} are the coefficients of the interaction terms of group and time) at all 24 atoms for each of the left hippocampus and the right hippocampus, and it turned out that the interaction terms were not significant for most atoms. Next we deleted the interaction terms and tried to look at whether there is a difference in the response across the three groups on average over time with respect to the null hypotheses $H_0 : \beta_3 = \beta_4 = 0$ at all

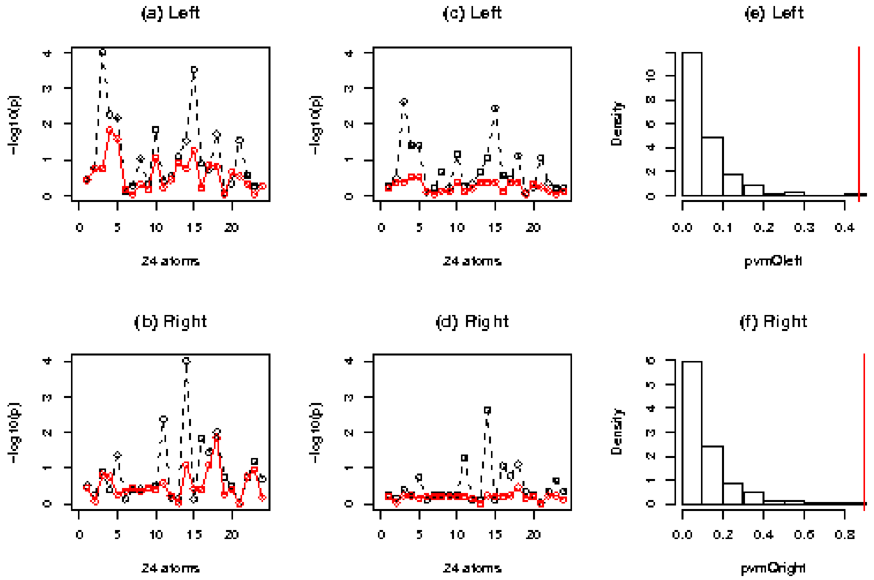

Figure 4.2: Results from the longitudinal schizophrenia study. Maps of $-\log_{10}(p)$-values for testing WBV as a type I time-dependent covariate (black) and a type II time-dependent covariate (red): uncorrected $-\log_{10}(p)$-values for the (a) left hippocampus and (b) right hippocampus; corrected $-\log_{10}(p)$-values for the (c) left hippocampus and (d) right hippocampus; (e) the goodness of fit test for $E\{\partial_{\beta_8}\mu_{i2}(\boldsymbol{\beta})[\mathbf{y}_{i3}-\mu_{i3}(\boldsymbol{\beta})]\} = 0$ for the 3-rd atom on the left hippocampus; (f) the goodness of fit test for $E\{\partial_{\beta_8}\mu_{i2}(\boldsymbol{\beta})[\mathbf{y}_{i3}-\mu_{i3}(\boldsymbol{\beta})]\} = 0$ for the 14-th atom on the right hippocampus.

24 atoms for each of the left hippocampus and the right hippocampus. Again we only found that there was a significant difference through time in the m-rep thickness at the lower central atoms in the left hippocampus across schizophrenia patients and healthy controls groups after correcting for multiple comparisons, but the differences were not significant at other atoms, nor at any atoms in the right hippocampus. The color-coded p-values of the LR_{Aetel} across the atoms of both left and right reference hippocampi are shown in Figure 4.3 e and f, and the corrected p-values are shown in Figure 4.3 g and h. Before correcting for multiple comparisons, there was a significant group difference in the m-rep thickness at the central atoms near the tail in the left hippocampus, and the significance level is larger than that of the baseline analysis. After correcting for multiple comparisons, there is still a significant group effect at the central atoms near the tail in the left hippocampus.

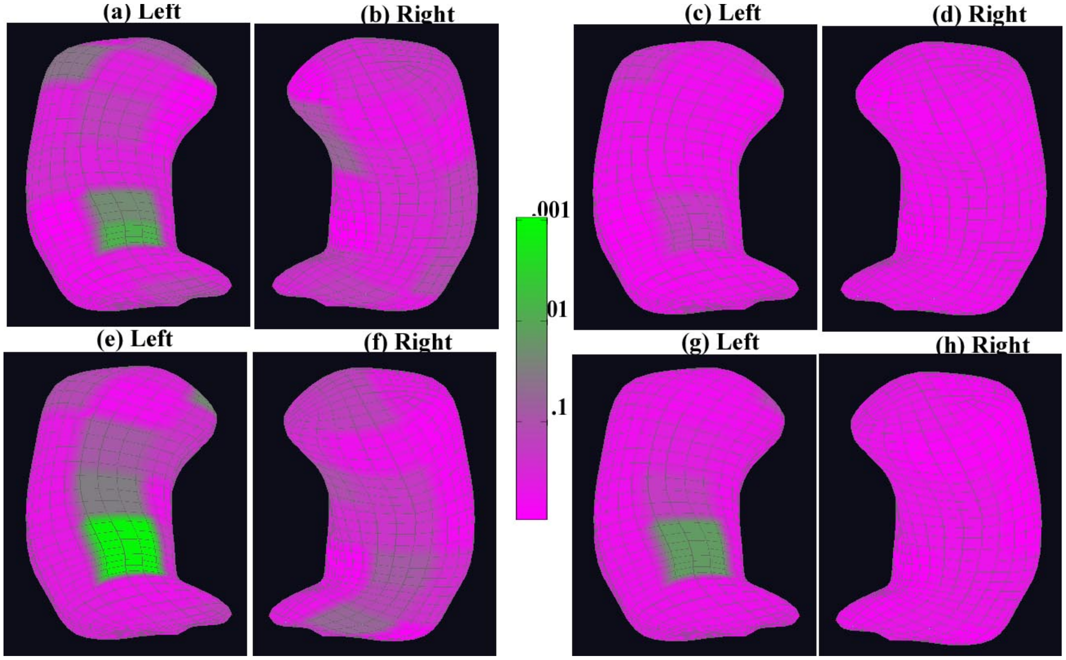

Figure 4.3: Results for the group effect from the longitudinal schizophrenia study. The top row is for the baseline analysis: the uncorrected $p-$value maps for the left hippocampus (a), and the right hippocampus (b); the corrected $p-$value maps for the left hippocampus (c), and the right hippocampus (d). The bottom row is for the longitudinal analysis: the uncorrected $p-$value maps for the left hippocampus (e), and the right hippocampus (f); the corrected $p-$value maps for the left hippocampus (g), and the right hippocampus (h).

We compared the results by making the assumption that WBV was a type II time-dependent and also a type III time-dependent covariate. Treating WBV as type II time-dependent lowered the $p-$values, making some non-significant $p-$values for the group effect significant. On the other hand, we found that all the standard errors associated with the parameter estimates treating WBV as a type II time-dependent covariate were uniformly less than those treating WBV as a type III, which confirms that treating WBV as type II gains efficiency by making use of more correct estimating equations. Table 4.5 compares the standard deviations of the parameter estimates between treating WBV as a type II time-dependent covariate and a type III time-dependent covariate at atom 11 of the left hippocampus.

The longitudinal analysis increased the significance level at those significant atoms for the group effect, compared to the baseline analysis. We were also able to observe the

Table 4.5: Estimate and standard error comparisons of the parameter estimates between treating WBV as a type II time-dependent covariate and a type III time-dependent covariate at atom 11 of the left hippocampus.

		Gender	Age	SC1	SC2	Race1	Race2	WBV	Time
Estimate	Type III	0.028	−.001	0.209	0.128	−.218	−.358	−.282	0.001
	Type II	0.022	−.001	0.212	0.131	−.216	−.356	−.279	0.001
Std Error	Type III	0.078	0.007	0.062	0.058	0.097	0.102	0.237	0.022
	Type II	0.075	0.005	0.058	0.054	0.094	0.100	0.221	0.018

change differences across groups through time, although it is not substantial. Both the baseline analysis and longitudinal analysis suggest that there is an asymmetric aspect in that the left hippocampus shows larger regions of significance than the right one, and the significant positions of group differences are around the middle atoms near the tail for the left hippocampus.

4.5 Discussion

We have developed an adjusted ETEL and associated test statistics for the analysis of cross-sectional and longitudinal neuroimaging data. Our adjusted ETEL not only avoids standard parametric assumptions for imaging data, but also dramatically improves its finite sample performance of the original ETEL. Our test statistics are very useful for determining the types of time-dependent covariates in longitudinal studies. Our simulation studies have shown good finite sample performance of the adjusted ETEL ratio statistic and goodness of fit statistics.

Our adjusted ETEL and associated test statistics have several distinctive features compared with linear mixed models used in the neuroimaging literature. Our adjusted ETEL is a powerful tool to carry out nonparametric statistical inference, whereas linear mixed models usually assume a Gaussian distribution of imaging measures with a specific covariance structure. Incorrectly specifying the distribution of the multivariate/univariate imaging measures can lead to seriously biased parameter estimates and incorrect standard errors of parameter estimates. Our adjusted ETEL can produce more efficient (or accurate) estimates by incorporating more estimating equations than the number of parameters. This advantage of having more estimating equations in our adjusted ETEL becomes obvious when handling time-dependent covariates, such as brain structure, function, cognition, and disease status in longitudinal studies. Moreover, the diagnostic methods associated with our adjusted ETEL provides useful tools for testing the specific type of time-dependent covariates, whereas it is unclear how to test the type of time-dependent covariates under linear mixed model. Thus, our adjusted ETEL can greatly maximize our ability and/or power of utilizing all information about longitudinal imaging data to understand normal brain development and aging, as well as evolution of pathology.

Many issues still merit further research. One major issue is to develop a test procedure, such as random field theory and resampling methods, to correct for multiple comparisons in order to control for family-wise error rates under the moment model (4.1). Another major issue is to extend the test procedure to conduct cluster size inference and examine its performance in controlling the Type I error rate. The test procedure may lead to a simple cluster size test (cluster size test assesses significance for all sizes of the connected regions greater than a given primary threshold).

4.6 Appendix

Assumptions and proofs:

ASSUMPTION A: $\mathbf{z}_i$ *forms an independent and identical sequence.*

ASSUMPTION B: *The true value* $\boldsymbol{\theta}_0$ *of* $\boldsymbol{\theta}$ *is the unique solution to* $E\{\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})\} = 0$ *and* $\boldsymbol{\theta}_0$ *is an interior point of the compact set* $\Theta \subset R^p$.

ASSUMPTION C: *In a neighborhood of the true value* $\boldsymbol{\theta}_0$, $\mathbf{g}(\mathbf{z}, \boldsymbol{\theta})$ *has a second-order* *continuous derivative with respect to* $\boldsymbol{\theta}$ *and* $||\partial_{\boldsymbol{\theta}}\mathbf{g}(\mathbf{z},\boldsymbol{\theta})||$, $||\partial^2_{\boldsymbol{\theta}}\mathbf{g}(\mathbf{z},\boldsymbol{\theta})||$, *and* $||\mathbf{g}(\mathbf{z},\boldsymbol{\theta})||^3$ *are bounded by some integrable function* $G(x)$ *with* $E_F\{G(x)\} < \infty$.

Assumption D: *The rank of* $E\{\partial_{\boldsymbol{\theta}}\mathbf{g}(\mathbf{z},\boldsymbol{\theta}_0)\}$ *is* p *and* $E\{\mathbf{g}(\mathbf{z},\boldsymbol{\theta}_0)\mathbf{g}'(\mathbf{z},\boldsymbol{\theta}_0)\}$ *is positive definite.*

Assumption E: $F_{\boldsymbol{\eta}}(u)d\boldsymbol{\eta}$ *is absolute continuous with respect to Lebesgue measure on* Π*, where* $F_{\boldsymbol{\eta}}(u)$ *is the true cumulative distribution function of* $\boldsymbol{\eta}'\mathbf{x}$.

Assumption F: $||a(\mathbf{x})||^3$ *is bounded by some integrable function* $G(x)$ *in assumption C.*

Notations. For notational simplicity, we define $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{Aetel}$, $\hat{\mathbf{t}} = \hat{\mathbf{t}}_{Aetel}$, and $\mathbf{h}_i(\boldsymbol{\theta}) = \mathbf{g}(\mathbf{z}_i,\boldsymbol{\theta})$ for $i = 1,\cdots,n$, and $\mathbf{h}_{n+1}(\boldsymbol{\theta}) = \mathbf{g}_{n+1}(\boldsymbol{\theta})$. For the adjusted ETEL, we define

$$G_n(\mathbf{t},\boldsymbol{\theta}) = -\ell_{Aetel}(\mathbf{t},\boldsymbol{\theta}) = -(n+1)^{-1}\sum_{i=1}^{n+1}\log[\frac{(n+1)\exp(\mathbf{t}'\mathbf{h}_i(\boldsymbol{\theta}))}{\sum\limits_{j=1}^{n+1}\exp(\mathbf{t}'\mathbf{h}_j(\boldsymbol{\theta}))}].$$

$$S_n(\boldsymbol{\theta}) = \begin{pmatrix} S_{n,11} & S_{n,12} \\ S_{n,21} & S_{n,22} \end{pmatrix} = (n+1)^{-1}\begin{pmatrix} \sum_{i=1}^{n+1}\mathbf{h}_i(\boldsymbol{\theta})\mathbf{h}_i'(\boldsymbol{\theta}) & -\sum_{i=1}^{n+1}(\partial_{\boldsymbol{\theta}}\mathbf{h}_i(\boldsymbol{\theta}))' \\ -\sum_{i=1}^{n+1}\partial_{\boldsymbol{\theta}}\mathbf{h}_i(\boldsymbol{\theta}) & 0 \end{pmatrix},$$

$$S = S(\mathbf{t}_0,\boldsymbol{\theta}_0) = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} E(\mathbf{g}^{\otimes 2}(\mathbf{z}_i,\boldsymbol{\theta}_0))) & -E(\partial_{\boldsymbol{\theta}}\mathbf{g}(\mathbf{z}_i,\boldsymbol{\theta}_0))' \\ -E(\partial_{\boldsymbol{\theta}}\mathbf{g}(\mathbf{z}_i,\boldsymbol{\theta}_0)) & 0 \end{pmatrix}.$$

Lemma A1. If assumptions A, C, and D are satisfied, then for any $1/3 < \eta < 1/2$ *and* $\mathcal{T}_n(\eta) = \{\mathbf{t} : ||\mathbf{t}|| \leq n^{-\eta}\}$, $\sup_{\boldsymbol{\theta}\in\Theta,\mathbf{t}\in\mathcal{T}_n,1\leq i\leq(n+1)}|\mathbf{t}'\mathbf{h}_i(\boldsymbol{\theta})| \to 0$ *and* $\mathcal{T}_n(\eta) \subset \mathcal{T}_n(a_1;\boldsymbol{\theta}) = \{\mathbf{t} : \mathbf{t}'\mathbf{h}_i(\boldsymbol{\theta}) \in [-a_1,a_1]\}$ *for all* $\boldsymbol{\theta} \in \Theta$*, where* $a_1 > 0$.

