© 2010
MINGYU LI
ALL RIGHTS RESERVED

NONPARAMETRIC AND SEMIPARAMETRIC REGRESSION, MISSING DATA, AND RELATED ALGORITHMS

BY MINGYU LI

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Department of Statistics and Biostatistics

Written under the direction of

Minge Xie

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2010

ABSTRACT OF THE DISSERTATION

Nonparametric and Semiparametric Regression, Missing Data, and Related Algorithms

by MINGYU LI

Dissertation Director: Minge Xie

This dissertation consists of two chapters:

- Chapter 1 develops nonparametric and semiparametric regression methodologies which relate the group testing responses to the individual covariates information. In this chapter, we extend the parametric regression model of Xie (2001) for binary group testing data to the nonparametric and semiparametric models. We fit nonparametric and semiparametric models and obtain estimators of the parameters by maximizing penalized likelihood function. For implementation, we apply EM algorithm considering the individual responses as complete data and the group testing responses as observed data. Simulation studies are performed to illustrate the methodologies and to evaluate the finite sample performance of our methods. In general, group testing involves a large number of subjects, hence, the computational aspect is also discussed. The results show that our estimation methods perform well for estimating both the individual probability of positive outcome and the prevalence rate in the population.

- Chapter 2 studies a partially linear regression model with missing response variable and develops semiparametric efficient inference for the parametric component

of the model. The missingness considered here includes a broad range of missing patterns. For the estimation method, we use the concept of least favorable curve, least favorable direction and the generalized profile likelihood in Severini and Wong (1992). Asymptotic distributions for the estimators of the parametric components are obtained. It is shown that the estimators are asymptotically normally distributed under some conditions. Furthermore, we prove that the asymptotic covariance of the estimators achieves the semiparametric lower bound under the regularity conditions and additional conditions given in the appendix. We also propose an algorithm which runs iteratively between fitting parametric components and fitting nonparametric components while holding the other fixed. EM algorithms are used in estimating the parametric components by a semiparametric estimating equation and in estimating the nonparametric components by smoothing methods. It is proved that the estimators from this iterative algorithm equal to the conditional expectations (conditioned on observed data) of the semiparametric efficient estimators from complete data. The methodology is illustrated and evaluated by numerical examples.

Acknowledgements

I would like express my sincerest gratitude to my advisor Professor Minge Xie for his insightful and thoughtful guidance and support through these years. Without his continuous encouragement, this dissertation would not be possible.

I would also like to thank all faculties, staff and friends in our department, because of whom my five years at Rutgers has been happy and unforgettable experience. I would especially like to thank my committee members, Professor John Kolassa, Professor Kesar Singh and Professor Tao Huang, for their precious advice and helpful comments.

Finally, I would like to thank my parents, my elder brother, Qiang Li and my husband, Chen Lu, for their support and encouragement.

Dedication

This dissertation is dedicated to my family.

Table of Contents

Chapter 1

Nonparametric and Semiparametric Regression Analysis of Group Testing Samples

1.1 INTRODUCTION

Group testing, or pooled testing, where the samples are tested in pools instead of individually, was first introduced by Dorfman (1943) to reduce cost and increase efficiency of tests. Since then, the group testing method has been widely used in blood or urine tests, chemical compound screening and infectious disease diagnostic tests; see, Cardoso et al. (1998), Kacena et al. (1998a), (1998b), Thorburn et al. (2001), Lindan et al. (2005) and Rours et al. (2005), among others. When the testing method is perfectly accurate, the group testing result is positive if at least one sample in the corresponding pool is positive, and negative if none of the samples are positive in that pool. Therefore, in a study of large population with rare disease, group testing can significantly reduce the total number of tests than individual testing.

Group testing has successful applications whether the objective of the study is to eliminate all positive individuals or to estimate the overall prevalence of the disease in large population. Chen and Swallow (1990) mentioned that group testing can substantially reduce the mean square error of the estimator of prevalence rate and the cost per unit information under some conditions. Vansteelandt, Goetghebeur, and Verstraeten (2000) pointed out that testing pools can lower false positive and false negative rates in low prevalence cases and yield more precise prevalence estimators. Depending on the purpose of study, the group testing schemes and the information contained in the group testing results may vary. If the objective is to identify all positive individuals, all samples in the pools with positive group testing results may be retested. In this case, the individual sample responses are all available. In many other cases, the individual

outcomes can not be implied completely from the group testing results, if the study is designed to estimate only the percentage of positive subjects or for the purpose of protecting privacy. The method developed in this chapter is especially for the latter case.

The optimal designs under various group testing schemes and the efficiency of group testing have also been studied in the literature. Dorfman (1943) calculated the optimal group size which minimizes the total number of tests, given selected prevalence rates in a study designed to weed out all syphilitic men. Yao and Hwang (1990) studied optimal nested group testing algorithms. Other publications include Hughes-Oliver and Swallow (1994), Phatarfod and Sudbury (1994) and Brookmeyer (1999).

In many studies, the individual covariate information, such as age, gender, and general health information, is available and it is of interest to explore whether such information is related to the responses or not. Vansteelandt et al. (2000), Xie (2001) and Chen et al. (2009) have each developed parametric regression methodologies to analyze the relationship between the group testing responses and the covariate variables. Vansteelandt et al. (2000) directly maximized the likelihood function of the group testing responses, while Xie (2001) considered the individual responses and group testing responses as the complete data and observed data respectively and applied the EM algorithm. Chen et al. (2009) studied heterogeneous populations and included a random effect covariate in the regression model. So far, the research on the regression method in group testing has focused on the parametric models, and nonparametric or semiparametric regression models have not been considered in the analysis of group testing samples.

In this chapter, we extend the parametric regression analysis of Xie (2001) to nonparametric and semiparametric regression analyses. We use the penalized maximum likelihood method and the EM algorithm. Penalized likelihood contains the likelihood function and a roughness penalty term and the smoothing parameter controls the trade-off between goodness-of-fit and smoothness. Green and Silverman (1994) provided a thorough discussion on the penalized maximum likelihood method for nonparametric and semi-parametric regression and generalized linear models. Green (1990) applied

the EM algorithm to the penalized maximum likelihood estimator and pointed out that the parameter can represents a smooth function that has been discretized. In our work, we will combine the algorithms in Green and Silverman (1994) and the methodologies in Green (1990) and apply the EM algorithm to the nonparametric and semiparametric regression. The results of numerical examples show that our estimation methods perform well in estimating both the individual probability of positive outcome and sample prevalence rate. In the simulation studies, we consider two pooling strategies, 'alike' and 'random' for comparison, and it turns out that 'alike' pooling provides notable improvement of the estimators, even for the multiple covariates models. However, we need to keep in mind that 'alike' pooling strategy may be impractical to implement in many studies. Bilder and Tebbs (2009) discussed and compared various grouping strategies, including 'alike' and 'random'.

The rest of the chapter is organized as follows. In section 1.2, we present the models, estimation methodology and algorithm; In section 1.3, simulation studies are conducted to illustrate the implementation and to evaluate the performance of the estimation methods for nonparametric and semiparametric models; Section 1.4 summarizes the results.

1.2 ESTIMATION METHOD

1.2.1 Notation and model

In a group testing experiment, samples from N subjects are grouped into, say, n pools and the entire pool is tested first. Then some individuals or the subsets of the n pools will be further tested. We use similar notation as in Xie (2001) in the following.

Let y_i denote whether the sample from the i^{th} individual is positive or not, which is equal to 1 if positive and equal to 0 if negative, for $i = 1, \cdots, N$. Suppose that m tests in total are performed on m (usually $m \geq n$) sets of individuals, say $g_1, g_2, \cdots, g_m$, where the sets correspond to the pools or the subsets of the pools depending on the group testing scheme or the purpose of the study. Denote the m testing results as $\mathbf{t} = \{t_1, \cdots, t_m\}$ corresponding to the sets $G = \{g_1, \cdots, g_m\}$. The testing result t_i is

equal to 1 if positive and 0 otherwise. In general, the testing methods are not perfectly accurate and sensitivity and specificity are used to specify the accuracy of a testing method. Let η and θ denote the sensitivity and specificity respectively, then we have $0 < \eta \leq 1$ and $0 < \theta \leq 1$. Under this assumption, t_i can be decided by

$$t_i = W_i 1_{(\sum_{j \in g_i} y_j > 0)} + (1 - V_i) 1_{(\sum_{j \in g_i} y_j = 0)},$$

where W_i and V_i are independent Bernoulli random variables equal to 1 with probability η and θ respectively and $1(\cdot)$ is the indicator function.

When covariate variables of individual subjects are available, we can fit generalized linear regression models for the individual responses. If we assume that the covariates are linearly related to the link function $h(\cdot)$, we can construct parametric GLM:

$$h\big[P(y_i = 1)\big] = x_i^T \beta; \tag{1.1}$$

and if we assume that a covariate may be related to the link function by an unknown smooth function, we can use nonparametric GLM:

$$h\big[P(y_i = 1)\big] = f(v_i); \tag{1.2}$$

or semiparametric GLM can be fitted if we think that some covariates are linearly related and some are related to the link function by an unknown smooth function,

$$h\big[P(y_i = 1)\big] = x_i^T \beta + f(v_i), \tag{1.3}$$

where x_i is $p \times 1$ covariate vector and v_i is covariate variable. The link function $h(\cdot)$ is a known monotonic function, which is differentiable. We are interested in estimating unknown smooth function $f(\cdot)$ in model (1.2) and β and $f(\cdot)$ in model (1.3). The most commonly used link function for binary responses is $h(p) = logit(p) = \log\big(p/(1-p)\big)$. Model (1.1) has been discussed in Xie (2001) and we will develop estimation methods for models (1.2) and (1.3) in this chapter.

Sometimes, it is possible to identify all the individual testing results $\mathbf{y} = (y_1, \cdots, y_N)$ from the testing results $\mathbf{t} = \{t_1, \cdots, t_m\}$. However, in many other cases, the individual testing results can not be fully determined from the group testing results, and our

focus is on the latter case. The earlier case can be regarded as a special case of the latter one. For some testing schemes, the explicit formula for the likelihood function of $(t_1, \cdots, t_m)$ (observed likelihood) may not be available or may be very complicated. Thus direct maximization of the observed likelihood function could be a tedious task, if not impossible. On the other hand, the log-likelihood function of complete data has simple form for generalized linear models. Therefore, EM algorithms for nonparametric and semiparametric GLMs are developed in the following.

1.2.2 Estimation method for nonparametric model

Under the model (1.2), the log-likelihood function of $(y_1, \cdots, y_N)$ is very simple:

$$l(f_1, \cdots, f_N | y_1, \cdots, y_N) = \sum_{i=1}^{N} \big[y_i f_i - \log(1 + e^{f_i}) \big],$$

where $f_i = f(v_i)$. We consider $\mathbf{y} = (y_1, \cdots, y_N)$ as the complete data, which is not completely observed and $\mathbf{t} = (t_1, \cdots, t_m)$ as the observed data.