Proof of Lemma A1. From assumptions A and C, $\max_{1\leq i\leq n+1}\sup_{\boldsymbol{\theta}\in\Theta}|\mathbf{h}_i(\boldsymbol{\theta})| = O(n^{1/3})$. Then, we have

$$\max_{1\leq i\leq n+1}\sup_{\mathbf{t}\in\mathcal{T}_n;\boldsymbol{\theta}\in\Theta}|\mathbf{t}'\mathbf{h}_i(\boldsymbol{\theta})| \leq O(n^{1/3})n^{-\eta} = O(n^{1/3-\eta}) \to 0$$

almost surely. Thus, Lemma A1 follows.

Lemma A2. If assumptions A and E are satisfied and $\overline{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + n^{-\eta_0}\mathbf{u}$*, then* $\overline{\mathbf{t}}(\overline{\boldsymbol{\theta}}) = \text{argmax}_{\mathbf{t}\in\mathcal{T}_n(a_1;\overline{\boldsymbol{\theta}})} F_n(\mathbf{t},\overline{\boldsymbol{\theta}})$ *exists and* $\overline{\mathbf{t}}(\overline{\boldsymbol{\theta}}) = O(n^{-\eta_0})$*, where* $||\mathbf{u}|| = 1$ *and* $F_n(\mathbf{t},\boldsymbol{\theta}) = -(n+1)^{-1}\sum_{i=1}^{n+1}\exp(\mathbf{t}^{'}\mathbf{h}_i(\boldsymbol{\theta}))$.

Proof of Lemma A2. It can be seen that $F_n(\mathbf{t},\overline{\boldsymbol{\theta}})$ is an analytical function of $\mathbf{t}$. Thus, $\tilde{\mathbf{t}} = \text{argmax}_{\mathbf{t}\in\mathcal{T}_n(\eta)} F_n(\mathbf{t},\overline{\boldsymbol{\theta}})$ exists. Using a Taylor's series expansion, we can show that

$$\begin{aligned} -1 &= F_n(\mathbf{0},\overline{\boldsymbol{\theta}}) \le F_n(\tilde{\mathbf{t}},\overline{\boldsymbol{\theta}}) \\ &= -1 - \tilde{\mathbf{t}}^{'}\sum_{i=1}^{n+1}\mathbf{h}_i(\overline{\boldsymbol{\theta}})/(n+1) - 0.5\tilde{\mathbf{t}}^{'}\sum_{i=1}^{n+1}\exp(\dot{\mathbf{t}}^{'}\mathbf{h}_i(\overline{\boldsymbol{\theta}}))\mathbf{h}_i(\overline{\boldsymbol{\theta}})^{\otimes 2}\tilde{\mathbf{t}}/(n+1), \quad (4.26) \end{aligned}$$

where $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{'}$, and $\dot{\mathbf{t}}$ is on the line joining $\tilde{\mathbf{t}}$ and 0. Moreover, because

$$(n+1)^{-1}\sum_{i=1}^{n+1}\exp(\dot{\mathbf{t}}^{'}\mathbf{h}_i(\overline{\boldsymbol{\theta}}))\mathbf{h}_i(\overline{\boldsymbol{\theta}})^{\otimes 2} \to E\{\mathbf{h}(\mathbf{z},\boldsymbol{\theta}_0)^{\otimes 2}\},$$

it follows from (4.26) that $||\tilde{\mathbf{t}}|| \le ||(n+1)^{-1}\sum_{i=1}^{n+1}\mathbf{h}_i(\overline{\boldsymbol{\theta}})|| = O(n^{-\eta_0}) = o(n^{-\eta})$ for all $\eta_0 > \eta$. Therefore, for large n, $\tilde{\mathbf{t}} \in \text{int}(\mathcal{T}_n(\eta)) \subset \mathcal{T}_n(a_1;\overline{\boldsymbol{\theta}})$ and $\partial_t F_n(\tilde{\mathbf{t}},\overline{\boldsymbol{\theta}}) = 0$. Because of the concavity of $F_n(\mathbf{t},\overline{\boldsymbol{\theta}})$ in $\mathbf{t}$, we have $\overline{\mathbf{t}}(\overline{\boldsymbol{\theta}}) = \tilde{\mathbf{t}}$ and $F_n(\tilde{\mathbf{t}},\overline{\boldsymbol{\theta}}) = \max_{\mathbf{t}\in\mathcal{T}_n(a_1;\overline{\boldsymbol{\theta}})} F_n(\mathbf{t},\overline{\boldsymbol{\theta}})$. Moreover, we have

$$\partial_{\mathbf{t}} F_n(\overline{\mathbf{t}}(\overline{\boldsymbol{\theta}}),\overline{\boldsymbol{\theta}}) = \partial_{\mathbf{t}} F_n(0,\overline{\boldsymbol{\theta}}) + \partial_{\mathbf{t}}^2 F_n(\dot{\mathbf{t}},\overline{\boldsymbol{\theta}})\overline{\mathbf{t}}(\overline{\boldsymbol{\theta}}).$$

Because $\max_{1\le i\le n+1}|\dot{\mathbf{t}}^{'}\mathbf{h}_i(\overline{\boldsymbol{\theta}})| = o(1)$, we have

$$\overline{\mathbf{t}}(\overline{\boldsymbol{\theta}}) = -[\sum_{i=1}^{n+1}\mathbf{h}_i(\overline{\boldsymbol{\theta}})^{\otimes 2}]^{-1}\sum_{i=1}^{n+1}\mathbf{h}_i(\overline{\boldsymbol{\theta}}) + o(n^{-\eta_0}).$$

Proof of Theorem 1.

The proof of Theorem 1 (i) consists of two steps as follows.

Step 1. $G_n(\overline{\mathbf{t}}(\boldsymbol{\theta}),\boldsymbol{\theta})$ attains its minimum value at some point $\tilde{\boldsymbol{\theta}}$ in the interior of the ball $||\boldsymbol{\theta} - \boldsymbol{\theta}_0|| \le n^{-\eta_0}$.

Step 2. $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges to ν_0 as described in Theorem 1 (i).

In Step 1, we can use assumptions (A)-(D) to show that $\sup_{\boldsymbol{\theta}\in\Theta} |h_{n+1}(\boldsymbol{\theta})+a_n E[\mathbf{g}(\mathbf{z},\boldsymbol{\theta})]| = O_p(a_n n^{-1/2})$ (van der Vaart and Wellner, 1996). Thus, the contribution from $h_{n+1}(\boldsymbol{\theta})$ is negligible. Then, we can follow the proof of Lemma 1 in Qin and Lawless (1996) to prove that $G_n(\overline{\mathbf{t}}(\overline{\boldsymbol{\theta}}), \overline{\boldsymbol{\theta}}) = O(n^{-2\eta_0})$ and $G_n(\overline{\mathbf{t}}(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0) = O(n^{-1}\log\log n) = o(n^{-2\eta_0})$. Since $G_n(\overline{\mathbf{t}}(\boldsymbol{\theta}), \boldsymbol{\theta})$ is a continuous function about $\boldsymbol{\theta}$ as $||\boldsymbol{\theta} - \boldsymbol{\theta}_0|| \leq n^{-2\eta_0}$, $G_n(\bar{t}(\boldsymbol{\theta}), \boldsymbol{\theta})$ has minimum value in the interior of this ball.

In Step 2, similar to Theorem 2 of Schennach (2007), we can obtain the first order conditions for $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{t}}$ as follows:

$$\begin{aligned}
Q_{1,n}(\mathbf{t}, \boldsymbol{\theta}) &= (n+1)^{-1} \sum_{i=1}^{n+1} \mathbf{h}_i(\boldsymbol{\theta}) \exp(\mathbf{t}' \mathbf{h}_i(\boldsymbol{\theta})) = 0, \\
Q_{2,n}(\mathbf{t}, \boldsymbol{\theta}) &= (n+1)^{-1} \sum_{i=1}^{n+1} \{1 - \frac{(n+1)\exp(\mathbf{t}'\mathbf{h}_i(\boldsymbol{\theta}))}{\sum_{j=1}^{n+1} \exp(\mathbf{t}'\mathbf{h}_j(\boldsymbol{\theta}))}\} \partial_{\boldsymbol{\theta}} \{\mathbf{h}_i(\boldsymbol{\theta})\mathbf{t}\} = 0.
\end{aligned}$$

Expanding the above first order conditions for $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{t}}$ around $\boldsymbol{\theta}_0$ and $\mathbf{t}_0 = 0$ leads to

$$\begin{aligned}
0 &= Q_{1,n}(0, \boldsymbol{\theta}_0) + [\{\partial_{\boldsymbol{\theta}} Q_{1,n}(0, \boldsymbol{\theta}_0)\}'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \partial_{\mathbf{t}} Q_{1,n}(0, \boldsymbol{\theta}_0)(\hat{\mathbf{t}} - 0)]\{1 + o_p(1)\}, \\
0 &= Q_{2,n}(0, \boldsymbol{\theta}_0) + [\partial_{\boldsymbol{\theta}} Q_{2,n}(0, \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \{\partial_{\mathbf{t}} Q_{2,n}(0, \boldsymbol{\theta}_0)\}'(\hat{\mathbf{t}} - 0)]\{1 + o_p(1)\}.
\end{aligned}$$

Therefore, it can be shown that

$$\begin{pmatrix} \hat{\mathbf{t}} \\ \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \end{pmatrix} = S_n^{-1}(\boldsymbol{\theta}_0) \begin{pmatrix} Q_{1,n}(0, \boldsymbol{\theta}_0) \\ 0 \end{pmatrix} \{1 + o_p(1)\}.$$

where

$$S_n(\boldsymbol{\theta}_0) = \begin{pmatrix} \partial_{\mathbf{t}} Q_{1,n} & (\partial_{\boldsymbol{\theta}} Q_{1,n})' \\ (\partial_{\mathbf{t}} Q_{2,n})' & \partial_{\boldsymbol{\theta}} Q_{2,n} \end{pmatrix}_{(0,\boldsymbol{\theta}_0)} = \frac{1}{n+1} \begin{pmatrix} \sum \mathbf{h}_i(\boldsymbol{\theta}_0)\mathbf{h}_i'(\boldsymbol{\theta}_0) & -\sum(\partial_{\boldsymbol{\theta}} \mathbf{h}_i(\boldsymbol{\theta}_0))' \\ -\sum \partial_{\boldsymbol{\theta}} \mathbf{h}_i(\boldsymbol{\theta}_0) & 0 \end{pmatrix}.$$

The law of large number ensures that $S_n(\boldsymbol{\theta}_0) \to S$. Thus, with some simple calculations, we can prove that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \sqrt{n} S_{22.1}^{-1} S_{21} S_{11}^{-1} \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}_0) + o_p(1), \tag{4.27}$$

where $S_{22.1} = -S_{21} S_{11}^{-1} S_{12}$. Applying the central limit theorem completes the proof of Theorem 1 (i).

Following the proof of Theorem 2 in Qin and Lawless (1994), we can finish the proof of Theorem 1 (ii) and (iii).

Proof of Lemma 1.

Let $\mathbf{g} = \mathbf{g}(\mathbf{z}, \boldsymbol{\theta}_0)$. If $E[\mathbf{g}|\mathbf{x}] = 0$, then $E[\mathbf{g}|\boldsymbol{\eta}^{'}\mathbf{x}] = E[E(\mathbf{g}|\mathbf{x})|\boldsymbol{\eta}^{'}\mathbf{x}] = 0$. Furthermore, if $E[\mathbf{g}|\boldsymbol{\eta}^{'}\mathbf{x}] = 0$, then it can be shown below that $E\{\mathbf{g} \times h(\mathbf{x})\} = 0$ holds for any continuous function $h(\mathbf{x})$. By a Fourier transformation, any continuous function $h(\mathbf{x})$ can be expressed as $\sum C_k \exp(iW_k\mathbf{x})$ for some constant C_k and W_k. Then, Part I of Theorem 1 in Bierens (1982) yields that $E[\mathbf{g} \times h(\mathbf{x})] = \sum C_k E[\mathbf{g} \times \exp(iW_k\mathbf{x})] = 0$. Furthermore, this is can be extended to any integrable function $h(\mathbf{x})$. Finally, letting $h(\mathbf{x}) = E[\mathbf{g}|\mathbf{x}]$, we have

$$E[\mathbf{g} \times h(\mathbf{x})] = E\{\mathbf{g} \times E[\mathbf{g}|\mathbf{x}]\} = E\{E[\mathbf{g} \times E(\mathbf{g}|\mathbf{x})]|\mathbf{x}\} = E\{[E(\mathbf{g}|\mathbf{x})]^2\} = 0,$$

which yields $E(\mathbf{g}|\mathbf{x}) = 0$.