We want to maximize the penalized observed log-likelihood function,

$$l(f_1, \cdots, f_N | t_1, \cdots, t_m) - \alpha/2 \int f''(v)^2 dv, \tag{1.4}$$

to obtain the estimator of $f(\cdot)$, $\hat{f}(\cdot)$. Here, $l(f_1, \cdots, f_N | t_1, \cdots, t_m)$ is the log-likelihood function of $\mathbf{t} = (t_1, \cdots, t_m)$ and α is the smoothing parameter. By taking conditional expectation, the first term of (1.4) can be written as

$$\begin{aligned} l(f_1, \cdots, f_N | t_1, \cdots, t_m) &= \log P(t_1, \cdots, t_m | f_1, \cdots, f_N) \\ &= \log P(\mathbf{y} | f_1, \cdots, f_N) - \log P(\mathbf{y} | \mathbf{t}, f_1, \cdots, f_N) \\ &= E\big[\log P(\mathbf{y} | f_1, \cdots, f_N) | \mathbf{t}, \tilde{f}_1, \cdots, \tilde{f_N} \big] \\ &\quad - E\big[\log P(\mathbf{y} | \mathbf{t}, f_1, \cdots, f_N) | \mathbf{t}, \tilde{f}_1, \cdots, \tilde{f_N} \big], \end{aligned} \tag{1.5}$$

where $P(\cdot)$ is the probability density function, and $\tilde{\mathbf{f}} = (\tilde{f}_1, \cdots, \tilde{f_N})$ are estimators of $(f_1, \cdots, f_N)$ in the previous iteration. By the information inequality, maximizing the first term of (1.5) for $\mathbf{f} = (f_1, \cdots, f_N)$ increases the value of the first term of (1.4). Hence, instead of maximizing the penalized observed log-likelihood (1.4) directly,

we maximize the conditional expectation of penalized complete log-likelihood given observed data, (1.6) iteratively until convergence.

$$\begin{aligned} & E\big[\log P(\mathbf{y}|f_1,\cdots,f_N)|\mathbf{t},\tilde{f}_1,\cdots,\tilde{f_N}\big]-\alpha/2\int f''(v)^2dv \\ = & \sum_{i=1}^{N}\big[E(y_i|\mathbf{t},\tilde{\mathbf{f}})f_i-\log(1+e^{f_i})\big]-\alpha/2\int f''(v)^2dv. \end{aligned} \tag{1.6}$$

Then we have the following EM algorithm to obtain the estimator, $\hat{f}(\cdot)$.

- Step 1. Select starting points $f^{[0]}(v_i)$ of $f(v_i)$ for $i=1,2,\cdots,N$.
- Step 2. (E-step) For given $f^{[k]}(v_i)$ for $i=1,\cdots,N$ at the $[k]^{th}$ iteration, calculate the conditional expectations

$$c_i^{[k]}=E\big[y_i|t_1,\cdots,t_m,f^{[k]}(v_1),\cdots,f^{[k]}(v_N)\big],\quad i=1,\cdots,N.$$

- Step 3. (M-step) Given $(c_1^{[k]},\cdots,c_N^{[k]})$ for fixed $k=0,1,2,\cdots$, update the estimator at the $[k+1]^{th}$ iteration, $f^{[k+1]}(v_i)$, for $i=1,\cdots,N$, by maximizing the following penalized log-likelihood function:

$$\sum_{i=1}^{N}\big[c_i^{[k]}f_i-\log(1+e^{f_i})\big]-\alpha/2\int f''(v)^2dv. \tag{1.7}$$

- Step 4. Repeat step 2 and 3 until $\big\|f^{[k+1]}-f^{[k]}\big\|$ is very small, that is, until the algorithm converges numerically.

The maximization of (1.7) in Step 3 will be discussed in the following subsection.

1.2.3 Computational consideration

For a fixed α, maximizing (1.7) can be solved via iterating on the penalized weighted least squares problem (refer to Gu (1992))

$$\min\sum_{i=1}^{N}\big[b_i''(z_i-f_i)^2\big]+\alpha\int f''(v)^2dv, \tag{1.8}$$

where $b_i''=e^{\tilde{f}_i}/(e^{\tilde{f}_i}+1)^2$, $z_i=\tilde{f}_i+(c_i^{[k]}-b_i')/b_i''$, $b_i'=e^{\tilde{f}_i}/(e^{\tilde{f}_i}+1)$ and $\tilde{f}_i=\tilde{f}(v_i)$ is evaluation of the $f(v_i)$ in last iteration.

By Green and Silverman (1994), the solution of problem (1.8) is natural cubic spline and the penalty term can be written as

$$\alpha \int f''(v)^2 dv = \alpha \mathbf{f}^T K \mathbf{f},$$

for natural cubic spline, where $K = QR^{-1}Q^T$ and $\mathbf{f} = (f(v_{(1)}), \cdots, f(v_{(N)}))$. Here Q is a $n \times (n-2)$ band matrix and R is a $(n-2) \times (n-2)$ symmetric band matrix and each element of these two matrices is a function of $(v_{(1)}, \cdots, v_{(N)})$, which is the ordered values of $(v_1, \cdots, v_N)$. The matrices, Q and R are given in Appendix 1.5.1. All the notations b_i'', b_i' and z_i are based on the ordered values of $(v_1, \cdots, v_N)$, $(v_{(1)}, \cdots, v_{(N)})$ afterwards. For instance, $b_i'' = e^{\tilde{f}(v_{(i)})}/(e^{\tilde{f}(v_{(i)})} + 1)^2$.

Let W is a diagonal matrix with $W_{ii} = b_i''$ and working response vector $\mathbf{z} = (z_1, \cdots, z_N)$, then the matrix form of problem (1.8) is

$$\min S(\mathbf{f}) = (\mathbf{z} - \mathbf{f})^T W (\mathbf{z} - \mathbf{f}) + \alpha \mathbf{f}^T K \mathbf{f}, \tag{1.9}$$

and the solution of (1.9) is

$$\mathbf{f}^{new} = (W + \alpha K)^{-1} W \mathbf{z}. \tag{1.10}$$

In group testing, the sample size N is usually very large, hence direct use of (1.10) is not appropriate and is too time consuming for general use. So we can apply the Reinsch algorithm for weighted smoothing (refer to Green and Silverman (1994)) to calculate (1.10). The steps of the algorithm are given in Appendix 1.5.2 and each step can be performed in $O(N)$ algebraic operations.

1.2.4 Choosing the smoothing parameter

Generalized cross-validation (GCV) is a common method for choosing the smoothing parameter, so we apply the following GCV criteria to choose α.

$$\min GCV(\alpha) = \frac{\left\| W^{\frac{1}{2}}(\mathbf{z} - \mathbf{f}) \right\|^2}{n\left\{1 - \frac{1}{n} tr\left[(W + \alpha K)^{-1} W^{\frac{1}{2}}\right]\right\}^2} = \frac{n\left\| W^{\frac{1}{2}}(\mathbf{z} - \mathbf{f}) \right\|^2}{\left[tr(\alpha W^{-\frac{1}{2}} K W^{-\frac{1}{2}})\right]^2}, \tag{1.11}$$

where W, $\mathbf{z}$ and $\mathbf{f}$ are all evaluated at the converged estimator, $\hat{f}(\cdot)$.

Other criteria, like cross validation and likelihood based cross validation, can also be used.

1.2.5 Estimation method for the semiparametric model

Semiparametric model can be analyzed by using similar estimation method and algorithm used in nonparametric model. Under the semiparametric model (1.3), the log-likelihood function of $(y_1, \cdots, y_N)$ is very simple:

$$l(\beta, f_1, \cdots, f_N | y_1, \cdots, y_N) = \sum_{i=1}^{N} \big[y_i (x_i^T \beta + f_i) - \log(1 + e^{x_i^T \beta + f_i}) \big],$$

where $f_i = f(v_i)$.

We want to maximize the penalized observed log-likelihood function,

$$l(\beta, f_1, \cdots, f_N | t_1, \cdots, t_m) - \alpha/2 \int f''(v)^2 dv, \tag{1.12}$$

to obtain the estimators of β and $f(\cdot)$, $\hat{\beta}$ and $\hat{f}(\cdot)$. Here, $l(\beta, f_1, \cdots, f_N | t_1, \cdots, t_m)$ is the log-likelihood function of $(t_1, \cdots, t_m)$, and α is the smoothing parameter. The first term of (1.12) can be written as

$$\begin{aligned}
& l(\beta, f_1, \cdots, f_N | t_1, \cdots, t_m) = \log P(t_1, \cdots, t_m | \beta, f_1, \cdots, f_N) \\
& = \log P(\mathbf{y} | \beta, f_1, \cdots, f_N) - \log P(\mathbf{y} | \mathbf{t}, \beta, f_1, \cdots, f_N) \\
& = E\big[\log P(\mathbf{y} | \beta, f_1, \cdots, f_N) | \mathbf{t}, \tilde{\beta}, \tilde{f}_1, \cdots, \tilde{f_N} \big] \\
& \quad - E\big[\log P(\mathbf{y} | \mathbf{t}, \beta, f_1, \cdots, f_N) | \mathbf{t}, \tilde{\beta}, \tilde{f}_1, \cdots, \tilde{f_N} \big],
\end{aligned} \tag{1.13}$$

where $\tilde{\beta}$ and $(\tilde{f}_1, \cdots, \tilde{f_N})$ are estimators of β and $(f_1, \cdots, f_N)$ in the previous iteration respectively. By the information inequality, maximizing the first term of (1.13) for β and $(f_1, \cdots, f_N)$ increases the value of the first term of (1.12). Hence, instead of maximizing the penalized observed log-likelihood (1.12), we maximize the conditional expectation of the penalized complete log-likelihood given observed data, (1.14) iteratively until converge.