Proof of Theorem 2.

(i) Using a Taylor's series expansion, we can show that $\mathrm{GF}_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}})$ can be approximated by

$$\begin{aligned}
&\mathrm{GF}_k(\boldsymbol{\eta}, u; \boldsymbol{\theta}_0) + n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) n^{-1} \textstyle\sum_{i=1}^{n} [\partial_{\boldsymbol{\theta}}^{'} g_k(\mathbf{z}_i, \tilde{\boldsymbol{\theta}}) - \partial_{\boldsymbol{\theta}}^{'} g_k(\mathbf{z}_i, \boldsymbol{\theta}_0)] \mathbf{1}(\boldsymbol{\eta}^{'}\mathbf{x}_i \le u) \\
&+ n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)[n^{-1} \textstyle\sum_{i=1}^{n} \partial_{\boldsymbol{\theta}}^{'} g_k(\mathbf{z}_i, \boldsymbol{\theta}_0) \mathbf{1}(\boldsymbol{\eta}^{'}\mathbf{x}_i \le u) - E\partial_{\boldsymbol{\theta}}^{'} g_k(\mathbf{z}_i, \boldsymbol{\theta}_0) \mathbf{1}(\boldsymbol{\eta}^{'}\mathbf{x}_i \le u)] \\
&+ n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) E[\partial_{\boldsymbol{\theta}}^{'} g_k(\mathbf{z}_i, \boldsymbol{\theta}_0) \mathbf{1}(\boldsymbol{\eta}^{'}\mathbf{x}_i \le u)],
\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}$ satisfies $||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2 \le ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||_2$. It follows from the law of large numbers that the second term and the third term in the last display are $o_p(1)$ (van der Vaart and Wellner 1996). We can prove that the marginals of $\mathrm{GF}_k(\boldsymbol{\eta}, u; \boldsymbol{\theta}_0)$ converge weakly to the corresponding marginals of the zero-mean Gaussian process $\mathrm{GF}_k(\boldsymbol{\eta}, u)$. This can be proved by using assumptions (C) and (D). Because $\mathscr{F} = \{f(\boldsymbol{\eta}, u) = g_k(\mathbf{z}, \boldsymbol{\theta}_0)\mathbf{1}(\boldsymbol{\eta}'\mathbf{x} \leq u) : (\boldsymbol{\eta}, u) \in \Pi\}$ is a VC class, which satisfies the universal entropy condition (van der Vaart and Wellner, 1996; Sections 2.5 and 2.6), the tightness of $\mathrm{GF}_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}})$ follows from the Donsker Theorem under the uniform entropy condition (Theorem 2.5.2 in van der Vaart and Wellner (1996); p.127). Because $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normal, it follows from the assumption that $E[\partial'_{\boldsymbol{\theta}} g_k(\mathbf{z}_i, \boldsymbol{\theta}_0)\mathbf{1}(\boldsymbol{\eta}'\mathbf{x}_i \leq u)]$ is asymptotically tight. This establishes that $\mathrm{GF}_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}})$ converges weakly to $\mathrm{G}_k(\boldsymbol{\eta}, u)$ as $n \to \infty$.

(ii) Applying the continuous mapping theorem yields Theorem 2 (ii).

Proof of Theorem 3.

Similar to the proof of Theorem 1, the proof of Theorem 3 (i) consists of two steps.

Step 1. $G_n(\overline{\mathbf{t}}(\boldsymbol{\theta}), \boldsymbol{\theta})$ attains its minimum value at some point in the interior of $||\boldsymbol{\theta} - \boldsymbol{\theta}_0|| \leq n^{-\eta_0}$.

Step 2. $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges to $\nu_0 + A_{1k}$ in distribution.

To prove Step 1, we can show that Lemmas A1 and A2 hold. Furthermore, we can show that $G_n(\overline{\mathbf{t}}(\overline{\boldsymbol{\theta}}), \overline{\boldsymbol{\theta}}) = O(n^{-2\eta_0})$ and $G_n(\overline{\mathbf{t}}(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0) = O(n^{-1}\log\log n) = o(n^{-2\eta_0})$, which yields Step 1.

To prove Step 2, we follow the exact same arguments in Theorem 1 and show that

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= \sqrt{n} S_{22.1}^{-1} S_{21} S_{11}^{-1} \frac{1}{n+1} \sum_{i=1}^{n} \{\mathbf{g}(\mathbf{z}_i, \boldsymbol{\theta}_0) - n^{-1/2} E[a(\mathbf{x}_{i1})]\} \\ &+ S_{22.1}^{-1} S_{21} S_{11}^{-1} E[a(\mathbf{x}_{i1})] + o_p(1),\end{aligned}$$

which leads to Theorem 3 (i). Particularly, $A_{1k} = S_{22.1}^{-1} S_{21} S_{11}^{-1} E[a(\mathbf{x}_{i1})]$.

The key step to Theorem 3 (ii) and (iii) is to show that $\mathrm{GF}_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}})$ can be approximated by

$$\mathrm{GF}_k(\boldsymbol{\eta}, u; \boldsymbol{\theta}_0) + n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)n^{-1}\sum_{i=1}^{n}[\partial'_{\boldsymbol{\theta}} g_k(\mathbf{z}_i, \tilde{\boldsymbol{\theta}}) - \partial'_{\boldsymbol{\theta}} g_k(\mathbf{z}_i, \boldsymbol{\theta}_0)]\mathbf{1}(\boldsymbol{\eta}'\mathbf{x}_i \leq u) + o_p(1).$$

Moreover, we note that $\sqrt{n}(\hat{\boldsymbol{\theta}}_{Aetel} - \boldsymbol{\theta}_0)$ converges to $\nu_0 + A_{1k}$ in distribution and

$$\mathrm{GF}_k(\boldsymbol{\eta}, u; \boldsymbol{\theta}_0) = n^{-1/2}\sum_{i=1}^{n}\{g_k(\mathbf{z}_i; \boldsymbol{\theta}_0) - n^{-1/2}a(\mathbf{x}_{i1})]\}\mathbf{1}(\boldsymbol{\eta}'\mathbf{x}_{i1} < u) + n^{-1}\sum_{i=1}^{n}[a(\mathbf{x}_{i1})\mathbf{1}(\boldsymbol{\eta}'\mathbf{x}_{i1} < u)].$$

Then, we can follow the same arguments in Theorem 2 to prove Theorem 3 (ii) and (iii) with $A_{2k}(\boldsymbol{\eta}, u) = E[a(\mathbf{x}_1)\mathbf{1}(\boldsymbol{\eta}'\mathbf{x}_1 < u)] - A_{1k}E[\partial'_{\boldsymbol{\theta}} g_k(\mathbf{z}, \boldsymbol{\theta}_0)\mathbf{1}(\boldsymbol{\eta}'\mathbf{x}_1 < u)]$.

Formula for computing the goodness of fit statistics.

With some algebraic calculations, we have

$$\begin{aligned}
\mathrm{PCvM}_{n,k} &= \int_{\Pi}[\mathrm{GF}_k(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}})]^2 F_{n,\boldsymbol{\eta}}(du)d\boldsymbol{\eta} \\
&= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} g_k(\mathbf{z}_i, \hat{\boldsymbol{\theta}})g_k(\mathbf{z}_j, \hat{\boldsymbol{\theta}})\int_{\Pi}\mathbf{1}(\boldsymbol{\eta}'\mathbf{x}_i \leq u)\mathbf{1}(\boldsymbol{\eta}'\mathbf{x}_j \leq u)F_{n,\boldsymbol{\eta}}(du)d\boldsymbol{\eta} \\
&= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{n} g_k(\mathbf{z}_i, \hat{\boldsymbol{\theta}})g_k(\mathbf{z}_j, \hat{\boldsymbol{\theta}})\int_{B^q}\mathbf{1}(\boldsymbol{\eta}'\mathbf{x}_i \leq \boldsymbol{\eta}'\mathbf{x}_r)\mathbf{1}(\boldsymbol{\eta}'\mathbf{x}_j \leq \boldsymbol{\eta}'\mathbf{x}_r)d\boldsymbol{\eta} \\
&= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{n} g_k(\mathbf{z}_i, \hat{\boldsymbol{\theta}})g_k(\mathbf{z}_j, \hat{\boldsymbol{\theta}})A_{ijr},
\end{aligned}$$

where

$$A_{ijr} = A'_{ijr}\frac{\pi^{d/2-1}}{\Gamma(d/2+1)} \quad \text{and} \quad A'_{ijr} = \left|\pi - \arccos(\frac{(\mathbf{x}_i - \mathbf{x}_r)'(\mathbf{x}_j - \mathbf{x}_r)}{|(\mathbf{x}_i - \mathbf{x}_r)|\,|(\mathbf{x}_j - \mathbf{x}_r)|})\right|.$$

The computation can be made simpler under some cases. For instance, if $\mathbf{x}_i = \mathbf{x}_j$ and $\mathbf{x}_i \neq \mathbf{x}_r$, then $A'_{ijr} = \pi$. If $\mathbf{x}_i = \mathbf{x}_j$ and $\mathbf{x}_i = \mathbf{x}_r$, then $A'_{ijr} = 2\pi$. When $\mathbf{x}_i \neq \mathbf{x}_j$ and

$\mathbf{x}_i = \mathbf{x}_r$ or $\mathbf{x}_j = \mathbf{x}_r$, we have $A'_{ijr} = \pi$ (Escanciano, 2006).

Using a resampling method that is similar to the one in Section 2.3.5, we calculate $\mathrm{GF}_k^{(q)}(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}}) = n^{-1/2} \sum_{i=1}^n v_i^{(q)} \{ g_k(\mathbf{z}_i, \hat{\boldsymbol{\theta}}) \mathbf{1}(\boldsymbol{\eta}' \mathbf{x}_i \leq u) - \hat{D}_k(\boldsymbol{\eta}, u) \mathbf{g}(\mathbf{z}_i, \hat{\boldsymbol{\theta}}) \}$ for $k = 1, \cdots, r$, where $\hat{D}_k(\boldsymbol{\eta}, u) = \{ n^{-1} \sum_{i=1}^n \partial_{\boldsymbol{\theta}} g_k(\mathbf{z}_i, \hat{\boldsymbol{\theta}})' \mathbf{1}(\boldsymbol{\eta}' \mathbf{x}_i \leq u) \} S_{n,22.1}^{-1} S_{n,21} S_{n,11}^{-1}$. For simplicity, we define $\hat{g}_k(\mathbf{z}_i) = g_k(\mathbf{z}_i, \hat{\boldsymbol{\theta}})$, $\hat{\mathbf{g}}(\mathbf{z}_i) = \hat{\mathbf{g}}(\mathbf{z}_i, \hat{\boldsymbol{\theta}})$ and $\tilde{S} = S_{n,22.1}^{-1} S_{n,21} S_{n,11}^{-1}$. Therefore, with some calculations, we have

$$\mathrm{PCvM}_{n,k}^{(q)} = \int_{\Pi} [\mathrm{GF}_k^{(q)}(\boldsymbol{\eta}, u; \hat{\boldsymbol{\theta}})]^2 F_{n,\boldsymbol{\eta}}(du) d\boldsymbol{\eta} = \mathrm{I}^{(q)} - 2 \times \ \mathrm{II}^{(q)} + \ \mathrm{III}^{(q)},$$

where

$$\begin{aligned}
\mathrm{I}^{(q)} &= \int_{\Pi} [\sum_{i=1}^n v_i^{(q)} \hat{g}_k(\mathbf{z}_i) \mathbf{1}(\boldsymbol{\eta}' \mathbf{x}_i \leq u)]^2 F_{n,\boldsymbol{\eta}}(du) d\boldsymbol{\eta} \\
&= n^{-1} \sum_{i=1}^n \sum_{j=1}^n v_i^{(q)} \hat{g}_k(\mathbf{z}_i) v_j^{(q)} \hat{g}_k(\mathbf{z}_j) \int_{\Pi} \mathbf{1}(\boldsymbol{\eta}' \mathbf{x}_i \leq u) \mathbf{1}(\boldsymbol{\eta}' \mathbf{x}_j \leq u) F_{n,\boldsymbol{\eta}}(du) d\boldsymbol{\eta} \\
&= n^{-2} \sum_{i=1}^n \sum_{j=1}^n \sum_{r=1}^n v_i^{(q)} v_j^{(q)} \hat{g}_k(\mathbf{z}_i) \hat{g}_k(\mathbf{z}_j) A_{ijr},
\end{aligned}$$

$$\begin{aligned}
\mathrm{II}^{(q)} &= n^{-1} \int_{\Pi} \sum_{i=1}^n v_i^{(q)} \hat{g}_k(\mathbf{z}_i) \mathbf{1}(\boldsymbol{\eta}' \mathbf{x}_i \leq u) \sum_{j=1}^n v_j^{(q)} n^{-1} \sum_{m=1}^n \partial_{\boldsymbol{\theta}} \hat{g}_k(\mathbf{z}_m)' \mathbf{1}(\boldsymbol{\eta}' \mathbf{x}_m \leq u)) \tilde{S} \hat{\mathbf{g}}(\mathbf{z}_j) \\
&\quad F_{n,\boldsymbol{\eta}}(du) d\boldsymbol{\eta} \\
&= n^{-2} \sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^n v_i^{(q)} \hat{g}_k(\mathbf{z}_i) v_j^{(q)} \partial_{\boldsymbol{\theta}} \hat{g}_k(\mathbf{z}_m)' \tilde{S} \hat{\mathbf{g}}(\mathbf{z}_j) \int_{\Pi} \mathbf{1}(\boldsymbol{\eta}' \mathbf{x}_i \leq u) \mathbf{1}(\boldsymbol{\eta}' \mathbf{x}_m \leq u) \\
&\quad F_{n,\boldsymbol{\eta}}(du) d\boldsymbol{\eta} \\
&= n^{-3} \sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^n \sum_{r=1}^n v_i^{(q)} v_j^{(q)} \hat{g}_k(\mathbf{z}_i) \partial_{\boldsymbol{\theta}} \hat{g}_k(\mathbf{z}_m)' \tilde{S} \hat{\mathbf{g}}(\mathbf{z}_j) A_{imr},
\end{aligned}$$