$$\begin{aligned}
& E\big[\log P(\mathbf{y} | \beta, f_1, \cdots, f_N) | \mathbf{t}, \tilde{\beta}, \tilde{f}_1, \cdots, \tilde{f_N} \big] - \alpha/2 \int f''(v)^2 dv \\
= & \sum_{i=1}^{N} \big[E(y_i | \mathbf{t}, \tilde{\beta}, \tilde{\mathbf{f}}) (x_i^T \beta + f_i) - \log(1 + e^{x_i^T \beta + f_i}) \big] - \alpha/2 \int f''(v)^2 dv \\
\equiv & \sum_{i=1}^{N} \big[c_i (x_i^T \beta + f_i) - \log(1 + e^{x_i^T \beta + f_i}) \big] - \alpha/2 \int f''(v)^2 dv,
\end{aligned} \tag{1.14}$$

where $c_i = E(y_i|\mathbf{t}, \tilde{\beta}, \tilde{\mathbf{f}})$ for $i = 1, \cdots, N$. Similar to the nonparametric GLM, the integration part in the second term of (1.14) can be written by $\mathbf{f}^T K \mathbf{f}$ for natural cubic splines. By Theorem 5.2 of Green and Silverman (1994), the Fisher scoring algorithm for maximizing the penalized log-likelihood (1.14) with respect β and $f(\cdot)$ for fixed α is given by solving

$$\begin{bmatrix} X^T W X & X^T W \\ W X & W + \alpha K \end{bmatrix} \begin{pmatrix} \beta \\ \mathbf{f} \end{pmatrix} = \begin{pmatrix} X^T W \mathbf{z} \\ W \mathbf{z} \end{pmatrix}, \tag{1.15}$$

where the working response vector $\mathbf{z} = (z_1, \cdots, z_N)$ has the form

$$z_i = x_i^T \tilde{\beta} + \tilde{f}_i + (c_i - b_i')/b_i'',$$

and $b_i'' = e^{x_i^T \tilde{\beta} + \tilde{f}_i}/(e^{x_i^T \tilde{\beta} + \tilde{f}_i} + 1)^2$, $b_i' = e^{x_i^T \tilde{\beta} + \tilde{f}_i}/(e^{x_i^T \tilde{\beta} + \tilde{f}_i} + 1)$, $X = (x_1, \cdots, x_N)^T$ and W is a diagonal matrix with $W_{ii} = b_i''$. Here $\tilde{\beta}$ and $\tilde{f}_i = \tilde{f}(v_i)$ are evaluations of β and $f(v_i)$ in the last iteration. The b_i'', b_i', z_i and $\tilde{f}_i$ are all based on the ordered values of $(v_1, \cdots, v_N)$ afterwards.

Equation (1.15) forms a system of $p + n$ equations, and it may not be convenient to solve this system directly. However, (1.15) can be written as a pair of simultaneous matrix equations (refer to Green and Silverman (1994)),

$$\begin{aligned} X^T W X \beta &= X^T W (\mathbf{z} - \mathbf{f}), \\ (W + \alpha K)\mathbf{f} &= W(\mathbf{z} - X\beta). \end{aligned}$$

Therefore, the semiparametric GLM can be fitted by the following algorithm which runs iteratively between fitting parametric components and fitting nonparametric components while holding the other fixed. This method is also known as back-fitting:

- Step 1. Select starting points $\beta^{[0]}$ and $f_i^{[0]}$ for $i = 1, 2, \cdots, N$.
- Step 2. (E-step for parametric part) For given $\beta^{[k]}$ and $f_i^{[k]}$ for $i = 1, \cdots, N$, update

$$c_i^{[k]} = E(y_i|t_1, \cdots, t_m, \beta^{[k]}, f_1^{[k]}, \cdots, f_N^{[k]}), \quad i = 1, \cdots, N.$$

- Step 3. (M-step for parametric part) Given $(c_1^{[k]}, \cdots, c_N^{[k]})$ for fixed $k = 0, 1, 2, \cdots$, update the estimator at the $[k+1]^{th}$ iteration, $\beta^{[k+1]}$, by

$$\beta^{[k+1]} = [X^T W X]^{-1} X^T W (\mathbf{z} - \mathbf{f}^{[k]}).$$

- Step 4. (E-step for nonparametric part) For given $\beta^{[k+1]}$ and $f_i^{[k]}$ for $i = 1, \cdots, N$ at the $[k]^{th}$ iteration, calculate

$$c_i^{[k]} = E(y_i | t_1, \cdots, t_m, \beta^{[k+1]}, f_1^{[k]}, \cdots, f_N^{[k]}), \quad i = 1, \cdots, N.$$

- Step 5. (M-step for nonparametric part) Given $(c_1^{[k]}, \cdots, c_N^{[k]})$ for fixed $k = 0, 1, 2, \cdots$, update the estimator at the $[k+1]^{th}$ iteration, $f_i^{[k+1]}$, for $i = 1, \cdots, N$, by

$$\mathbf{f}^{[k+1]} = (W + \alpha K)^{-1} W (\mathbf{z} - X\beta^{[k+1]}).$$

- Step 6. Repeat Step 2 to Step 5 until both the $\left\|\beta^{[k+1]} - \beta^{[k]}\right\|$ and $\left\|f^{[k+1]} - f^{[k]}\right\|$ are very small; that is, until the algorithm converges numerically.

The Reinsch algorithm for weighted smoothing can be applied in Step 3, and the GCV criteria can be used to choose the smoothing parameter.

1.3 SIMULATION STUDIES

In this section we conduct simulation studies to evaluate the finite sample performance of the penalized maximum likelihood estimation methodology proposed in Section 1.2. We apply the Gastwirth-Hammick (GH) group testing scheme proposed by Gastwirth and Hammick (1989) for illustration.

Under the GH group testing scheme, individual samples to be tested are batched into pools first. Then a screening test is performed for each pool. After that, those pools classified as positive are given confirmatory tests. In general, the screening test is cheap but not quite accurate while the confirmatory test is almost perfect with higher cost. Gastwirth and Hammick (1989) noted that in blood testing practice for screening

HIV positives, the commonly used screening test is the ELISA kit and the standard confirmatory test is the Western blot (WB) analysis.

Without loss of generality, we assume that total $N = nk$ individual samples are grouped into n pools of size k. The same group size is used in the simulation studies for simplicity. However, the proposed algorithm can be applied to different group sizes in the same way. Denote the screening testing results by $t_1^{(s)}, \cdots, t_n^{(s)}$ ($t_i^{(s)}$ is equal to 1 if positive; 0 otherwise) corresponding to pools $g_1, \cdots, g_n$ respectively. Suppose there are r positive outcomes $t_{j1}^{(s)}, \cdots, t_{jr}^{(s)}$ of n screening tests, and they correspond to the pools $g_{j1}, \cdots, g_{jr}$. Let $t_{j1}^{(c)}, \cdots, t_{jr}^{(c)}$ denote the r confirmatory testing results. Therefore, we have testing results $\mathbf{t} = \{t_1^{(s)}, \cdots, t_n^{(s)}, t_{j1}^{(c)}, \cdots, t_{jr}^{(c)}\}$ from pools $G = \{g_1, \cdots, g_n, g_{j1}, \cdots, g_{jr}\}$ and the total number of tests is $m = n + r$.

For the screening tests, the testing results can be written as

$$t_j^{(s)} = W_j^{(s)} 1_{(\sum_{i \in g_j} y_i > 0)} + (1 - V_j^{(s)}) 1_{(\sum_{i \in g_j} y_i = 0)}, \tag{1.16}$$

where $W_j^{(s)}$ and $V_j^{(s)}$ are independent Bernoulli random variables equal to 1 with probability $\eta^{(s)}$ and $\theta^{(s)}$ respectively; for the confirmatory tests, the testing results can be expressed as

$$t_{jl}^{(c)} = W_l^{(c)} 1_{(\sum_{i \in g_{jl}} y_i > 0)} + (1 - V_l^{(c)}) 1_{(\sum_{i \in g_{jl}} y_i = 0)}, \tag{1.17}$$

where $W_l^{(c)}$ and $V_l^{(c)}$ are independent Bernoulli random variables equal to 1 with probability $\eta^{(c)}$ and $\theta^{(c)}$ respectively. In fact, $(\eta^{(s)}, \theta^{(s)})$ are sensitivity and specificity of screening tests and $(\eta^{(c)}, \theta^{(c)})$ are sensitivity and specificity of confirmatory tests.

We carry out two simulation studies, one is for nonparametric model, and the other is for semiparametric model. The simulation study for the semiparametric model is based on the chlamydia data collected by the state of Nebraska as part of the Infertility Prevention Project.

1.3.1 Nonparametric model

Suppose that individual i belongs to the group g_j, then by Bayes formula, it is easy to verify that the conditional expectation of y_i given $(t_1, \cdots, t_m)$ and $(f_1, \cdots, f_N)$ has the

following explicit formula,

$$
\begin{aligned}
& E(y_i|f_1,\cdots,f_N,t_1,\cdots,t_m) \\
= \; & \frac{(1-\eta^{(s)})p_i}{(1-\eta^{(s)})\big[1-\prod_{i'\in g_j}(1-p_{i'})\big]+\theta^{(s)}\big[\prod_{i'\in g_j}(1-p_{i'})\big]}1_{(t_j^{(s)}=0)} \\
& +\frac{\eta^{(s)}(1-\eta^{(c)})p_i}{\eta^{(s)}(1-\eta^{(c)})\big[1-\prod_{i'\in g_j}(1-p_{i'})\big]+(1-\theta^{(s)})\theta^{(c)}\big[\prod_{i'\in g_j}(1-p_{i'})\big]}1_{(t_j^{(s)}=1,t_j^{(c)}=0)} \\
& +\frac{\eta^{(s)}\eta^{(c)}p_i}{\eta^{(s)}\eta^{(c)}\big[1-\prod_{i'\in g_j}(1-p_{i'})\big]+(1-\theta^{(s)})(1-\theta^{(c)})\big[\prod_{i'\in g_j}(1-p_{i'})\big]}1_{(t_j^{(s)}=t_j^{(c)}=1)},
\end{aligned}
$$

where $p_i = \exp(f_i)/\big(1+\exp(f_i)\big)$ for $i = 1,\cdots,N$.

The assumption of $\eta^{(c)} = \theta^{(c)} = 1$, which was used in Gastwirth and Hammick (1989), that is the confirmatory test is perfectly accurate, is also adopted in the simulation studies. In this case,

$$
\begin{aligned}
& E(y_i|f_1,\cdots,f_N,t_1,\cdots,t_m) \\
& = \frac{(1-\eta^{(s)})p_i}{(1-\eta^{(s)})\big[1-\prod_{i'\in g_j}(1-p_{i'})\big]+\theta^{(s)}\big[\prod_{i'\in g_j}(1-p_{i'})\big]}1_{(t_j^{(s)}=0)} \\
& +\frac{p_i}{1-\prod_{i'\in g_j}(1-p_{i'})}1_{(t_j^{(s)}=t_j^{(c)}=1)}. \qquad (1.18)
\end{aligned}
$$

For each replication, we generate independent random samples of size N of (v_i, y_i) for $i = 1,\cdots,N$, where v_i is from uniform distribution $U(-6.28, 6.28)$ and y_i is Bernoulli random variable, which is equal to 1 with probability p_i. We take $logit(p_i) = f(v_i) = a + b\sin(v_i/2)$, where $a = -2.65$ and $b = 0.6$. Under this sin curve setting, the mean probability that y equal to 1 is about 7.08% (range from 3.73% to 11.41%). After that the N individuals are pooled into $n = N/5$ groups $(g_1,\cdots,g_n)$ of size 5. For simplicity, we use the same size for all the pools and take pool size 5 for illustration. After grouping, the results of screening tests are generated according to (1.16) with $\eta^{(s)} = 0.923$ and $\theta^{(s)} = 0.996$, the same sensitivity and specificity used in the simulation study of Xie (2001). Furthermore, the results of confirmatory tests are generated by (1.17) assuming $\eta^{(c)} = \theta^{(c)} = 1$ for the pools with positive screening test results. Our purpose is to estimate $f(\cdot)$ given $(v_1,\cdots,v_N)$ and $\mathbf{t} = (t_1,\cdots,t_n,t_{j1},\cdots,t_{jr})$ and then estimate the overall prevalence based on the estimators of $f(\cdot)$. Under this group testing scheme, only 21.4% of tests are needed compared to the individual tests.