$$\begin{aligned}
\mathrm{III}^{(q)} &= n^{-1}\int_{\Pi}[\sum_{i=1}^{n} v_i^{(q)} n^{-1}\sum_{m=1}^{n}\partial_{\boldsymbol{\theta}}\hat{g}_k(\mathbf{z}_m)^{'}\mathbf{1}(\boldsymbol{\eta}^{'}\mathbf{x}_m \leq u))\tilde{S}\hat{\mathbf{g}}(\mathbf{z}_i)]^2 F_{n,\boldsymbol{\eta}}(du)d\boldsymbol{\eta} \\
&= n^{-3}\sum_{i=1}^{n}\sum_{j=1}^{n} v_i^{(q)}\sum_{m=1}^{n}\partial_{\boldsymbol{\theta}}\hat{g}_k(\mathbf{z}_m)^{'}\tilde{S}\hat{\mathbf{g}}(\mathbf{z}_i)v_j^{(q)}\sum_{l=1}^{n}\partial_{\boldsymbol{\theta}}\hat{g}_k(\mathbf{z}_l)^{'}\tilde{S}\hat{\mathbf{g}}(\mathbf{z}_j) \\
&\quad \times\int_{\Pi}\mathbf{1}(\boldsymbol{\eta}^{'}\mathbf{x}_m \leq u)\mathbf{1}(\boldsymbol{\eta}^{'}\mathbf{x}_l \leq u)F_{n,\boldsymbol{\eta}}(du)d\boldsymbol{\eta} \\
&= n^{-4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{m=1}^{n}\sum_{l=1}^{n}\sum_{r=1}^{n} v_i^{(q)}v_j^{(q)}\partial_{\boldsymbol{\theta}}\hat{g}_k(\mathbf{z}_m)^{'}\tilde{S}\hat{\mathbf{g}}(\mathbf{z}_i)\partial_{\boldsymbol{\theta}}\hat{g}_k(\mathbf{z}_l)^{'}\tilde{S}\hat{\mathbf{g}}(\mathbf{z}_j)A_{mlr}.
\end{aligned}$$

Chapter 5

Intrinsic Regression Models for Medial Representation of Subcortical Structures

5.1 Introduction

Statistical shape modeling and analysis have become important tools for understanding the geometric variability of the anatomical structures in various neuroimaging studies. Statistical shape models provide an efficient description (or measurement) of the morphology of the cortical and subcortical structures (e.g., hippocampus). For instance, linear shape models including the active shape model and landmark method describe shape changes as a combination of local translations (Bookstein, 1986; Cootes et al., 1995). The medial representation (m-rep) of shape provides a useful framework for describing shape variability in local thickness, bending, and widening (Fletch et al., 2004). Statistical analysis of these shape models are crucial for characterizing differences in brain structure across groups of healthy individuals and persons with various diseases, and changes of brain structure across time (Thompson and Toga, 2002; Thompson et al., 2002; Chung et al., 2005; Styner et al., 2005; Zhu et al., 2007b).

In the m-rep framework, a geometric object is represented as a set of connected continuous medial primitives, called medial atoms (See Figure 5.1 for a hippocampus example). For 3-dimensional objects, these medial atoms are formed by the centers of the inscribed spheres and by the associated spokes from the sphere centers to the two respective tangent points on the object boundary. Specifically, a medial atom $\mathbf{m} = (\mathbf{O}, r, \mathbf{n}_0, \mathbf{n}_1)$ is formed by a position $\mathbf{O}$, the center of the inscribed sphere; a radius r, the common spoke length; and $(\mathbf{n}_0, \mathbf{n}_1)$, the two unit spoke directions (Styner et al., 2004). A medial atom can be regarded as a point on a Riemannian manifold, $M(1) = R^3 \times R^+ \times S^2 \times S^2$, where S^2 is the sphere in R^3 with radius one. An m-rep model consisting of k medial atoms can be described as the direct product of k copies of $M(1)$, i.e., $M(k) = \prod_{i=1}^k M(1)$. The existing statistical analytical methods for the m-rep include principal geodesic analysis, the estimation of extrinsic and intrinsic means, and a permutation test for comparing m-rep data from two groups (Fletch et al., 2004; Styner et al., 2007). The scientific interests of some neuroimaging studies, however, typically focus on establishing the associations between a set of covariates, particularly diagnostic status, age, and gender, and shape differences in a population, thus requiring a regression modeling framework for m-rep models.

There are several important issues including multiple directions on S^2 and the correlation structure among different components of $M(1)$ in developing m-rep regression models with a set of covariates. Although there is a sparse literature on regression modeling of a single directional observation from each subject (Mardia and Jupp, 1983; Jupp and Mardia, 1989), these regression models of directional data are based on particular parametric distributions, such as the von Mises-Fisher distribution. For instance, existing circular regression models assume that the angular response follows the von Mises-Fisher distribution with either the angular mean η_i or the concentration parameter κ_i being associated with the covariates $\mathbf{x}_i$ (Gould, 1969; Johnson and Wehrly,

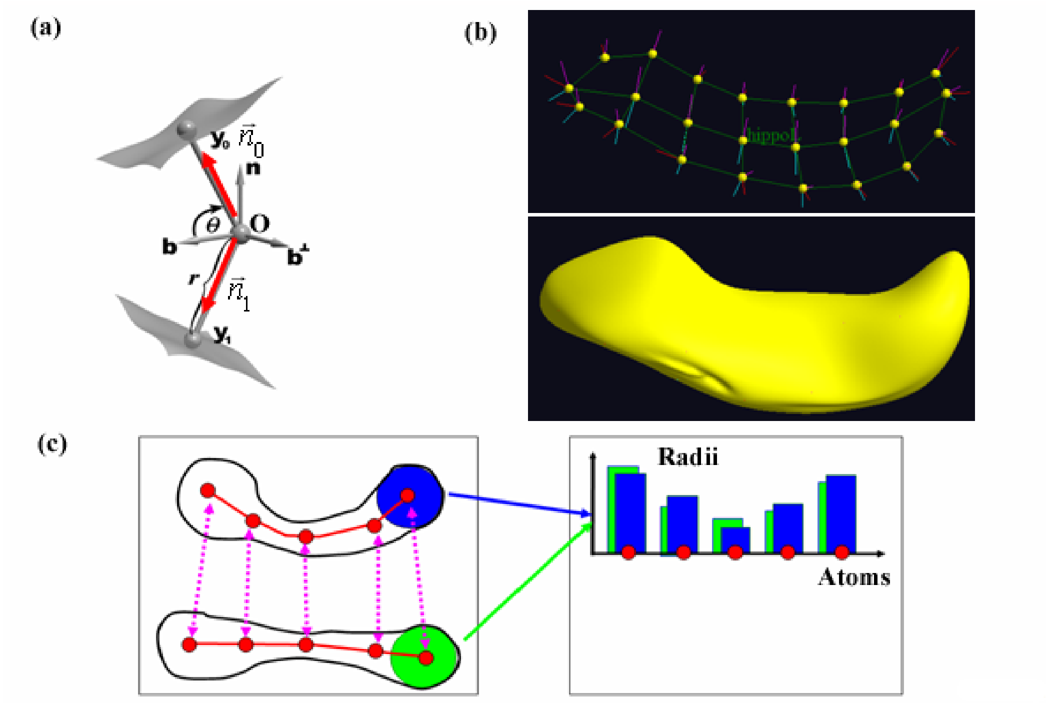

Figure 5.1: (a) An m-rep model; (b) a skeleton and the smoothed surface of an m-rep model of the hippocampus; (c) m-rep radius comparison at the five atoms between two m-rep objects

1978; Fisher and Lee 1992). This circular regression model can be generalized to high-dimensional spherical data using the Fisher-Bingham family (Mardia, 1975; Mardia and Jupp, 1983). Furthermore, the spherically projected linear model for directional data assumes an offset normal distribution (Presnell, Morrison, and Littell 1998). However, it remains unknown whether it is appropriate to use these parametric models for a single directional measure to simultaneously characterize the two spoke directions at each atom, which are correlated among themselves. Moreover, the two spoke directions may be correlated with other components of each atom and this provides futher challenges in modeling the dependence structure of all components at each atom.

This paper develops a semiparametric regression model with m-rep as responses on a Riemannian manifold and covariates in Euclidean space. Our regression model avoids specifying any parametric distributions. We propose an estimation procedure based on the annealing evolutionary stochastic approximation Monte Carlo (AESAMC) algorithm to obtain regression coefficient estimates in this semi-parametric model. We establish asymptotic properties, including consistency and asymptotic normality, of the estimates of the regression coefficients, and develop Wald statistics to test linear hypotheses of unknown parameters.

The rest of this paper is organized as follows. In Section 5.2, we formulate the semiparametric regression model, introduce the AESAMC algorithm for estimating the regression coefficients, establish asymptotic properties of our estimates, and then develop Wald statistics to carry out hypothesis testing. Simulation studies in Section 5.3 are used to assess the finite sample performance of the parameter estimates, the AESAMC algorithm and Wald test statistics. In Section 5.4, we illustrate the application of our statistical methods to the detection of the difference in morphological changes of the hippocampi between schizophrenia patients and healthy controls in an MRI study of schizophrenia, before making some concluding remarks in Section 5.5.

5.2 Theory

5.2.1 Model

We formally introduce a semiparametric regression model for m-rep data from n subjects. Suppose we have an exogenous $q \times 1$ covariate vector $\mathbf{x}_i$ for the $i-$th subject and m-rep measures, denoted by $\mathbf{M}_i = \{\mathbf{m}_i(d) : d \in \mathcal{D}\}$, for the $i-$th subject, where d represents an atom of an m-rep on $\mathcal{D}$, a specific brain region. For notational simplicity, we temporarily drop atom d from our notation.

The regression model involves modeling a 'conditional mean' of an m-rep response at an atom $\mathbf{m}_i$ given $\mathbf{x}_i$, denoted by $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = \boldsymbol{\mu}(\mathbf{x}_i, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients in $\mathcal{B} \subset R^p$. Thus, $\boldsymbol{\mu}(\cdot, \cdot)$ is a map from $R^q \times R^p$ to $M(1)$ and $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = (\boldsymbol{\mu}_{Oi}(\boldsymbol{\beta}), \mu_{ri}(\boldsymbol{\beta}), \boldsymbol{\mu}_{0i}(\boldsymbol{\beta}), \boldsymbol{\mu}_{1i}(\boldsymbol{\beta}))'$, which is a 10×1 vector and $\boldsymbol{\mu}_{Oi}(\boldsymbol{\beta})$, $\mu_{ri}(\boldsymbol{\beta})$, $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$, and $\boldsymbol{\mu}_{1i}(\boldsymbol{\beta})$ are the 'conditional means' of the location O_i, the radius r_i, and the two spoke directions $\mathbf{n}_{0i}$ and $\mathbf{n}_{1i}$ respectively, given $\mathbf{x}_i$, for the i-th subject. Note that we just borrow the term 'conditional mean' from Euclidean space.

We need to formalize this notion of 'conditional mean' explicitly. For the location component of an m-rep, we may set $\boldsymbol{\mu}_{Oi}(\boldsymbol{\beta}) = (g(\mathbf{x}_i, \boldsymbol{\beta}_1), g(\mathbf{x}_i, \boldsymbol{\beta}_2), g(\mathbf{x}_i, \boldsymbol{\beta}_3))'$, where $\boldsymbol{\beta}_k$ $(k = 1, 2, 3)$ is a $p_k \times 1$ coefficient vector. There are many different ways of specifying $g(\mathbf{x}_i, \boldsymbol{\beta}_k)$. The simplest one is the linear link function $g(\mathbf{x}_i, \boldsymbol{\beta}_k) = \mathbf{x}_i' \boldsymbol{\beta}_k$. We may also represent $g(\mathbf{x}_i, \boldsymbol{\beta}_k)$ as a linear combination of basis functions $\{\psi_j(\mathbf{x}_i) : j = 1, \cdots, J\}$ (such as B-splines), that is $g(\mathbf{x}_i, \boldsymbol{\beta}_k) = \sum_{j=1}^{J} \psi_j(\mathbf{x}_i)\boldsymbol{\beta}_{kj}$. For the radius component, we may use $\mu_{ri}(\boldsymbol{\beta}) = g(\mathbf{x}_i, \boldsymbol{\beta}_4)$, where $\boldsymbol{\beta}_4$ is a $p_4 \times 1$ coefficient vector for an m-rep radius. Since a radius is always positive, a natural link function is $g(\mathbf{x}_i, \boldsymbol{\beta}_k) = \exp(\mathbf{x}_i' \boldsymbol{\beta}_k)$, among other possible choices.

As for the two directions on an m-rep, they are more complex and will be our focus here. In the existing literature, the circular regression models (Gould, 1969; Johnson and Wehrly, 1978; Fisher and Lee, 1992) assume that the angular representation of a direction follows the von Mises distribution with either the angular mean η_i or the concentration parameter κ_i associated with $\mathbf{x}_i$. Gould's (1969) regression model for a circular response takes the form of $\eta_i = \eta + \mathbf{x}_i' \boldsymbol{\beta}$, where η_i is the angular representation of the circular response for the $i-$th subject and $(\eta, \boldsymbol{\beta})$ are unknown parameters. A major critism of Gould's model is its identifiability problem, that is, the likelihood function has infinitely many maxima of comparable size (Fisher and Lee, 1992; Presnell, Morrison, and Littell, 1998). To avoid this problem, Fisher and Lee (1992) replaced the linear link function by a suitable one-to-one function $g : (-\infty, \infty) \to (-\pi, \pi)$ satisfying $g(0) = 0$. Two such link functions are the inverse tangent link, $g(x) = 2\arctan(x)$, and the scaled probit link, $g(x) = 2\pi[\Phi(x) - 0.5]$, where arctan is the inverse of the tangent function. Johnson and Wehrly (1978) used the link function $g(x) = 2\pi F(x)$, where F is a cumulative distribution function. These link functions can be generalized to spherical data as detailed below.