We generate 200 replications with sample size $N = 5000$ and 10000. For each replication, we estimate $\hat{f}(\cdot)$. Since we know the true $f(\cdot)$, we can choose the smoothing parameter $\alpha = \alpha_{MISE}$ by minimizing the mean integrated squared error (MISE) of the estimators $\hat{f}(\cdot)$. We also use generalized cross validation criteria (1.11) to choose $\alpha = \alpha_{GCV}$ and compare the estimators using α_{GCV} to those using α_{MISE} and true $f(\cdot)$ values. The optimum smoothing parameter α is searched on the grid 0.1(0.05)1.

Bilder and Tebbs (2009) compared 3 different pooling strategies — 'alike', 'random' and 'different'. In this chapter we consider 'alike' and 'random' grouping methods. In 'alike' pooling strategy, samples with similar covariates are grouped together. This can be done by ordering the covariate first and then forming groups, when there is only one covariate. For multiple covariates models, Vansteelandt et al. (2000) suggested that one can sort by 'the most important' covariate first, and then sort the second most important covariate within sorted values of the first one. This approach continues until all covariates have been sorted. In 'random' pooling strategy, samples are randomly assigned to pools, regardless of their covariate values.

The simulation results are summarized in Table 1.1. Table 1.1 shows the integrated relative bias, the integrated standard error, the integrated MISE and the estimator of prevalence rate. In the table, $\alpha = \alpha_{GCV}$ ('random') means the 'random' pooling strategy is used and the smoothing parameter α is selected by minimizing GCV score. Similarly, $\alpha = \alpha_{MISE}$ ('alike') means the 'alike' pooling strategy is used and the smoothing parameter α is selected by minimizing MISE value and so on.

Insert Table 1.1 here.

From Table 1.1, we can see that 'alike' method provides better estimators than 'random' method for both $\alpha = \alpha_{MISE}$ and $\alpha = \alpha_{GCV}$, which is intuitive. In addition, using $\alpha = \alpha_{MISE}$ gives a little better estimator than using $\alpha = \alpha_{GCV}$ for both pooling strategies, which is also expected. When sample size N is equal to 10000, the relative bias and empirical MISE are reduced about half compared to those when N is equal to 5000 for both $\alpha = \alpha_{MISE}$ and $\alpha = \alpha_{GCV}$. Compared to the true prevalence 7.08%, the estimators of the prevalence are very close to true value for all cases.

The point-wise average of the estimated nonparametric curves $\hat{f}(\cdot)$ over 200 replications are displayed in Figure 1.1. The left panel of Figure 1.1 shows the estimators using $N = 5000$ and the right panel is for the estimators using $N = 10000$.

Insert Figure 1.1 here.

In Figure 1.1, the blue dotted curve represents the true values; the red solid curve is for the estimator using α_{GCV} and 'random' pooling, while the red dashed curve represents the estimator using smoothing parameter α_{MISE} and 'random' pooling; the green solid curve is for the estimator using α_{GCV} and 'alike' pooling, while the green dashed curve represents the estimator using smoothing parameter α_{MISE} and 'alike' pooling. First of all, the point-wise average curves of the estimators from 4 methods are all close to the true curve. Second, the estimator using α_{MISE} is closer to the true curve than the one using α_{GCV} given the same pooling strategy, which is expected, however, the difference becomes smaller as the sample size N increases. In addition, the 'alike' pooling method has notable improvement compared to the 'random' pooling method. When sample size increases, the difference from pooling strategies and smoothing parameter selection criteria becomes smaller and all the estimators are very close to the true curve in the whole support of the covariate.

Figure 1.2 illustrates the point-wise variances of the estimators of $\hat{f}(\cdot)$ over 200 replications, with the left panel for $N = 5000$ and right panel for $N = 10000$.

Insert Figure 1.2 here.

In Figure 1.2, the red solid curve is for the estimator using α_{GCV} and 'random' pooling, while the red dashed curve represents the estimator using smoothing parameter α_{MISE} and 'random' pooling; the green solid curve is for the estimator using α_{GCV} and 'alike' pooling, while the green dashed curve represents the estimator using smoothing parameter α_{MISE} and 'alike' pooling. All the variance curves have similar trend. They have larger variances in the margin of the support of v and when the corresponding probability of positive, p_i is low. Furthermore, the variances decrease dramatically when the sample size increases, and 'alike' method has smaller point-wise variances than 'random' method.

In conclusion, the simulation studies demonstrate that our proposed estimation algorithm for nonparametric model performs well and generalized cross validation criteria chooses proper smoothing parameters in these settings.

1.3.2 Semiparametric model

In this section we conduct a simulation study based on the chlamydia data example studied in Chen et al. (2009). Chen et al. (2009) developed regression method to fit mixed effect models for group testing samples, and applied their method to the chlamydia data collected by the state of Nebraska. The data set consists of chlamydia infection statuses for 6138 subjects, and the risk covariates like age, gender, urethritis status and infection symptoms status. The sample prevalence is 7.8 percent.

In our example, we consider two covariates, age and some continuous covariate V, and assume that age is linearly related to the link function, while V has nonparametric relationship with the link function. We fit the semiparametric GLM:

$$logit\big\{P(y_i = 1)\big\} = \beta * age_i + f(v_i), \tag{1.19}$$

and estimate β and $f(\cdot)$.

Under the assumption that $\eta^{(c)} = \theta^{(c)} = 1$, $E(y_i|\beta, f_1, \cdots, f_N, t_1, \cdots, t_m)$ has the same formula as (1.18), where $p_i = \exp(x_i^T\beta + f_i)/\big(1 + \exp(x_i^T\beta + f_i)\big)$ for $i = 1, \cdots, N$. For simplicity, we take the total number of subjects N equal to 6140 and group the samples into 1228 pools with group size 5. Again, the smoothing parameter α is selected by minimizing GCV, and is searched on the grid 0.1(0.05)1. We use both 'alike' and 'random' pooling strategies. For the 'alike' grouping, there are two approaches: 'alike' by par-non and 'alike' by non, depending on sorting by which covariate first. The 'alike' by par-non means that we sort by the age, and then sort by V in the same value of Age; while 'alike' by non means that we sort the samples by V (assume that there are no ties in V). For model (1.19), the covariate Age is generated randomly from $\{15 : 45\}$, V is a continuous random variable from uniform distribution $U(1.57, 7.85)$ and $f(v) = -1.25 + \sin(v)$. Assume that true β is equal to -0.05 and $\eta^{(s)} = 0.95$ and $\theta^{(s)} = 0.98$. Under these settings, the overall positive percentage is about 7.8 percent

and only 21.6% tests are needed compared to the individual testing. For the model (1.19), we estimate β and $f(\cdot)$ by the EM algorithm proposed in Section 1.2.5.

In this example, we generate 200 replications. Table 1.2 shows the average and standard error of 200 estimators $\hat{\beta}$, and the integrated relative bias, the integrated S.E. and the integrated MISE of $\hat{f}(\cdot)$. The estimator of prevalence rate is also calculated.

Insert Table 1.2 here.

Table 1.2 shows that for the parametric part β, all of the three pooling strategies —'random', 'alike' by par-non and 'alike' by non, provide good estimators, which are very close to the true value -0.05 with small standard errors. Among them, 'random' and 'alike' by non have similar S.Es and 'alike' by par-non has smallest S.E. For the nonparametric part, 'alike' by non has smallest relative bias and empirical MISE and 'alike' by par-non has the smallest empirical SE. In addition, all the three pooling methods provide the estimators of prevalence rate very close to the true value, 7.8%.

The box-plot of the estimators of β is displayed in Figure 1.3 for 'random', 'alike' by par-non and 'alike' by non.

Insert Figure 1.3 here.

This plot shows clearly that averages of the estimators of β are all very close to the true value -0.05 (the dotted line) for 3 pooling strategies.The standard error of 'alike' by par-non is the smallest and 'random' method has similar standard error with the 'alike' by non approach.

The point-wise average (left panel) and point-wise variance (right panel) of the estimators $\hat{f}(\cdot)$ over 200 replications are displayed in Figure 1.4.

Insert Figure 1.4 here.

In Figure 1.4, the blue dotted curve represents the true values; the red solid curve is for 'random'; and the green solid curve is for 'alike' by par-non, while the green dashed curve is for 'alike' by non pooling strategy.

From left panel of 1.4, we can notice that the point-wise average curves from two 'alike' methods are a little closer to the true curve than 'random' method. For the point-wise variance curve, the 'alike' by par-non method gives smallest variances in the whole support of v, and the 'random' and 'alike' by non methods have similar variance curves.

In conclusion, our estimation methodology gives good estimators for both the parametric component and nonparametric component and prevalence rate in semiparametric model.

1.4 DISCUSSION

In this chapter, we generalized the parametric model in Xie (2001) and fitted nonparametric and semiparametric models for group testing responses using the covariate information. We maximize the penalized likelihood function of group testing results and apply the EM algorithm, considering the group testing as the missing data case. By the information inequality, the EM algorithm can be used in both nonparametric and semiparametric models.

For the group testing experiment, since the number of subjects is usually very large, direct use of available software may not be practical. Therefore, the computational aspect has been discussed, and the method of choosing the smoothing parameter has also been considered.

The simulation studies confirm that our proposed estimation methodologies perform very well for both nonparametric and semiparametric models for group testing samples. In simulation studies, we use 'random' and 'alike' pooling strategies, and the results show that 'alike' method improves the estimators significantly, which agrees with the results from other research paper.

1.5 APPENDICES

1.5.1 Q and R matrices

Let $(v_{(1)}, \cdots, v_{(N)})$ are ordered values of $(v_1, \cdots, v_N)$ and assume that there is no tie. Let $h_i = v_{(i+1)} - v_{(i)}$ for $i = 1, \cdots, N-1$. Then Q is a $N \times (N-2)$ band matrix with entries q_{ij}, for $i = 1, \cdots, N$ and $j = 2, \cdots, N-1$, given by

$$q_{j-1,j} = h_{j-1}^{-1}, q_{jj} = -h_{j-1}^{-1} - h_j^{-1}, \quad \text{and } q_{j+1,j} = h_j^{-1},$$

for $j = 2, \cdots, N-1$, and $q_{ij} = 0$ for $|i-j| \geq 2$. The columns of Q are numbered starting at $j = 2$, so that the top left element of Q is q_{12}.

$$Q = \begin{pmatrix} q_{12} & q_{13} & \cdots & q_{1,N-1} \\ q_{22} & q_{23} & \cdots & q_{2,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N-1,2} & q_{N-1,3} & \cdots & q_{N-1,N-1} \\ q_{N2} & q_{N3} & \cdots & q_{N,N-1} \end{pmatrix}$$

$$= \begin{pmatrix} h_1^{-1} & & & & \\ -h_1^{-1} - h_2^{-1} & h_2^{-1} & & 0 & \\ h_2^{-1} & -h_2^{-1} - h_3^{-1} & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & 0 & \ddots & -h_{N-3}^{-1} - h_{N-2}^{-1} & h_{N-2}^{-1} \\ & & & h_{N-2}^{-1} & -h_{N-2}^{-1} - h_{N-1}^{-1} \\ & & & & h_{N-1}^{-1} \end{pmatrix}$$

The symmetric band matrix R is $(N-2) \times (N-2)$ with elements r_{ij}, for i and j both from 2 to $(N-1)$, given by

$$r_{ii} = \frac{1}{3}(h_{i-1} + h_i) \quad \text{for } i = 2, \cdots, N-1,$$
$$r_{i,i+1} = r_{i+1,i} = \frac{1}{6}h_i \quad \text{for } i = 2, \cdots, N-2,$$

and $r_{ij} = 0$ for $|i-j| \geq 2$.

$$
\begin{aligned}
R &= \begin{pmatrix} r_{22} & r_{23} & \cdots & r_{2,N-1} \\ r_{32} & r_{33} & \cdots & r_{3,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N-1,2} & r_{N-1,3} & \cdots & r_{N-1,N-1} \end{pmatrix} \\
&= \begin{pmatrix} \frac{h_1+h_2}{3} & \frac{h_2}{6} & & 0 & \\ \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} & & \\ & \frac{h_3}{6} & \frac{h_3+h_4}{3} & \ddots & \\ & & \ddots & \ddots & \frac{h_{N-2}}{6} \\ & 0 & & \frac{h_{N-2}}{6} & \frac{h_{N-2}+h_{N-1}}{3} \end{pmatrix}
\end{aligned}
$$

1.5.2 Reinsch algorithm for weighted smoothing

Define the $(N-2)$-vector γ as $\gamma_i = \partial^2 g(v_{(i)})/\partial v_{(i)}^2$ for $i = 2, \cdots, N-1$, then we have $Q^T\mathbf{f} = R\gamma$ for natural cubic spline (refer to Green and Silverman (1994)).

The solution of (1.8) satisfies $\mathbf{f} = (W + \alpha QR^{-1}Q^T)^{-1}W\mathbf{z}$, which implies

$$W\mathbf{f} = W\mathbf{z} - \alpha QR^{-1}Q^T\mathbf{f} = W\mathbf{z} - \alpha Q\gamma.$$

Therefore, $\mathbf{f} = \mathbf{z} - \alpha W^{-1}Q\gamma$. Again, by $Q^T\mathbf{f} = R\gamma$,

$$
\begin{aligned}
Q^T\mathbf{f} &= Q^T\mathbf{z} - \alpha Q^T W^{-1}Q\gamma \\
R\gamma &= Q^T\mathbf{z} - \alpha Q^T W^{-1}Q\gamma \\
(R + \alpha Q^T W^{-1}Q)\gamma &= Q^T\mathbf{z}.
\end{aligned}
$$

The algorithm for weighted spline smoothing is

- Step 1 Evaluate the vector $Q^T\mathbf{z}$.
- Step 2 Find the non-zero diagonals of $R + \alpha Q^T W^{-1}Q$, and its Cholesky decomposition factors L and D.
- Step 3 Solve $LDL^T\gamma = Q^T\mathbf{z}$ for γ by forward and back substitution.
- Step 4 Use $\mathbf{f} = \mathbf{z} - \alpha W^{-1}Q\gamma$ to find $\mathbf{f}$.

Table 1.1: Simulation results for nonparametric model based on 200 replications.

	Relative bias[1]	*Empirical S.E.*[2]	*Empirical MISE*[3]	*prev.*[4](7.08%)
	$\alpha = \alpha_{GCV}$ ('random')			
$N = 5000$	0.026	0.422	0.0076	6.68
$N = 10000$	0.014	0.301	0.0022	6.86
	$\alpha = \alpha_{MISE}$ ('random')			
$N = 5000$	0.020	0.341	0.0041	6.79
$N = 10000$	0.011	0.247	0.0014	6.93
	$\alpha = \alpha_{GCV}$ ('alike')			
$N = 5000$	0.008	0.237	0.0006	6.99
$N = 10000$	0.005	0.180	0.0002	7.03
	$\alpha = \alpha_{MISE}$ ('alike')			
$N = 5000$	0.005	0.188	0.0002	7.02
$N = 10000$	0.003	0.143	0.0001	7.05

1. Relative bias: $\int \left|[\hat{f}(v) - f(v)]/f(v)\right| dF(v)$.
2. Empirical SE: $\int \hat{SE}\{\hat{f}(v)\}dF(v)$.
3. Empirical MISE: $\int [\hat{f}(v) - f(v)]^2 dF(v)$.
4. prevalence: $\int exp(\hat{f}(v))/[1 + exp(\hat{f}(v))]dF(v)$.

Figure 1.1: Point-wise average of the estimated nonparametric curve $\hat{f}(\cdot)$ for nonparametric model based on 200 replications. Left panel is for $N = 5000$ and right panel is for $N = 10000$: the blue dotted curve is the true values; the red solid curve is for α_{GCV} and 'random' pooling, while the red dashed curve is for α_{MISE} and 'random' pooling; the green solid curve is for α_{GCV} and 'alike' pooling, while the green dashed curve is for α_{MISE} and 'alike' pooling.

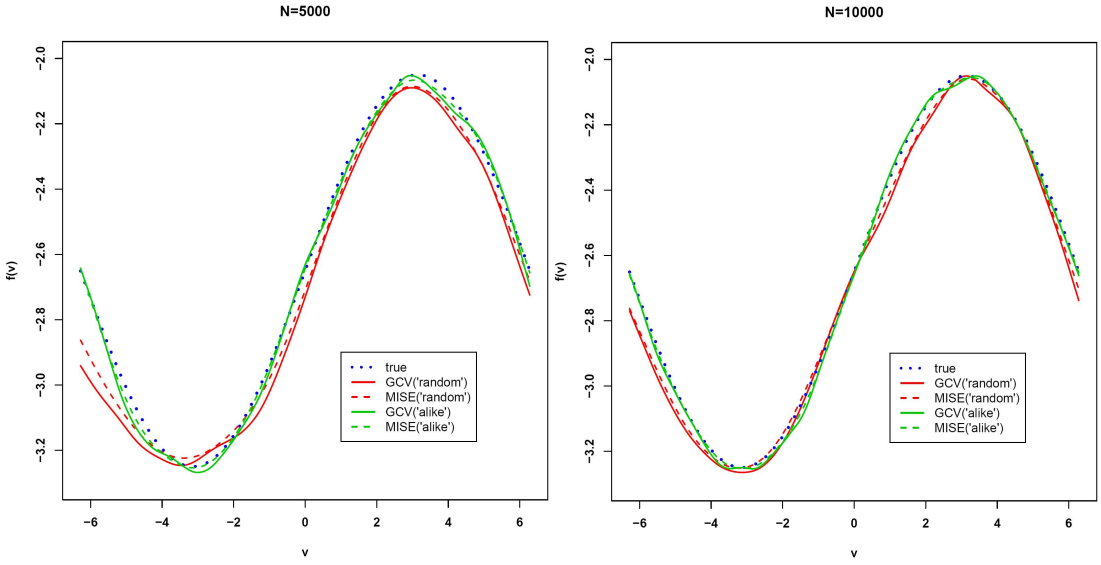


Figure 1.2: Empirical point-wise variances of the estimated nonparametric curve $\hat{f}(\cdot)$ for nonparametric model based on 200 replications. Left panel is for $N = 5000$ and right panel is for $N = 10000$: the red solid curve is for α_{GCV} and 'random' pooling, while the red dashed curve is for α_{MISE} and 'random' pooling; the green solid curve is for α_{GCV} and 'alike' pooling, while the green dashed curve is for α_{MISE} and 'alike' pooling.

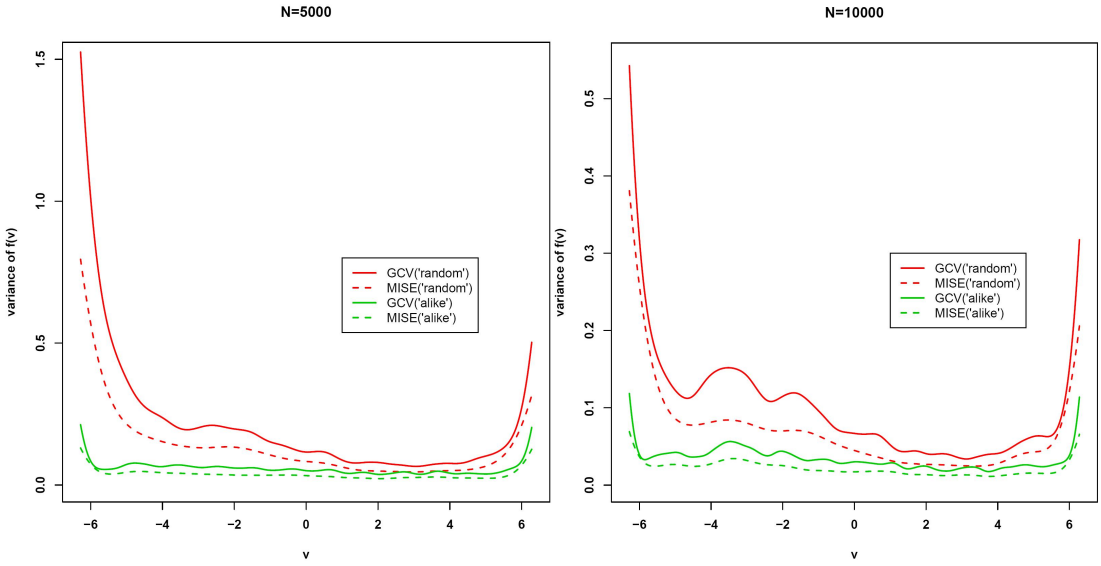


Table 1.2: Simulation results for semiparametric model based on 200 replications.