We now develop the link functions for a spherical response. For notational simplicity, we use $\mathbf{n}_0$ as an example throughtout. We need to specify the explicit form of $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$, the 'conditional mean' function of $\mathbf{n}_{0i}$ for the i-th subject. We can use spherical polar coordinates to represent $\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$ as

$$\boldsymbol{\mu}_{0i}(\boldsymbol{\beta}) = \begin{pmatrix} \cos(\phi_i) \\ \sin(\phi_i)\cos(\eta_i) \\ \sin(\phi_i)\sin(\eta_i) \end{pmatrix}, \tag{5.1}$$

where ϕ_i denotes the colatitude (so that $\pi/2 - \phi_i$ is the latitude) and η_i denotes the longitude for the $i-$th subject. Following Fisher and Lee (1992), we may assume that

$$\begin{aligned} \phi_i &= \mathbf{x}_{i,d}^{'}\boldsymbol{\beta}_{1d} + \arctan(\mathbf{x}_i^{'}\boldsymbol{\beta}_{1c}), \\ \eta_i &= \mathbf{x}_{i,d}^{'}\boldsymbol{\beta}_{2d} + 2\text{arctan}(\mathbf{x}_i^{'}\boldsymbol{\beta}_{2c}), \end{aligned} \tag{5.2}$$

where $\mathbf{x}_{i,d}$ includes all the discrete covariates and the intercept, and $\mathbf{x}_{i,c}$ are all the centered continuous covariates.

So far, we have defined link functions for all the components of an m-rep. Now, we introduce a definition of a 'residual' to ensure that $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ is the proper 'conditional mean' of $\mathbf{m}_i$ given $\mathbf{x}_i$. For instance, in the classical linear model, the response is the sum of the regression function and the residual. Then, the regression function is the conditional mean of the response only when the conditional mean of residual equals zero. Given two points $\mathbf{m}_i$ and $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ on the manifold, we need to define the residual or 'difference' between them. At $\boldsymbol{\mu}_i(\boldsymbol{\beta})$, we have the tangent space of the manifold, denoted by $T_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}M(1)$, which is a Euclidean space representing a first order approximation of the manifold $M(1)$ near $\boldsymbol{\mu}_i(\boldsymbol{\beta})$. Then we calculate the projection of $\mathbf{m}_i$ onto $T_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}M(1)$, denoted by $\text{Log}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i)$, which is given by

$$\text{Log}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i) = (O_i - \boldsymbol{\mu}_{Oi}(\boldsymbol{\beta}), \log(r_i/\mu_{ri}(\boldsymbol{\beta})), \text{Log}_{\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})}(\mathbf{n}_{0i}), \text{Log}_{\boldsymbol{\mu}_{1i}(\boldsymbol{\beta})}(\mathbf{n}_{1i})), \tag{5.3}$$

where $\text{Log}_{\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})}(\mathbf{n}_{0i}) = \arccos(\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})'\mathbf{n}_{0i})\mathbf{v}/||\mathbf{v}||$, in which $\mathbf{v} = \mathbf{n}_{0i} - (\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})'\mathbf{n}_{0i})\boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$ and $\|\cdot\|$ is the Euclidean norm. Thus, $\text{Log}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i)$ can be regarded as the difference between $\mathbf{m}_i$ and $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ in $T_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}M(1)$. Since $\text{Log}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i)$ are in different tangent spaces, we must translate them to the same tangent space. We can use a rotation matrix, $R_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}$, to translate $\text{Log}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i) \in T_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}M(1)$ into the same tangent space, say $T_{P_0}M(1)$, in which $P_0 = (0,0,0,1,0,0,1,0,0,1)'$, and define $\mathcal{E}_i(\boldsymbol{\beta}) = R_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}\text{Log}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i)$ for $i = 1, \cdots, n$.

5.2.2 Geometric structure of an m-rep manifold

We summarize some basic results about the geometric structure of an m-rep as a Riemannian manifold. We first introduce tangent vectors and tangent spaces at any $p_0 \in M(1)$. For a small scalar $\delta > 0$, let $p(t)$ be a differentiable map from $(-\delta, \delta)$ to $M(1)$ such that it passes through $p(0) = p_0$. The tangent vector at p_0 is defined as the derivative of the smooth curve $p(t)$ with respect to t. The set of all tangent vectors at p_0 forms the tangent sapce of $M(1)$ at p_0, denoted by $T_{p_0}M(1)$. Secondly, we use the Euclidean norm as the inner product of any two tangent vectors in the same tangent space, which varies smoothly along the manifold.

Let $\gamma_{p_0}(t;\theta)$ be the geodesic on $M(1)$ passing through p_0 in the direction of the tangent vector $\theta \in T_{p_0}M(1)$. The Riemannian Exponential map, denoted by $\text{Exp}_{p_0}(\cdot)$, maps the tangent vector θ at p_0 to a point $p_1 \in M(1)$ and $\text{Exp}_{p_0}(\theta) = \gamma_{p_0}(1;\theta)$; the Riemannian Logarithm map, denoted by $\text{Log}_{p_0}(p_1)$, maps $p_1 \in M(1)$ onto the tangent vector $\theta \in T_{p_0}M(1)$. The Exponential map and Logarithm map are inverses of each other.

Because an m-rep is the product of several spaces, i.e., $M(1) = R^3 \times R^+ \times S^2 \times S^2$, the Exponential/Logarithm map for $M(1)$ is the product of the Exponential/Logarithm map for each space. Let $p_0 = (p_{01}, p_{02}, p_{03}, p_{04})'$ and $p_1 = (p_{11}, p_{12}, p_{13}, p_{14})'$, where $p_{01} \in R^3$ and $p_{11} \in R^3$ are the location components, $p_{02} \in R^+$ and $p_{12} \in R^+$ are the radius components, $p_{03} \in S^2$ and $p_{13} \in S^2$ are the first direction components, and $p_{04} \in S^2$ and $p_{14} \in S^2$ are the second direction components. Further, let the tangent vector $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)^{'}$, where $\theta_1 \in R^3$ is the locational tangent component, $\theta_2 \in R$ is the radius tangent component, $\theta_3 \in R^3$ and $\theta_4 \in R^3$ are the two directional tangent components. We discuss the calculation of the Exponential and Logarithm maps for each space of interest. For R^3, the tangent space is R^3 itself, and $\mathrm{Exp}_{p_{01}}(\theta_1) = p_{01} + \theta_1$, whereas $\mathrm{Log}_{p_{01}}(p_{11}) = p_{11} - p_{01}$. For the space R^+, the non-negative real numbers, the tangent space is R and $\mathrm{Exp}_{p_{02}}(\theta_2) = p_{02}\exp(\theta_2)$, whereas $\mathrm{Log}_{p_{02}}(p_{12}) = \log(p_{12}/p_{02})$, which involve the classical exponential and logarithm functions. For the space S^2,

$$\mathrm{Exp}_{p_{03}}(\theta_3) = cos(\|\theta_3\|)p_{03} + \sin(\|\theta_3\|)\theta_3 \|\theta_3\| .$$

Let $\upsilon = p_{13} - (p_{03}^{'}p_{13})p_{03} \neq 0$. If $p_{03},\ p_{13}$ are not antipodal ($p_{03} \neq -p_{13}$), then we can get

$$\mathrm{Log}_{p_{03}}(p_{13}) = \arccos(p_{03}^{'}p_{13}) \cdot \upsilon / \|\upsilon\|_2 .$$

Thus, for $M(1)$, we have the Exponential map as

$$\mathrm{Exp}_{p_0}(\theta) = (p_{01} + \theta_1, p_{02}\exp(\theta_2), \mathrm{Exp}_{p_{03}}(\theta_3), \mathrm{Exp}_{p_{04}}(\theta_4)). \tag{5.4}$$

Likewise, the Logarithm map for $M(1)$ is

$$\mathrm{Log}_{p_0}(p_1) = (p_{11} - p_{01}, \log(p_{12}/p_{02}), \mathrm{Log}_{p_{03}}(p_{13}), \mathrm{Log}_{p_{04}}(p_{14})). \tag{5.5}$$

A nice property of $M(1)$ is that a group action of G on $M(1)$ can relate any two points $p_0,\ p_1 \in M(1)$ and the tangent spaces at p_0 and p_1. Specifically, the group action of G $= R^3 \times R^+ \times SO(3) \times SO(3)$, where $SO(3)$ denotes the 3×3 rotation matrices.

For an element of G, $g = (g_1, g_2, g_3, g_4)'$, the group action on any point p_0 is given by

$$g(p_0) = (p_{01} + g_1, g_2 \cdot p_{02}, g_3 \cdot p_{03}, g_4 \cdot p_{04}). \tag{5.6}$$

For any $p_0, p_1 \in M(1)$, there exists a $g \in \mathrm{G}$ such that $g(p_0) = p_1$. The group action of G on $M(1)$ induces a group action between $T_{p_0}M(1)$ and $T_{G_g(p_0)}M(1)$. Explicitly, if $\theta \in T_{p_0}M(1)$, then $g(\theta) \in T_{G_g(p_0)}M(1)$.

We consider the geodesic distance between any two points $p_0,\ p_1 \in M(1)$. The geodesic distance between p_0 and p_1 is uniquely given by

$$d(p_0, p_1) = \sqrt{(\mathrm{Log}_{p_0}(p_1)'\mathrm{Log}_{p_0}(p_1)} = \left\|\mathrm{Log}_{p_0}(p_1)\right\|. \tag{5.7}$$

The geodesic distance has many nice properties. For instance, the geodesic distance is a proper metric satisfying positive definiteness, symmetry, and the triangle inequality. Particularly, $d(p_0, p_1) = d(p_1, p_0)$. The geodesic distance is also invariant under group actions, that is $d(p_0, p_1) = d(g(p_0), g(p_1))$ for any $g \in \mathrm{G}$.

5.2.3 Estimation

The square of the geodesic distance between $\mathbf{m}_i$ and $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ equals

$$d^2(\mathbf{m}_i, \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathrm{Log}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i)'\mathrm{Log}_{\boldsymbol{\mu}_i(\boldsymbol{\beta})}(\mathbf{m}_i). \tag{5.8}$$

We calculate an intrinsic least squares estimator of the parameter $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$, by minimizing the square of the distance,

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} D_n(\boldsymbol{\beta}) = \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} D_{ni}(\boldsymbol{\beta}) = \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} d^2(\mathbf{m}_i, \boldsymbol{\mu}_i(\boldsymbol{\beta})). \tag{5.9}$$

We do not assume a parametric distribution for $\mathbf{m}_i$ given $\mathbf{x}_i$, and thus we allow for a large class of distributions. Because the true distribution may deviate from any parametric distribution, it may be very stringent to assume a parametric distribution, such as a Gaussian distribution for the error. In addition, we do not need homogeneous variances across all i. This is desirable for real applications, since the between-subject and between-atom variabilities in the imaging measures can be substantial.

We now develop an annealing evolutionary stochastic approximation Monte Carlo (AESAMC) algorithm for obtaining $\hat{\boldsymbol{\beta}}$. Quite recently, the stochastic approximation Monte Carlo (SAMC) algorithm (Liang et al., 2007) has been proposed in the literature as a general simulation technique, which possesses a nice feature in that the moves are self-adjustable and thus not likely to get trapped by local energy minima. The annealing evolutionary SAMC (AESAMC) algorithm (Liang, 2008) represents a further improvement of SAMC for optimization problems by incorporating some features of simulated annealing (Kirkpatrick et al., 1983) and the genetic algorithm (Goldberg, 1998) into its search process. AESAMC has been tested on a large number of benchmark optimization problems and produced favorable results in comparison with many other optimization metaheuristics, such as simulated annealing, the genetic algorithm (Ali et al., 2005), continuous GRASP (Hirsch et al., 2007), tabu search (Hedar and Fukushima, 2006), and the scatter search (Laguna and Marti, 2005). The AESAMC algorithm can be described as follows.

Like the genetic algorithm, AESAMC works on a population of samples. Let $\boldsymbol{\beta}^l = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_l)$ denote the population, where l is the population size, and $\boldsymbol{\beta}_i = (\boldsymbol{\beta}_{i1}, \ldots, \boldsymbol{\beta}_{ip})$ is a p-dimensional vector called an individual or chromosome in terms of genetic algorithms. Thus, the minimum of the objective function $D_n(\boldsymbol{\beta})$, $\boldsymbol{\beta} \in \mathcal{B}$, can be obtained by minimizing the function $\boldsymbol{U}(\boldsymbol{\beta}^l) = \sum_{i=1}^{l} D_n(\boldsymbol{\beta}_i)$. An unnormalized Boltzmann density can be defined for the population as follows,

$$\boldsymbol{\psi}(\boldsymbol{\beta}^l) = \exp\{-\boldsymbol{U}(\boldsymbol{\beta}^l)/\tau\}, \quad \boldsymbol{\beta}^l \in \mathcal{B}^l, \tag{5.10}$$

where $\tau = 1$ is called the temperature, and $\mathcal{B}^l = \mathcal{B} \times \cdots \times \mathcal{B}$ is a product sample space.