	$\hat{\beta}$		$\hat{f}(\cdot)$			
pooling strategy	Mean (-0.05)	*S.E.*	*Relative bias*	*Empirical S.E.*	*Empirical MISE*	*prev.* (7.8%)
'random'	-0.050	0.011	0.052	0.410	0.0018	7.58
'alike' by par-non	-0.050	0.006	0.036	0.226	0.0006	7.73
'alike' by non	-0.049	0.013	0.031	0.400	0.0006	7.87

Figure 1.3: Box-plot of the estimated β for semiparametric model for 'random', 'alike' by par-non and 'alike' by non pooling strategies based on 200 replications. The horizontal dotted line correspondes to the true value of β -0.05.

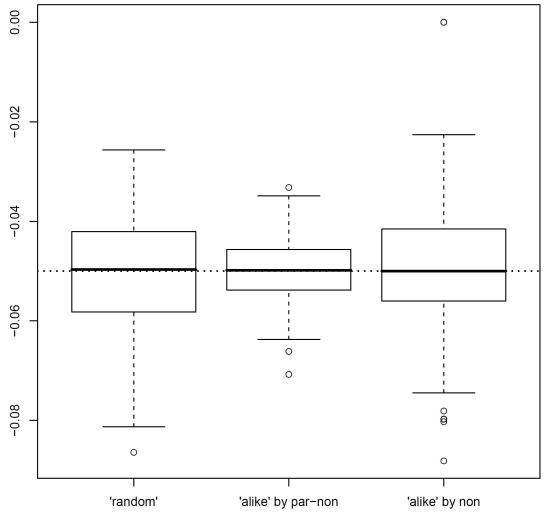
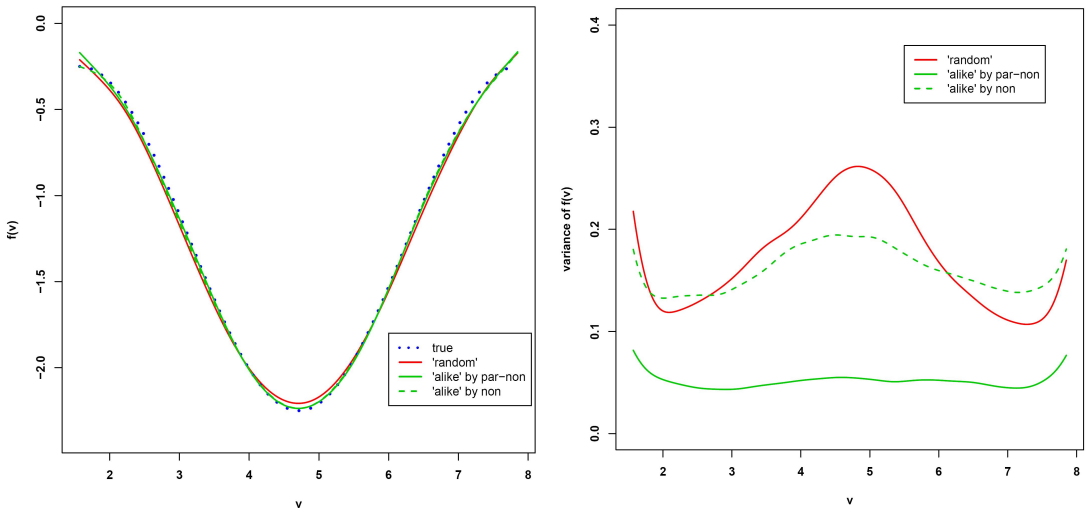


Figure 1.4: Simulation results of the estimated nonparametric curve $\hat{f}(\cdot)$ for semiparametric model based on 200 replications. Left panel is for point-wise average and right panel is for point-wise variance: the blue dotted curve is for the true values; the red solid curve is for 'random'; and the green solid curve is for 'alike' by par-non, while the green dashed curve is for 'alike' by non.



References

[1] Bilder, C.R. and Tebbs, J.M. (2009), "Bias, efficiency, and agreement for group-testing regression models," *Journal of Statistical Computation and Simulation* **79** No. 1, 67-80.

[2] Brookmeyer R. (1999). "Analysis of multistage pooling studies of biological specimens for estimating disease incidence and prevalence," *Biometrics* **55**, 608-612.

[3] Cardoso, M., Koerner, K., and Kubanek, B. (1998). "Mini-pool screening by nucleic acid testing for hepatitis B virus, hepatitis C virus, and HIV: Preliminary results," *Transfusion* **38**, 905-907.

[4] Chen, C., swallow, W. (1990). "Using group testing to estimate a proportion, and to test the binomial model," *Biometrics* **46**, 1035-1046.

[5] Chen, P, Tebbs, J.M. and Bilder C.R. (2009), "Group testing regression models with fixed and random effects," *Biometrics* **65** No. 4, 1270-1278.

[6] Dorfman, R. (1943). "The detection of defective members of large populations," *Annals of Mathematical Statistics* **14**, 436-440.

[7] Gastwirth, J.L. and Hammick P.A. (1989). "Estimation of prevalence of a rare disease, preserving anonymity of subjects by group testing: application to estimating the prevalence of AIDS antibodies in blood donors," *Journal of Statistical Planning and Inference* **22**, 15-27.

[8] Green, P. (1990). "On Use of the EM for Penalized Likelihood Estimation," *Journal of the Royal Statistical Society B* **52** No.3, 443-452.

[9] Green, P., and Silverman, B. (1994). "Nonparametric Regression and Generalized Linear Models," *Champman and Hall.*

[10] Gu, C. (1992). "Cross-validating Non-Gaussian Data," *Journal of Computational and Graphical Statistics* **1** No. 2, 169-179.

[11] Hughes-Oliver J., and Swallow, W. (1994). "A two-stage adaptive group-testing procedure for estimating small proportions," *Journal of the American Statistical Association* **89**, 982-993.

[12] Kacena, K., Quinn, S., Hartman, S., Quinn, T., and Gaydos, C. (1998a). "Pooling of urine samples for screening for *Neisseria gonorrhoeae* by ligase chain reaction: Accuracy and application," *Journal of Clinical Microbiology* **36**, 3624-3628.

[13] Kacena, K., Quinn, S., Howell, M., Madico, G., Quinn, T., and Gaydos, C. (1998b). "Pooling urine samples for ligase chain reaction screening for genital *Chlamydia trachomatis* infection in asymptomatic women," *Journal of Clinical Microbiology* **36**, 481-485.

[14] Lindan, C., Mathur, M., Kumta, S., Jerajani, H., Gogate, A., Schachter, J., and Moncada, J. (2005). "Utility of pooled urine specimens for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in men attending public sexually transmitted infection clinics in Mumbai, India, by PCR," *Journal of Clinical Microbiology* **43**, 1674-1677.

[15] Phatarfor R., and Sudbury A. (1994). "The use of a square array scheme in blood testing," *Statistics in Medicien* **13**, 2337-2343.

[16] Rours, G., Verkooyen, R., Willemse, H., van der Zwaan, E., van Belkum, A., de Groot, R., Verbrugh, H., and Ossewaarde, J. (2005). "Use of pooled urine samples and automated DNA isolation to achieve to improved sensitivity and coset-effectiveness of large-scale tesing for *Chlamydia trachomatis* in pregnant women," *Journal of Clinical Microbiology* **43**, 4684-4690.

[17] Thorburn, D., Dunda, D, McCruden, E., Cameron, S., Goldberg, D., Syminton, I., Kirk, A., and Mills, P. (2001). "A study of hepatitis C prevalence in healthcare workers in the west of Scotland," *Gut* **48**, 116-120.

[18] Vansteelandt, S., Goetghebeur, E., and Verstraeten, T. (2000). "Regression models for disease prevalence with diagnostic tests on pools of serum samples," *Biometrics* **56**, 1126-1133.

[19] Xie, M. (2001), "Regression Analysis of Group Testing Samples," *Statistics in Medicine* **20**, 1957-1969.

[20] Xie, M., Tatsuoka, K., Sacks, J., and Young S. (2001), "Group Testing With Blockers and Synergism," *Journal of the American Statistical Associatoin* **96** No. 453, 92-102.

[21] Yao, Y. and Hwang, F. (1990). "On optimal nested group testing algorithms," *Journal of Statistical Planning and Inference* **24**, 167-175.

Chapter 2

Semiparametric Efficient Estimation and the EM algorithm for Partially Linear Models with Missing Data

2.1 INTRODUCTION

Semiparametric models, which incorporate both the parametric and nonparametric components, have been studied extensively in statistics and econometrics since their introduction by Stein (1956). The available literature on semiparametric regression models mainly discusses estimation methods in complete data cases, but little literature studies efficient semiparametric inference in the presence of general missing data. In this chapter we prove that the asymptotic covariance of the estimator, which maximized the generalized profile likelihood, achieves the semiparametric efficiency bound under some conditions, and propose an estimation algorithm for the estimator of parametric component and nonparametric component in a partially linear regression model with general missing response values.

Consider a partially linear regression model with homoscedastic Gaussian error,

$$Y_i = W_i^T \beta + g(V_i) + \epsilon_i, \qquad \text{for} \quad i = 1, 2, \cdots, n, \tag{2.1}$$

where Y_i is the response variable, W_i is a $q \times 1$ vector of covariate variable, β is a $q \times 1$ vector of unknown parameter, $g(\cdot)$ is an unknown smooth function taking values in a compact subset of the real line, V_i is a $r \times 1$ vector of covariate variables, and ϵ_i has a normal distribution with mean 0 and variance σ^2. This model assumes that Y_i depends on covariate W_i in a linear way and depends on V_i in a nonparametric way by the unspecified smooth function $g(\cdot)$. The partially linear regression model was introduced by Engle et al. (1986) to study the effect of weather on electricity demand and has been widely used.

An efficient estimator of the parameter β for model (2.1) without missing data was given in Ma et al. (2006). In this chapter, we assume that $\mathbf{y} = (y_1, \cdots, y_n)$ are not completely observed, and denote the observed data by $\mathbf{z} = (z_1, \cdots, z_m)^T$. Our goal is to estimate the parameter of interest β in the presence of the infinite-dimensional nuisance parameter $g(\cdot)$ and missing data.

Missing data is a common problem in practice. Little and Rubin (2002) describe missing data types in detail. However, the existing literature mainly focuses on parametric regression model with missing data, but little literature discusses nonparametric or semiparametric regression with missing data. Wang et al. 2004 studied semiparametric regression with outcomes missing at random, but the missing patterns they considered only includes the case where each outcome is observed with certain probability, and missing otherwise. The missing structure we consider here is very general and is the same as the one studied in Wu (1983) and Green (1990). Suppose we have two sample spaces $\mathscr{Y}$ and $\mathscr{Z}$ and there is a many-to-one mapping from $\mathscr{Y}$ to $\mathscr{Z}$. Instead of observing the complete data $\mathbf{y}$ in $\mathscr{Y}$, we observe the incomplete data $\mathbf{z} = \mathbf{z}(\mathbf{y})$ in $\mathscr{Z}$. Let the density function of $\mathbf{y}$ be $f(\mathbf{y}|\theta)$ with parameters $\theta \in \Theta$ and let the density function of $\mathbf{z}$ be given by $f(\mathbf{z}|\theta) = \int_{\mathscr{Y}(z)} f(\mathbf{y}|\theta) d\mathbf{y}$, where $\mathscr{Y}(z) = \{\mathbf{y} : \mathbf{z}(\mathbf{y}) = \mathbf{z}\}$. This type of missingness includes group testing studied in Xie (2001) and is broader than what considered in Wang et al. (2004).