The sample space can be partitioned according to the function $\boldsymbol{U}(\boldsymbol{\beta}^l)$ into b subregions: $\mathbb{E}_1 = \{\boldsymbol{\beta}^l : \boldsymbol{U}(\boldsymbol{\beta}^l) \leq u_1\}, \mathbb{E}_2 = \{\boldsymbol{\beta}^l : u_1 < \boldsymbol{U}(\boldsymbol{\beta}^l) \leq u_2\}, \ldots, \mathbb{E}_{b-1} = \{\boldsymbol{\beta}^l : u_{b-2} < \boldsymbol{U}(\boldsymbol{\beta}^l) \leq u_{b-1}\}$, and $\mathbb{E}_b = \{\boldsymbol{\beta}^l : \boldsymbol{U}(\boldsymbol{\beta}^l) > u_{b-1}\}$, where $u_1 < u_2 < \ldots < u_{b-1}$ are $b-1$ known real numbers. We note that here the sample space is not necessarily partitioned according to the function $\boldsymbol{U}(\boldsymbol{\beta}^l)$, for example, the function $\lambda(\boldsymbol{\beta}^l) = \min\{D_n(\boldsymbol{\beta}_1), \ldots, D_n(\boldsymbol{\beta}_l)\}$ also works.

Let $\varpi(u)$ denote the index of the subregion that a sample with energy u belongs to. For example, if $\boldsymbol{\beta}^l \in \mathbb{E}_j$, then $\varpi(\boldsymbol{U}(\boldsymbol{\beta}^l)) = j$. Let $\mathcal{B}^{(t)}$ denote the sample space at iteration t. AESAMC initiates its search in the entire sample space $\mathcal{B}_0 = \bigcup_{i=1}^{b} \mathbb{E}_i$, and then iteratively searches in the set

$$\mathcal{B}_t = \bigcup_{i=1}^{\varpi(\boldsymbol{U}_{\min}^{(t)} + \aleph)} \mathbb{E}_i, \quad t = 1, 2, \ldots, \tag{5.11}$$

where $\boldsymbol{U}_{\min}^{(t)}$ is the best function value obtained until iteration t, and $\aleph > 0$ is a user specified parameter which determines the broadness of the sample space at each iteration. Note that in AESAMC, the sample space shrinks iteration by iteration. To ensure the convergence of the algorithm to the set of global minima, the moves at each itertaion are required to admit the following distribution as the invariant distribution,

$$\boldsymbol{f}_{\theta^{(t)}}(\boldsymbol{\beta}^l) \propto \sum_{i=1}^{\varpi(\boldsymbol{U}_{\min}^{(t)} + \aleph)} \frac{\boldsymbol{\psi}(\boldsymbol{\beta}^l)}{e^{\theta_i^{(t)}}} I(\boldsymbol{\beta}^l \in \mathbb{E}_i), \quad \boldsymbol{\beta}^l \in \mathcal{B}_t^l, \tag{5.12}$$

where $\theta_i^{(t)}$ are the working parameters which will be updated from itertaion to iteration as described in the algorithm below.

AESAMC includes five types of moves, the MH-Gibbs mutation, K-point mutation, K-point crossover, snooker crossover, and linear crossover operators. See Liang (2008) for the details of the moves. Let $\rho_1, \ldots, \rho_5$, $0 < \rho_i < 1$ and $\sum_{i=1}^{5} \rho_i = 1$, denote the respective working probabilities of the five types of moves. The AESAMC algorithm can be summarized as follows.

AESAMC algorithm:

(a) (Initialization) Partition the sample space $\mathcal{B}^l$ into b disjoint subregions $\mathbb{E}_1, \ldots, \mathbb{E}_b$; choose the threshold value $\aleph$ and the working probabilities $\rho_1, \ldots, \rho_5$; initialize a population $\boldsymbol{\beta}^{l(0)}$ at random; and set $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_b^{(0)}) = (0, 0, \ldots, 0)$, $\mathcal{B}_0^l = \bigcup_{i=1}^b \mathbb{E}_i$, $\boldsymbol{U}_{\min}^{(0)} = \boldsymbol{U}(\boldsymbol{\beta}^{l(0)})$ and $t = 0$.

(b) (Sampling) Update the current population $\boldsymbol{\beta}^{l(t)}$ using the MH-Gibbs mutation, K-point mutation, K-point crossover, snooker crossover, and linear crossover operators according to the respective working probabilities.

(c) (Working weight updating) Update the working weight $\theta^{(t)}$ by setting

$$\theta_i^* = \theta_i^{(t)} + \gamma_{t+1} H_i(\theta^{(t)}, \boldsymbol{\beta}^{l(t+1)}), \quad i = 1, \ldots, \varpi(\boldsymbol{U}_{\min}^{(t)} + \aleph),$$

where $H_i(\theta^{(t)}, \boldsymbol{\beta}^{l(t+1)}) = I(\boldsymbol{\beta}^{l(t+1)} \in \mathbb{E}_i)$ for the crossover operators, $H_i(\theta^{(t)}, \boldsymbol{\beta}^{l(t+1)}) = \sum_{j=1}^l I(\boldsymbol{\beta}^{l(t+1,j)} \in \mathbb{E}_i)/l$ for the mutation operators, and γ_{t+1} is called the gain factor. If $\theta^* \in \Theta$, set $\theta^{(t+1)} = \theta^*$; otherwise, set $\theta^{(t+1)} = \theta^* + \boldsymbol{c}^*$, where $\boldsymbol{c}^* = (c^*, \ldots, c^*)$ and c^* is chosen such that $\theta^* + \boldsymbol{c}^* \in \Theta$.

(d) (Termination Checking) Check the termination condition, e.g., whether a fixed number of iterations has been reached. Otherwise, set $t \to t + 1$ and go to step (b).

In this article, we follow Liang (2008) to set $\rho_1 = \rho_2 = 0.05$, $\rho_3 = \rho_4 = \rho_5 = 0.3$, and the gain factor sequence

$$\gamma_t = \frac{t_0}{\max(t_0, t)}, \quad t = 0, 1, 2, \ldots, \tag{5.13}$$

with $t_0 = 5000$. In general, a large value of t_0 will allow the sampler to reach all the subregions very quickly even for a large system. As shown in Liang (2008), AESAMC can converge weakly toward a neighboring set of global minima of $\boldsymbol{U}(\boldsymbol{\beta}^l)$ in the space of energy. More precisely, the sample $\boldsymbol{\beta}^{l(t)}$ converges in distribution to a random population with the density function

$$\boldsymbol{f}_\theta(\boldsymbol{\beta}) \propto \sum_{i=1}^{\varpi(\boldsymbol{U}_{\min}+\aleph)} \frac{\boldsymbol{\psi}(\boldsymbol{\beta}^l)}{\int_{\mathbb{E}_i} \boldsymbol{\psi}(\boldsymbol{\beta}^l) d\boldsymbol{\beta}^l} I(\boldsymbol{\beta}^l \in \mathbb{E}_i), \tag{5.14}$$

where $\boldsymbol{U}_{\min}$ is the global minimum value of $\boldsymbol{U}(\boldsymbol{\beta})$,

Regarding the setting of the other parameters, we have the following suggestions. In AESAMC, the moves are reduced to the Metropolis-Hastings moves (Metropolis et al., 1953; Hastings, 1970) within the same subregions. Hence, the sample space should be partitioned such that the MH moves within the same subregion have a reasonable acceptance rate. In this article, we set $u_{i+1} - u_i \equiv 0.2$ for $i = 1, \ldots, b-1$.

In AESAMC, the crossover operator has been modified to serve as a proposal for the moves, and it is no longer as critical as to the genetic algorithm. Hence, the population size l is usually set to a moderate number, ranging from 10 to 100. Since $\aleph$ determines the size of the neighboring set toward which AESAMC converges, $\aleph$ should be chosen carefully for efficiency of the algorithm. If $\aleph$ is too small, it may take a long time for the algorithm to locate the global minima. In this case, the sample space may contain a lot of separated regions, and most of the proposed transitions will be rejected if the proposal distribution is not spread out enough. If $\aleph$ is too large, it may also take a long time for the algorithm to locate the global energy minimum due to the broadness of the sample space. In practice, the values of l and $\aleph$ can be determined through a trial and error process based on the diagnosis for the convergence of the algorithm. If it fails to converge, the parameters should be tuned to larger values. As suggested by Liang (2008), the convergence of AESAMC can be diagnosed by examining the difference of the patterns of the working weights obtained in multiple runs. In this article, we set $l = 50$ and $\aleph = 50$.

5.2.4 Asymptotic properties

We introduce some notation to present the limiting behavior of our estimates. Let $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}$, $\mathcal{B}$ denote the parameter space for $\boldsymbol{\beta}$, and $||\cdot||$ denote the Euclidean norm of a vector or a matrix.

We establish consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}$. We obtain the following theorems, whose detailed assumptions can be found in the Appendix.

Theorem 1

(a) If assumptions A1, A2, and A3 in the Appendix are true, then $\hat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}_0$ in probability.

(b) Under assumptions A1-A4, we have

$$\{E\sum_{i=1}^{n}[\partial_{\boldsymbol{\beta}} D_{ni}(\hat{\boldsymbol{\beta}})^{\otimes 2}]\}^{-1/2}E[-\partial^2_{\boldsymbol{\beta}} D_n(\hat{\boldsymbol{\beta}})](\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \to^L N(0,\mathbf{I}_p) \tag{5.15}$$

as $n\to\infty$, where $\to^L$ denotes convergence in distribution.

Theorem 1 has several important applications. Theorem 1 (a) establishes consistency of $\hat{\boldsymbol{\beta}}$. According to Theorem 1 (b), the covariance matrix of $\hat{\boldsymbol{\beta}}$ can be consistently estimated by

$$\hat{\Sigma}_{\boldsymbol{\beta}} = \{E[-\partial^2_{\boldsymbol{\beta}} D_n(\hat{\boldsymbol{\beta}})]\}^{-1}\{E\sum_{i=1}^{n}[\partial_{\boldsymbol{\beta}} D_{ni}(\hat{\boldsymbol{\beta}})^{\otimes 2}]\}\{E[-\partial^2_{\boldsymbol{\beta}} D_n(\hat{\boldsymbol{\beta}})]\}^{-1}. \tag{5.16}$$

Moreover, we can use Theorem 1 (b) to construct confidence cones of $\boldsymbol{\beta}$ and its functions. Since Theorem 1 only establishes the asymptotic properties of $\hat{\boldsymbol{\beta}}$ when the sample size is large, these properties may be inadequate to characterize the finite sample behavior of $\hat{\boldsymbol{\beta}}$ for relatively small samples. In the case of small samples, we may have to resort to higher order approximations, such as saddlepoint approximations, and bootstrap methods (Butler, 2007; Davison and Hinkley, 1997).

Test linear hypotheses

Our choices of which hypotheses to test are motivated by scientific questions, which involve a comparison of m-rep components across diagnostic groups. These questions usually can be formulated as testing linear hypotheses of $\boldsymbol{\beta}$ as follows:

$$H_0 : A\boldsymbol{\beta} = \mathbf{b}_0 \quad \text{vs.} \quad H_1 : A\boldsymbol{\beta} \neq \mathbf{b}_0, \tag{5.17}$$

where A is an $r \times p$ matrix of full row rank and $\mathbf{b}_0$ is an $r \times 1$ specified vector.

We test the null hypothesis $H_0 : A\boldsymbol{\beta} = \mathbf{b}_0$ using a Wald test statistic W_n defined by

$$W_n = (A\hat{\boldsymbol{\beta}} - \mathbf{b}_0)^{'} (A\hat{\Sigma}A)^{-1}(A\hat{\boldsymbol{\beta}} - \mathbf{b}_0), \tag{5.18}$$

Theorem 2

If the assumptions in the Appendix are true, then the statistic W_n is asymptotically distributed as $\chi^2(r)$, a chi-square distribution with r degrees of freedom, under the null hypothesis H_0.

An asymptotically valid test can be obtained by comparing sample values of the test statistic with the critical value of an $\chi^2(r)$ distribution at a pre-specified significance level α. That is, we reject H_0 if $W_n \geq \chi^2_\alpha(r)$, and do not reject H_0 otherwise, where $\chi^2_\alpha(r)$ is the upper α-percentile of the $\chi^2(r)$ distribution.

5.3 Simulation studies

The simulation studies presented here focus on directional data. The goal of the first set of simulations was to evaluate the accuracy of the parameter estimates and their associated variance estimates for the proposed intrinsic regression model. The goal of the second set of simulations was to examine the finite sample performance of the Wald statistics.

Table 5.1: Bias ($\times 10^{-3}$), RMS($\times 10^{-2}$), SD($\times 10^{-2}$), and RS of all parameters. Bias denotes the bias of the mean of the estimates; RMS denotes the root-mean-square error; SD denotes the mean of the standard deviation estimates; RS denotes the ratio of RMS over SD.

	$n = 40$				$n = 80$				$n = 120$			
	Bias	RMS	SD	RS	Bias	RMS	SD	RS	Bias	RMS	SD	RS
β_{10}	19.57	31.00	42.78	1.38	3.76	7.99	9.67	1.21	3.21	7.55	8.38	1.11
β_{11}	35.77	53.22	68.65	1.29	1.51	15.08	17.34	1.15	1.01	14.13	14.98	1.06
β_{20}	57.19	46.60	62.44	1.34	3.12	15.88	18.74	1.18	2.39	14.91	15.66	1.05
β_{21}	37.24	33.65	40.72	1.21	2.42	14.03	15.29	1.09	0.40	8.32	8.49	1.02

We generated the simulated data as follows:

$$\mathrm{R}_{\boldsymbol{\mu}_{0i}}\mathrm{Log}_{\boldsymbol{\mu}_{0i}}(\mathbf{n}_{0i}) = \mathcal{E}_i,$$

where $\boldsymbol{\mu}_{0i} = \boldsymbol{\mu}_{0i}(\boldsymbol{\beta})$ are defined in (5.1) and (5.2). We considered having only one covariate of interest, x_i. We generated the x_i independently from a $N(0,1)$ distribution. We generated some errors $\mathcal{E}_i$ independently from a $N_2(0, 0.5 \times \mathbf{1})$ distribution on the tangent space, $T_{P_0}(S^2)$, of the north pole $P_0 = (0,0,1)'$ on a unit sphere, rotated the errors onto the tangent space of $\boldsymbol{\mu}_{0i}$, $T_{\boldsymbol{\mu}_{0i}}(S^2)$, then used the Exp map to get the response $\mathbf{n}_{0i}$. Here, $\boldsymbol{\beta} = (\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21})'$ is a 4×1 vector, with β_{10} and β_{20} being the mean of the two angles of the responses, and β_{11} and β_{21} corresponding to the effect of x_i on each of the two angles. We fixed the true values of $\boldsymbol{\beta}$ to be $\boldsymbol{\beta}_0 = (\pi/2, 1, \pi, 1)'$. We set $n = 40$, 80, and 120 and then simulated 500 datasets for each case to examine the finite sample performance of $\hat{\boldsymbol{\beta}}$ and $\hat{\Sigma}(\hat{\boldsymbol{\beta}})$.