The motivation of our work is from several aspects, for instance, income report and education score report. In the annal household income report, it is common that only the median income of each town is recorded instead of each household for the privacy reason; in the education score report, average score of each class is available instead of individual scores. In these situations, we want to estimate the parameters given that only the median or mean of each group is observed. Our estimation method can be applied to more general missing data cases as long as there is a many to one mapping between complete data and observed data.

One of the effective methods for dealing with the missing data is the EM algorithm. The EM algorithm is a general approach to maximum likelihood estimation, and it can be used both in parametric regression models, see Dempster et al. (1977) and Wu

(1983), and penalized likelihood estimation when the parameter represents a smooth function that has been discretized, see Green (1990). Silverman et al. (1990) modified the EM approach by introducing a simple smoothing step at each EM iteration and developed the EMS algorithm.

Much work has been done on modeling semiparametric regression for complete data. In most cases, the main interest or objective is to estimate the finite dimensional parametric part, while the nonparametric component is considered as the infinite dimensional nuisance parameter. For some particular classes of semiparametric models, efficient estimator of the parametric component is given in the literature. Severini and Wong (1992) proposed an estimation method maximizing the generalized profile likelihood under the conditionally parametric model and proved that the estimation method leaded to an asymptotically efficient estimator of the parameter of interest; Ahmad et al. (2005) proposed a general series method to estimate semiparametric partially linear varying coefficient model and and showed that the estimator of the finite dimensional parameters is semiparametrically efficient when the error is conditionally homoskedastic; Ma et al. (2006) proposed a family of consistent estimators and showed that the optimal semiparametric efficiency bound can be reached by a semiparametric kernel estimator in this family; Boente et al. (2006) introduced a family of robust estimates under a generalized partially linear model and showed that their estimates of parametric component had root n convergence rate; Xie et al. (2008) developed efficient semiparametric inference for the parametric component under a class of heteroscedastic generalized linear regression models in which a subset of the regression parameters were rescaled nonparametrically; Lam and Fan (2008) considered the generalized varying coefficient partially linear model allowing the number of predictors to increase with the sample size and established root-n asymptotic results. Here efficiency refers to the usual asymptotic efficiency, see Newey (1990) for the detailed discussion on semiparametric efficiency bounds.

In this chapter we use the estimation method, which maximized the generalized profile likelihood and prove that the estimator is root-n consistent and efficient under the conditions given in appendix. We also propose an estimation algorithm, which runs

iteratively between fitting parametric components and fitting nonparametric components while holding the other fixed. The estimators from this iterative algorithm are conditional expectation (conditioned on the observed data) of the semiparametric efficient estimator without missing data. The algorithm utilizes EM algorithm to estimate the parametric components by a semiparametric estimating equation and to estimate the nonparametric components by smoothing methods.

The rest of the chapter is organized as follows. In Section 2.2 we present our estimation method and the large sample properties of the estimator, including consistency and efficiency. Estimation algorithm using EM algorithm is given in Section 2.3. After that we evaluate the finite sample performance of proposed algorithm by two simulation studies in Section 2.4. Discussion is given in Section 2.5 and Appendix provides technical details and assumptions.

2.2 GENERALIZED PROFILE LIKELIHOOD APPROACH

2.2.1 Generalized profile likelihood and regularity conditions

Let $(y_i, w_i, v_i, \epsilon_i)$, for $i = 1, 2, \cdots, n$, be n independently and identically distributed replicates of (Y, W, V, ϵ) and denote $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\mathbf{w} = (w_1^T, \cdots, w_n^T)^T$, $\mathbf{v} = (v_1, \cdots, v_n)^T$, $\mathbf{g} = g(\mathbf{v}) = \big(g(v_1), \cdots, g(v_n)\big)^T$ and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T$. The complete log-likelihood function for model (2.1) is

$$l\big(\beta, g(\cdot); \mathbf{y}, \mathbf{w}, \mathbf{v}\big) = \log f\big(\mathbf{y}, \mathbf{w}, \mathbf{v}; \beta, g(\cdot)\big) = \log f\big(\mathbf{y}; \mathbf{w}, \mathbf{v}, \beta, g(\cdot)\big) + \log f(\mathbf{w}, \mathbf{v}),$$

where

$$\log f\big(\mathbf{y}; \mathbf{w}, \mathbf{v}, \beta, g(\cdot)\big) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\big[y_i - w_i^T\beta - g(v_i)\big]^2,$$

and $f(\mathbf{w}, \mathbf{v})$ is the marginal density function of $(\mathbf{w}, \mathbf{v})$. The observed log-likelihood function is

$$l\big(\beta, g(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) = \log f\big(\mathbf{z}, \mathbf{w}, \mathbf{v}; \beta, g(\cdot)\big) = \log f\big(\mathbf{z}; \mathbf{w}, \mathbf{v}, \beta, g(\cdot)\big) + \log f(\mathbf{w}, \mathbf{v}).$$

Since $f(\mathbf{w}, \mathbf{v})$ does not depend on β or $g(\cdot)$, we consider $\log f\big(\mathbf{y}; \mathbf{w}, \mathbf{v}, \beta, g(\cdot)\big)$ as the complete log-likelihood and $\log f\big(\mathbf{z}; \mathbf{w}, \mathbf{v}, \beta, g(\cdot)\big)$ as observed log-likelihood and denote them by $l\big(\beta, g(\cdot); \mathbf{y}, \mathbf{w}, \mathbf{v}\big)$ and $l\big(\beta, g(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)$ respectively afterward.

We use the same estimation method as in Severini and Wong (1992). For any fixed β, suppose g_β is any least favorable curve and $\hat{g}_\beta$ is a consistent estimator of g_β for that β. Then $l\big(\beta, \hat{g}_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) = \log f\big(\mathbf{z}; \mathbf{w}, \mathbf{v}, \beta, \hat{g}_\beta(\cdot)\big)$ is called a generalized profile likelihood for β. The estimator of β can then be obtained by maximizing $\log f(\mathbf{z}; \mathbf{w}, \mathbf{v}, \beta, \hat{g}_\beta(\cdot))$. The concept of least favorable curve and the least favorable direction have been discussed in Severini and Wong (1992) in detail.

For each m, define $\hat{\beta} \equiv \hat{\beta}_m$ to be any element of parameter space B satisfying

$$l\big(\hat{\beta}, \hat{g}_{\hat{\beta}}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) = \sup_{\beta \in B} l\big(\beta, \hat{g}_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big),$$

then the the estimator $\hat{\beta}$ is a consistent and semiparametric efficient estimator of true β, β_0 under some coditions. To prove this, we need the following assumptions and the conditions given in Appendix 2.6.1.

First, we require that the joint probability density function of $\mathbf{z} = (z_1, \cdots, z_m)$ satisfies the following identifiability (I) and continuity (C) conditions.

CONDITIONS I. For a fixed but arbitrary β, and any least favorable curve $g_\beta(\cdot)$, where $\beta \in B$, $g_\beta(v_i) \in R$, and $g_\beta(\cdot)|_{\beta=\beta_0} = g_0(\cdot)$, suppose that

$$m^{-1}\Big\{ l\big(\beta, g_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) - E_0\big[l\big(\beta, g_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)\big] \Big\} \to_p 0$$

and

$$m^{-1} E_0\big[l\big(\beta, g_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)\big] \to_p l_0(\beta),$$

where $l_0(\beta)$ is the limiting function of $m^{-1} l\big(\beta, g_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)$, as $m \to \infty$.
The limiting function is used to identify the true parameters (β_0, g_0) and we assume that $l_0(\beta)$ is maximized at true β, β_0.

CONDITIONS C. Suppose that $\log f\big(\mathbf{z}; \mathbf{w}, \mathbf{v}, \beta, g(\cdot)\big)$ is Lipschitz continuous in β and $g(\cdot)$. Therefore, there exist A_m and B_m such that

$$\begin{aligned} &\frac{1}{m} \left| \log f\big(\mathbf{z}; \mathbf{w}, \mathbf{v}, \beta_1, g_1(\cdot)\big) - \log f\big(\mathbf{z}; \mathbf{w}, \mathbf{v}, \beta_2, g_2(\cdot)\big) \right| \\ &\leq A_m \left| \beta_1 - \beta_2 \right| + B_m \left\| g_1 - g_2 \right\|, \end{aligned}$$

where A_m and B_m are bounded by constants A and B respectively.

Second, the estimator of the nonparametric part $\hat{g}_\beta(\cdot)$ must satisfy the Nuisance parameter conditions, Conditions NP in Severini and Wong (1992).

CONDITIONS NP. (a) For each v in a finite interval $[a, b]$ and each $\beta \in B$, $\hat{g}_\beta(v)$ converges in probability to some constant as $m \to \infty$; denote that constant by $\tilde{g}_\beta(v)$. Assume that for each $\beta \in B, \tilde{g}_\beta \in \Lambda = \left\{h \in C^2[a, b] : h(v) \in int(R) \text{ for all } v \in [a, b]\right\}$, and that for all $r, s = 0, 1, 2, r + s \leq 2$,

$$\frac{\partial^{r+s}\tilde{g}_\beta(v)}{\partial v^r \partial \beta^s} \quad \text{and} \quad \frac{\partial^{r+s}\hat{g}_\beta(v)}{\partial v^r \partial \beta^s}$$

exist. Let

$$\tilde{g}_0 = \tilde{g}_\beta\Big|_{\beta=\beta_0} \quad \text{and} \quad \tilde{g}_0' = \frac{d}{d\beta}\tilde{g}_\beta\Big|_{\beta=\beta_0}.$$

Then suppose

$$\|\hat{g}_0 - \tilde{g}_0\| = o_p(m^{-\alpha}) \quad \text{and} \quad \left\|\hat{g}_0' - \tilde{g}_0'\right\| = o_p(m^{-\beta}),$$

where $\alpha + \beta \geq 1/2$ and $\alpha \geq 1/4$.

Furthermore, suppose that $\sup_{\beta\in B} \|\hat{g}_\beta - \tilde{g}_\beta\|$, $\sup_{\beta\in B} \left\|\hat{g}_\beta' - \tilde{g}_\beta'\right\|$ and $\sup_{\beta\in B} \left\|\hat{g}_\beta'' - \tilde{g}_\beta''\right\|$ are all of order $o_p(1)$ as $m \to \infty$.