Based on 500 parameter estimates, we calculated the bias, the root-mean-square error (RMS), and the mean of the estimated standard error estimates (SD) (Table 5.1). All relative efficiencies (the ratio of the mean of the standard deviation estimates to the root mean-square error) are close to 1.0, indicating that it is an accurate estimate. As expected, the root mean-square error decreases as the sample size increases.

To examine the finite sample performance of the Wald statistic W_n, we used the same setup as the first set of simulations except that we varied the true value of β_{21}. To assess

Table 5.2: Comparisons of the rejection rates for Wald test statistics.

	$n = 40$		$n = 80$		$n = 120$	
	5%	1%	5%	1%	5%	1%
1	0.085	0.044	0.080	0.042	0.052	0.014
1.4	0.156	0.122	0.304	0.164	0.468	0.210
1.8	0.398	0.162	0.722	0.364	0.918	0.744
2.2	0.586	0.214	0.960	0.778	0.998	0.940

the Type I and II error rates for W_n, we tested the following hypotheses

$$H_0 : \beta_{21} = 1 \quad \text{and} \quad H_1 : \beta_{21} \neq 1.$$

We set five different β_{21} at 1.0, 1.2, 1.4, 1.6, and 1.8, and set n = 40, 80, and 120 and then simulated 500 datasets for each case.

The Wald statistic W_n performs reasonably well for relatively small sample sizes (Table 5.2). The Type I error rates are not too excessive even for both the 5% and 1% significance levels at $n = 40$. Thus, increasing the sample size can increase the power for rejecting the null hypothesis.

5.4 Schizophrenia study of the hippocampus

We consider a neuroimaging dataset about the m-rep shape of the hippocampus structure in the left and right brain hemispheres in schizophrenia patients and healthy controls, collected at 14 academic medical centers in North America and western Europe.

In this study, 238 schizophrenia patients were enrolled who met the following criteria: age 16 to 40 years; onset of psychiatric symptoms before age 35; diagnosis of schizophrenia, schizophreniform, or schizoaffective disorder according to DSM-IV criteria; and various treatment and substance dependence conditions. 56 healthy control subjects were also enrolled.

The aim of our study was to investigate the difference of m-rep shape between schizophrenia patients and healthy controls while controlling for other factors, such as gender and age. The response of interest was the hippocampus m-rep shape at the 24 medial atoms of the left and right brain (Figure 5.1). Covariates of interest were Whole Brain Volume (WBV), race (Caucasian, African American and others), age (in years), gender, and diagnostic status (patient or control).

We used the square of the geodesic distance with $\mathbf{x}_i$ being an 7×1 vector given by $\mathbf{x}_i = (1, \text{gender}_i, \text{age}_i, \text{diag}_i, \text{race1}_i, \text{race2}_i, \text{WBV}_i)'$, where diag is the dummy variable for patients versus healthy controls, and race1 and race2 are, respectively, dummy variables for Caucasians and African Americans versus other races. For the location component on the m-rep, we set $\boldsymbol{\mu}_O(\mathbf{x}, \boldsymbol{\beta}) = (\mathbf{x}'\boldsymbol{\beta}_{O1}, \mathbf{x}'\boldsymbol{\beta}_{O2}, \mathbf{x}'\boldsymbol{\beta}_{O3})'$, where $\boldsymbol{\beta}_{Ok}\ \ k = 1, 2, 3$ are 7×1 coefficient vectors. For the radius component on the m-rep, we set $\mu_r(\mathbf{x}, \boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta}_r)$, where $\boldsymbol{\beta}_r$ is a 7×1 coefficient vector. For the directional components on the m-rep, we used $\boldsymbol{\mu}_0(\mathbf{x}_i, \boldsymbol{\beta})$ as defined in (5.1) and (5.2), where $\boldsymbol{\beta}_{n0} = (\boldsymbol{\beta}_{1d,n0}', \boldsymbol{\beta}_{1c,n0}', \boldsymbol{\beta}_{2d,n0}', \boldsymbol{\beta}_{2c,n0}')'$ for $\mathbf{n}_0$ and $\boldsymbol{\beta}_{n1} = (\boldsymbol{\beta}_{1d,n1}', \boldsymbol{\beta}_{1c,n1}', \boldsymbol{\beta}_{2d,n1}', \boldsymbol{\beta}_{2c,n1}')'$ for $\mathbf{n}_1$. Therefore, we have the coefficient vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_{O1}', \boldsymbol{\beta}_{O2}', \boldsymbol{\beta}_{O3}', \boldsymbol{\beta}_r', \boldsymbol{\beta}_{n0}', \boldsymbol{\beta}_{n1}')'$. Then we minimized the square of the geodesic distance to obtain the estimates of $\boldsymbol{\beta}$ and conducted hypothesis testing using Wald statistics. Since the primary goal of the study is to investigate the difference of m-rep shape between schizophrenia patients and healthy controls, we paid special attention to the terms in $\boldsymbol{\beta}$ associated with diagnostic status.

For the radius component of the m-rep, the color-coded p-values of the diagnostic status effects across the atoms of both the left and right reference hippocampi are shown in Figure 5.2 a and b. The false discovery rate approach was used to correct for multiple comparisons, and the resulting adjusted p-values were shown in Figure 5.2 c and d. Before correcting for multiple comparisons, there was a significant diagnostic status difference in the m-rep thickness at the central atoms near the tail in the left hippocampus, and some area in the right hippocampus. However, there was not much of a significant diagnostic

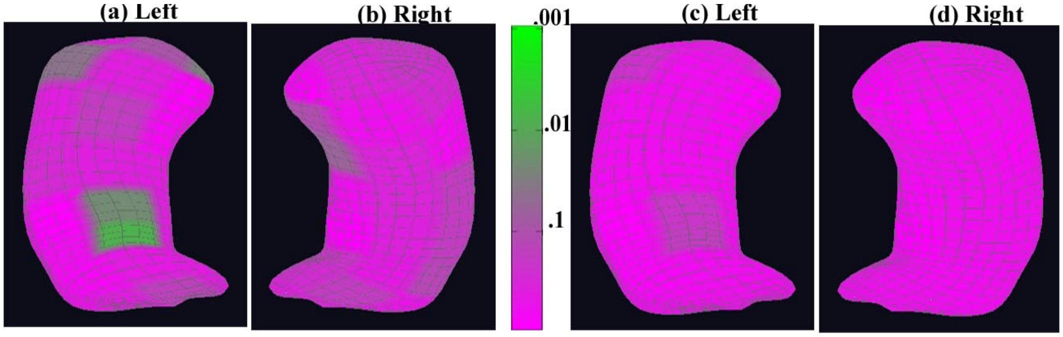

Figure 5.2: Results for the m-rep radius component from the schizophrenia study of hippocampus: the color-coded uncorrected $p-$value maps of the diagnostic status effects for (a) the left hippocampus and (b) the right hippocampus; the color-coded corrected $p-$value maps of the diagnostic status effects for (c) the left hippocampus and (d) the right hippocampus after correcting for multiple comparisons.

status effect after correcting for multiple comparisons.

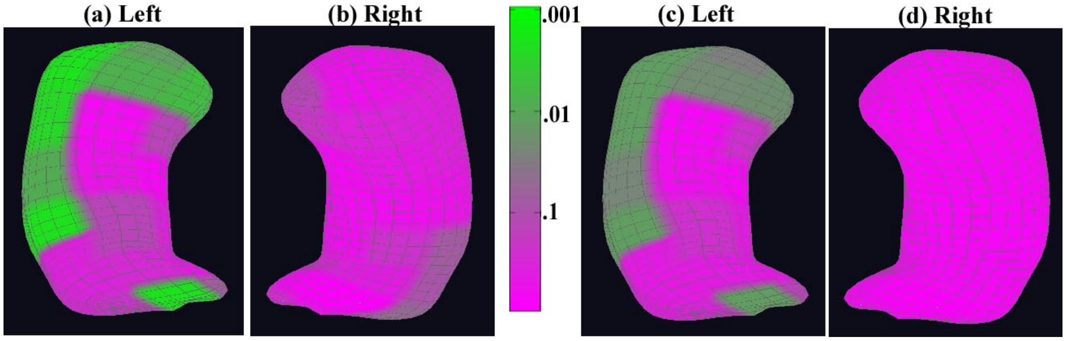

Figure 5.3: Results for the m-rep location component from the schizophrenia study of the hippocampus: the color-coded uncorrected $p-$value maps of the diagnostic status effects for (a) the left hippocampus and (b) the right hippocampus; the color-coded corrected $p-$value maps of the diagnostic status effects for (c) the left hippocampus and (d) the right hippocampus after correcting for multiple comparisons.

For the location component of the m-rep, the color-coded p-values of the diagnostic status effects across the atoms of both the left and right reference hippocampi are shown in Figure 5.3 a and b, and the resulting adjusted p-values are shown in Figure 5.3 c and d. Before correcting for multiple comparisons, there was some significant area around the top and the left side of the left hippocampus, but not much in the right hippocampus. There was still some significance for diagnostic status effect around the same areas in the left hippocampus after correcting for multiple comparisons, but nothing in the right

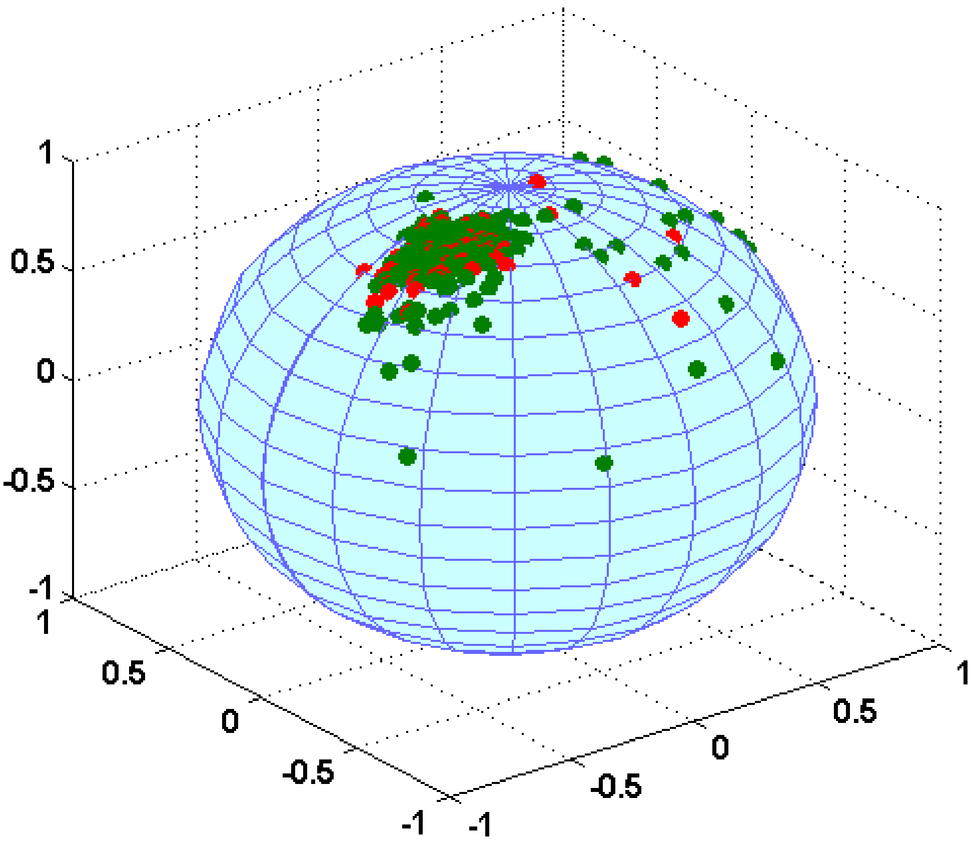

Figure 5.4: A scatter plot of the directional data of n_0 at the first atom in the left hippocampus. The green dots represent the schizophrenia patients and the red dots represents the healthy controls.

hippocampus.

For both the two spoke directions on the m-rep, i.e., $\mathbf{n}_0$ and $\mathbf{n}_1$, we did not see much of a diagnostic status effect at any atoms. This confirms the results of earlier studies in the literature. Figure 5.4 displays the directional data of n_0 at the first atom in the left hippocampus, with the green dots representing schizophrenia patients and the red dots representing healthy controls. Most of the directions are clustered around some area near the north pole, and there is not obvious difference in the distribution pattern between the two diagnostic status groups, the schizophrenia patients and healthy controls.

5.5 Appendix

The following assumptions are needed to facilitate the technical details, although they are not the weakest possible conditions.

ASSUMPTION A1: *The data form an independent and identical sequence.*

ASSUMPTION A2: *The true value $\boldsymbol{\beta}_0$ is the unique minimum point solution, and $\boldsymbol{\beta}_0$ is an interior point of the compact set $\mathcal{B} \subset R^p$.*

ASSUMPTION A3: *In a neighborhood of the true value $\boldsymbol{\beta}_0$, $D_{ni}(\boldsymbol{\beta})$ has a second-order continuous derivative with respect to $\boldsymbol{\beta}$ and $||\partial_{\boldsymbol{\beta}} D_{ni}(\boldsymbol{\beta})||$ and $||\partial^2_{\boldsymbol{\beta}} D_{ni}(\boldsymbol{\beta})||$ are bounded by some integrable function $G(x)$ with $E_F\{G(x)\} < \infty$.*

ASSUMPTION A4: *The rank of $E[\partial_{\boldsymbol{\beta}} D_{ni}(\boldsymbol{\beta})]$ is p and $E[\partial_{\boldsymbol{\beta}} D_{ni}(\boldsymbol{\beta})]^{\otimes 2}$ ($a^{\otimes 2} = aa^{'}$) is positive definite.*

References

Ali, M.M., Khompatraporn, C., and Zabinsky, Z.B. (2005). A numerical evaluation of several stochastic algorithms on selected continuous global optimization test problems. *Journal of Global Optimization*, **31**(4), 635-672.

Andrews, D. W. K. (1994). Empirical process methods in econometrics. *Handbook of Econometrics*, Volume IV. Edited by Engle, R. F. and McFadden, D. L., 2248-2292.

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica*, **67**, 1341-1383.

Ashburner, J. and Friston, K. J. (2000). Voxel-based morphometry: the methods. *NeuroImage*, **11**, 805-821.

Beckman, R. J., Nachtsheim, C. J., and Cook, R. D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, **29**, 413-426.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Ser. B, **57**, 289-300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165-1188.

Bierens, H. (1982). Consistent model specification tests. *Journal of Econometrics*, **20**, 105-134.

Bierens, H. and Ploberger W. (1997). Asymptotic theory of integrated conditional moment tests. *Econometrica*, **65**, 1129-1152.

Bookstein, F. L. (1986). Size and shape spaces for landmark data in two dimensions (with discussion). *Statistical Science*, **1**, 181-242.

Butler, R. W. (2007). *Saddlepoint Approximations with Applications.* New York, Cambridge University Press.

Chen, J., Variyath, A. M., and Abraham, B. (2008). Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics*, **17**, 426-443.

Chen, M. H., Shao, Q. M. and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation.* Springer-Verlag, New York.

Chung, M. K., Robbins, S., Dalton, K. M., Davidson, R. J., Alexander, A. L., and Evans, A. C. (2005). Cortical thickness analysis in autism via heat kernel smoothing. *NeuroImage*, **25**, 1256-1265.

Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society*, Ser. B, **48**, 133-169.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, London.

Cootes, T., Taylor, C., Cooper, D., and Graham, J. (1995). Active shape models - their training and application. *Computer Vision and Image Understanding*, **61**, 38-59.

Copas, J. B. and Eguchi, S. (2001). Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society*, Ser. B, **63**, 871-895.

Copas, J. B. and Eguchi, S. (2005). Local model uncertainty and incomplete-data bias (with discussion). *Journal of the Royal Statistical Society*, Ser. B, **67**, 459-513.

Copas, J. B. and Li, H. G. (1997). Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society*, Ser. B, **59**, 55-96.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society*, Ser. B, **49**, 1-39.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society*, Ser. B, **30**, 248-75.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press, New York.

Davison, A. C. and Tsai, C. L. (1992). Regression model diagnostics. *International Statistical Review*, **60**, 337-55.

Diggle, P., Heagerty, P., Liang, K. Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*, 2nd edn. Oxford University Press, New York.

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory*, **22**, 1030-1051.

Fisher, N. I. and Lee, A. J. (1992). Regression models for an angular response. *Biometrics* **48**, 665-677.

Fletcher P. T., Lu C., Pizer S. M., and Joshi S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape," *Medical Imaging* **23**, 995-1005.

Friston, K. J. (2007). Statistical parametric mapping: the analysis of functional brain images. Academic Press, London.

Goldberg, D.E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison Wesley.

Gould, A. L. (1969). A regression technique for angular variates. *Biometrics*, **25**, 683-700.

Gustafson, P. (2001). On measuring sensitivity to parametric model misspecification. *Journal of the Royal Statistical Society*, Ser. B, **63**, 81-94.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029-1054.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97-109.

Hedar, A. R. and Fukushima, M. (2006). Tabu search directed by direct search methods for nonlinear global optimization. *European Journal of Operational Research* **170**, 329-349.

Hens, N., Aerts, Molenberghs, G., Thijs, H., and Verbeke, G. (2005). Kernel weighted influence measures. *Computational Statistics and Data Analysis*, **48**, 467-487.

Hirsch, M. J., Meneses, C. N., Pardalos, P. M., and Resende, M. G. C. (2007). Global optimization by continuous GRASP. *Optimization Letters*, **1**, 201-212.

Huettel, S. A., Song, A. W., and McCarthy, G. (2004). Functional magnetic resonance imaging. Sinauer Associates, Inc.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, **85**, 765-769.

Ibrahim, J. G., Chen, M. H., and Lipsitz, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics*, **55**, 591-596.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association*, **100**, 332-346.

Ibrahim, J. G. and Lipsitz, S. R. (1996). Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. *Biometrics*, **52**, 1071-1078.

Ibrahim, J. G., Lipsitz, S. R., and Chen, M. H. (1999). Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *Journal of the Royal Statistical Society*, Ser. B, **61**, 173-190.

Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G., and Kenward, M. G. (2006). The nature of sensitivity in monotone missing not at random models. *Computational Statistics and Data Analysis*, **50**, 830-858.

Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., and Van Steen, K. (2003). A local influence approach to binary data from a psychiatric study. *Biometrics*, **59**, 410-419.

Johnson, R. A. and Wehrly, T. E. (1978). Some angular-linear distributions and related regression models. *Journal of the American Statistical Association*, **73**, 602-606.

Jupp, P. E. and Mardia, K. V. (1989). A unified view of the theory of directional statistics, 1975-1988. *International Statistical Review*, **57**, 261-294.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, **220**, 671-680.

Kitamura, Y. (2006). Empirical likelihood methods in econometrics: theory and practice. Cowles Foundation Discussion Paper No. 1569.

Kosorok, M. R. (2003). Bootstraps of sums of independent but not identically distributed stochastic Processes. *Journal of Multivariate Analysis*, **84**, 299-318.

Kosorok, M. R. (2007). *Introduction to Empirical Processes and Semiparametric Inference.* Springer-Verlag, New York.

Lehmann, E. L. and Romano, J. P. (2006). *Testing Statistical Hypotheses*, 3rd edn.

Springer, New York.

Laguna, M. and Martı́, R. (2005). Experimental testing of advanced scatter search designs for global optimization of multimodal functions. *Journal of Global Optimization*, **33**, 235-255.

Lai, T. L., and Small, D. (2007). Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. *Journal of the Royal Statistical Society*, Ser. B, **69**, 79-99.

Lau, J. C., Lerch, J. P. , Sled, J. G., Henkelman, R. M, Evans, A. C., Bedell, B. J. (2008). Longitudinal neuroanatomical changes determined by deformation-based morphometry in a mouse model of Alzheimer's disease. *NeuroImage*, **42**, 19-27.

Liang, F. (2008). Annealing evolutionary stochastoc approximation Monte Carlo for global optimization. *Journal of Global Optimization*, in press.

Liang, F., Liu, C., and Carroll, R. J. (2007). Stochastic approximation in Monte Carlo computation. *Journal of the American Statistical Association*, **102**, 305-320.

Liang, K. Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Lieberman, J. A., Tollefson, G. D., Charles, C., Zipursky, R., Sharma, T., Kahn, R. S., Keefe, R. S. E., Green, A. I., Gur, R. E., McEvoy, J., Perkins, D., Hamer, R. M., Gu, H., and Tohen, M. (2005). Antipsychotic drug effects on brain morphology in first-episode psychosis. *Archives of General Psychiatry*, **62**, 361-70.

Lin, D. Y., Wei, L. J., and Ying, Z. L. (2002). Model-checking techniques based on cumulative residuals. *Biometrics*, **58**, 1-12.

Lipsitz, S. R. and Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika*, **83**, 916-922.

Lipsitz, S. R. and Ibrahim, J. G. (1998). Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics*, **54**, 1002-1013.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*, 2nd edn. John Wiley, New York.

Little, R. J. A. and Schluchter, M. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, **72**, 497-512.

Liu, J. (2003). *Monte Carlo Strategies in Scientific Computing.* Springer, New York.

Luo, W. and Nichols, T. (2003). Diagnosis and exploration of massively univariate fMRI models. *NeuroImage*. **19**, 1014-1032.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn. Chapman and Hall, London.

Mardia, K. V. (1975). Statistics of directional data (with discussion). *Journal of the Royal Statistical Society*, Ser. B, **37**, 349-393.

Mardia, K. V. and Jupp, P. E. (1983). *Directional Statistics.* Academic Press, John

Wiley.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087-1091.

Newey, W. and Smith, R. J. (2004). Higher-order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, **72**, 219-255.

Ossiander, M. (1987). A central limit theorem under metric entropy with bracketing. *The Annals of Probability*, **15**, 897-919.

Owen, A.B. (2001). Empirical Likelihood. Chapman and Hall/CRC, New York.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, **22**, 300-325.

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalized estimating equations using quadratic inference functions. *Biometrika*, **87**, 823-836.

Presnell B., Morrison S. P., and Littell R. C. (1998). Projected multivariate linear models for directional data. *Journal of the American Statistical Association*, **93**, 1068-1077.

Rogers, B. P., Morgan, V.L., Newton, A. T., and Gore, J. C. (2007). Assessing Functional Connectivity in the Human Brain by FMRI. *Magnetic Resonance Imaging*, **25**, 1347-1357.

Salmond, C. H., Ashburner, J., Vargha-Khadem, F., Connelly, A., Gadian, D. G., and Friston, K. J. (2002). Distributional assumptions in voxel-based morphometry," *NeuroImage*, **17**, 1027-1030.

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, **35**, 634-672.

Schluchter, M. and Jackson, K. (1989). Log-linear analysis of censored survival data with partially observed covariates. *Journal of the American Statistical Association*, **84**, 42-52.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics*, **25**, 613-41.

Stute, W., Gonzlez-Manteiga, W., and Presedo-Quindimil, M. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association*, **93**, 141-149.

Stute, W. and Zhu, L. X. (2001). Model checks for generalized linear models. *Scandinavia Journal of Statistics*, **29**, 535-545.

Styner, M., and Gerig, G., (2003). Automatic and robust computation of 3d medial models incorporating object variability," *International Journal of Computer Vision*, **55**, 107-122.

Styner, M., Lieberman, J.A., McClure, R. K., Weinberger, D. R., Jones, D. W., Gerig, G. (2005). Morphometric analysis of lateral ventricles in Schizophrenia and healthy controls regarding genetic and disease-specific factors. *Proceedings of the National Academy of Sciences*, **102**, 4872-4877.

Styner, M., Lieberman, J. A., Pantazis, D. and Gerig, G. (2004). Boundary and medial shape analysis of the hippocampus in schizophrenia. *Medical Image Analysis*, **8**, 197-203.

Styner, M., Oguz, I., Xu, S., Pantazis, D., and Gerig, G. (2007). Statistical group differences in anatomical shape analysis using hotelling T^2 metric. *Proc SPIE, Medical Imaging Conference*, 65123, Z, 1-11.

Su, J. Q. and Wei, L. J. (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association*, **86**, 420-426.

Thomas, W. and Cook, R. D. (1989). Assessing influence on regression coefficients in generalized linear models. *Biometrika*, **76**, 741-749.

Thompson, P. M., Cannon, T. D., and Toga, A. W. (2002). Mapping genetic influences on human brain structure. *The Annals of Medicine*. **24**, 523-536.

Thompson, P. M. and Toga, A. W. (2002). A framework for computational anatomy. *Computing and Visualization in Science*. **5**, 13-34.

Troxel, A. B. (1998). A comparative analysis of quality of life data from a southwest oncology group randomized trial of advanced colorectal cancer. *Statistics in Medicine*, **17**, 767-779.

Troxel, A. B., Ma, G., and Heitjan, D. F. (2004). An index of local sensitivity to nonignorability. *Statistica Sinica*, **14**, 1221-1237.

Tsao, M. (2004). Bounds on coverage probabilities of the empirical likelihood ratio confidence regions. *The Annals of Statistics*. **32**, 1215-1221.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Van Steen, K., Molenberghs, G., and Thijs, H. (2001). A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modelling: An International Journal* **1**, 125-142.

Verbeke, G., Molenberghs, G., Thijs, H., Lasaffre, E., and Kenward, M. G. (2001). Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics*, **57**, 43-50.

Wager, T. D., Keller, M. C., Lacey, S.C., and Jonides, J. (2005). Increased sensitivity in neuroimaging analyses using robust regression. *NeuroImage*, **26**, 99-113.

Worsley, K. J., Taylor, J. E., Tomaiuolo, F., and Lerch, J. (2004). Unified univariate and multivariate random field theory. *NeuroImage*, **23**, 189-195.

Zhu, H. T., Ibrahim, J. G., Lee, S., and Zhang, H. P. (2007a). Assessment of local influence by invariant measures. *The Annals of Statistics*, **35**, 2565-2588.

Zhu, H.T., Ibrahim, J.G., Tang, N.S., Daniel, R., Hao, X., Bansal, R., and Peterson, B. (2007b). A statistical analysis of brain morphology using wild boostrapping. *IEEE*

Trans Med Imaging, **26**, 954-966.

Zhu, H. T. and Lee, S. Y. (2001). Local influence for incomplete-data models. *Journal of the Royal Statistical Society*, Ser. B, **63**, 111-126.

Zhu, H. T., Lee, S. Y., Wei, B. C., and Zhou, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika*, **88**, 727-37.