For some $\delta > 0$, assume that

$$\left\|\frac{\partial\hat{g}_0}{\partial v} - \frac{\partial\tilde{g}_0}{\partial v}\right\| = o_p(m^{-\delta}) \quad \text{and} \quad \left\|\frac{\partial\hat{g}_0'}{\partial v} - \frac{\partial\tilde{g}_0'}{\partial v}\right\| = o_p(m^{-\delta}).$$

(b)The curve $\tilde{g}_\beta$ is a least favorable curve.

2.2.2 Theoretical results

The estimator, which maximizes the generalized profile likelihood function has the following large sample properties.

Theorem 1 states that if $\hat{\beta}$ maximizes the generalized profile likelihood, then it is a consistent estimator of true β, β_0, under some conditions.

Theorem 1 (Consistency). Under the Conditions I, C and NP,

$$\hat{\beta} \to_p \beta_0 \quad \text{as } m \to \infty.$$

The proof is given in Appendix 2.6.2.

Theorem 2 establishes that $\hat{\beta}$ is asymptotically normally distributed with asymptotic variance equal to the semiparametric efficiency bound, i_β^{-1}.

Theorem 2 (Efficiency). Under the conditions $C1 - C10$ given in Appendix,

$$\sqrt{m}(\hat{\beta} - \beta_0) \rightarrow_{\mathscr{D}} N(0, i_\beta^{-1}),$$

where i_β^{-1} is the semiparametric efficiency bound.

Let

$$\hat{i}_\beta = -\frac{1}{m} \frac{dl^2\big(\beta, \hat{g}_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)}{d\beta^2}\Big|_{\beta=\hat{\beta}}.$$

Then,

$$\hat{i}_\beta \rightarrow_p i_\beta \quad \text{as } m \rightarrow \infty.$$

The proof is given in Appendix 2.6.2.

2.3 ESTIMATION ALGORITHM

2.3.1 Iterative algorithm

The observed log-likelihood $\log f\big(\mathbf{z}; \mathbf{w}, \mathbf{v}, \beta, g(\cdot)\big)$ depends on the missing structure and can be very complicated. Therefore, direct maximization of the observed log-likelihood function may not be practical. However, the complete log-likelihood function has a simple form and hence the EM algorithm is an effective way to deal with missing data in this case. We propose an EM algorithm for estimating the parametric part β making only smoothness assumptions on the unknown function $g(\cdot)$ in model (2.1) with missing data. The key idea of the algorithm is similar to the backfitting algorithm and is based on the concept of generalized profile likelihood. In this algorithm, first we fix the parametric component and estimate the nonparametric component using EM algorithm and some smoothing method; this estimator depends on the value at which the parametric component is held fixed, that is nonparametric part can be considered as a function of parametric part. This estimator of nonparametric component is then

used to create a generalized profile likelihood of parametric part using the observed log-likelihood function. Then, the estimator of parametric component can be obtained by using EM algorithm and semiparametric estimating equation. In implementation, the algorithm iterates between estimating parametric component and estimating nonparametric component while holding the other fixed until converge, and utilizes the EM algorithm in each iteration to deal with the missing data.

For simplicity, we consider the case in which the covariate V is one-dimensional. Extension to multivariate V involves no fundamentally new idea. Then the estimation algorithm iterates between the following two modules:

- Estimating parametric component: Fix the current estimator of nonparametric function and its first derivative with respect to β, say $\hat{g}^{cur}(\cdot)$ and $\hat{g}'^{cur}(\cdot)$. Then update the estimate of parametric part $\hat{\beta}^{new}$ and $\hat{\sigma}^{new}$ using the estimating equation and EM algorithm.

- Estimating nonparametric component: Fix the current estimators of β and σ, say $\hat{\beta}^{cur}$ and $\hat{\sigma}^{cur}$. Update the estimator of nonparametric function $\hat{g}(\cdot)$ and its first derivative to β, say $\hat{g}'(\cdot)$ iteratively until converge to get $\hat{g}^{new}(\cdot)$ and $\hat{g}'^{new}(\cdot)$ using EM algorithm and smoothing methods.

In the first module, consider updating the estimator of parametric component β given $\hat{\beta}^{cur}$, $\hat{\mathbf{g}}^{cur}$, $\hat{\mathbf{g}}'^{cur}$ and $\hat{\sigma}^{cur}$. By taking conditional expectation, the observed log-likelihood can be written as

$$
\begin{aligned}
&\log f(\mathbf{z};\mathbf{w},\mathbf{v},\beta,\hat{\mathbf{g}}^{cur}) \\
&= E\big[\log f(\mathbf{y};\mathbf{w},\mathbf{v},\beta,\hat{\mathbf{g}}^{cur})|\mathbf{z},\hat{\beta}^{cur},\hat{\mathbf{g}}^{cur}\big] - E\big[\log f(\mathbf{y};\mathbf{z},\mathbf{w},\mathbf{v},\beta,\hat{\mathbf{g}}^{cur})|\mathbf{z},\hat{\beta}^{cur},\hat{\mathbf{g}}^{cur}\big] \\
&= Q(\beta,\hat{\mathbf{g}}^{cur}|\hat{\beta}^{cur},\hat{\mathbf{g}}^{cur}) - H(\beta,\hat{\mathbf{g}}^{cur}|\hat{\beta}^{cur},\hat{\mathbf{g}}^{cur}),
\end{aligned}
$$

where

$$
\begin{aligned}
&Q(\beta,\hat{\mathbf{g}}^{cur}|\hat{\beta}^{cur},\hat{\mathbf{g}}^{cur}) \\
&= -\frac{n}{2}\log(2\pi\hat{\sigma}^{cur2}) - \frac{1}{2\hat{\sigma}^{cur2}}\sum_{i=1}^{n}\big[E(y_i^2|\mathbf{z},\mathbf{w},\hat{\beta}^{cur},\hat{\mathbf{g}}^{cur},\hat{\sigma}^{cur2}) \\
&-2E(y_i|\mathbf{z},\mathbf{w},\hat{\beta}^{cur},\hat{\mathbf{g}}^{cur},\hat{\sigma}^{cur2})\big(w_i^T\beta+\hat{g}^{cur}(v_i)\big) + \big(w_i^T\beta+\hat{g}^{cur}(v_i)\big)^2\big].
\end{aligned}
$$

By information inequality, maximizing Q function for β increases observed log-likelihood. Hence $\hat{\beta}^{new}$ can be updated by the following E-step and M-step:

- E-step. Calculate

$$\mathbf{u1} = E(\mathbf{y}|\mathbf{z}, \mathbf{w}, \hat{\beta}^{cur}, \hat{\mathbf{g}}^{cur}, \hat{\sigma}^{cur}).$$

- M-step. Solve the following estimating equation to get $\hat{\beta}^{new}$,

$$\big[\mathbf{u1} - \mathbf{w}\beta - \hat{g}^{cur}(\mathbf{v})\big]^T \big[\mathbf{w} + \hat{g}'^{cur}(\mathbf{v})\big] = \mathbf{0}_{1\times q}. \tag{2.2}$$

Here $E(\mathbf{y}|\mathbf{z}, \mathbf{w}\beta, \mathbf{g}, \sigma) = \big(E(y_1|\mathbf{z}, \mathbf{w}\beta, \mathbf{g}, \sigma), \cdots, E(y_n|\mathbf{z}, \mathbf{w}, \beta, \mathbf{g}, \sigma)\big)$ and $\hat{g}'(\mathbf{v}) = \partial\hat{g}(\mathbf{v})/\partial\beta$. The estimating equation (2.2) is similar to the estimating equation (c) for complete data in Ma et al. (2006), with $\mathbf{y}$ replaced by $\mathbf{u1}$, and is derived from the efficient score function. The $\hat{\sigma}^{new}$ can be updated by maximum likelihood

$$\begin{aligned}\hat{\sigma}^{new2} &= \frac{1}{n}\sum_{i=1}^{n}\big[E(y_i^2|\mathbf{z}, \mathbf{w}, \hat{\beta}^{new}, \hat{\mathbf{g}}^{cur}, \hat{\sigma}^{cur}) - 2E(y_i|\mathbf{z}, \mathbf{w}, \hat{\beta}^{new}, \hat{\mathbf{g}}^{cur}, \hat{\sigma}^{cur})\big(w_i^T\hat{\beta}^{new} \\ &\quad + \hat{g}^{cur}(v_i)\big) + \big(w_i^T\hat{\beta}^{new} + \hat{g}^{cur}(v_i)\big)^2\big]. \end{aligned}\tag{2.3}$$

Repeat (2.3) until converge to get the estimator of σ corresponding to $\hat{\beta}^{new}$, $\hat{\sigma}^{new}$.

In the second module, maximum likelihood approach is used to estimate the function $g(\cdot)$ and $g'(\cdot)$ nonparametrically for a fixed β and a fixed σ. Given $\hat{\beta}^{new}$, $\hat{\mathbf{g}}^{cur}$ and $\hat{\sigma}^{new}$, the observed log-likelihood can be written as

$$\begin{aligned}&\log f(\mathbf{z}; \mathbf{w}, \hat{\beta}^{new}, \mathbf{g}) \\ &= E\big[logf(\mathbf{y}; \mathbf{w}, \hat{\beta}^{new}, \mathbf{g})|\mathbf{z}, \hat{\beta}^{new}, \hat{\mathbf{g}}^{cur}\big] - E\big[logf(\mathbf{y}; \mathbf{z}, \mathbf{w}, \hat{\beta}^{new}, \mathbf{g})|\mathbf{z}, \hat{\beta}^{new}, \hat{\mathbf{g}}^{cur}\big] \\ &= Q(\hat{\beta}^{new}, \mathbf{g}|\hat{\beta}^{new}, \hat{\mathbf{g}}^{cur}) - H(\hat{\beta}^{new}, \mathbf{g}|\hat{\beta}^{new}, \hat{\mathbf{g}}^{cur}),\end{aligned}$$

where

$$\begin{aligned}&Q(\hat{\beta}^{new}, \mathbf{g}|\hat{\beta}^{new}, \hat{\mathbf{g}}^{cur}) \\ &= -\frac{n}{2}\log(2\pi\hat{\sigma}^{new2}) - \frac{1}{2\hat{\sigma}^{new2}}\Big\{\Big\|\mathbf{g} - \big[E(\mathbf{y}|\mathbf{z}, \mathbf{w}, \hat{\beta}^{new}, \hat{\mathbf{g}}^{cur}, \hat{\sigma}^{new2}) - \mathbf{w}\hat{\beta}^{new}\big]\Big\|^2 \\ &+ \sum_{i=1}^{n}\big[E(y_i^2|\mathbf{z}, \mathbf{w}, \hat{\beta}^{new}, \hat{\mathbf{g}}^{cur}, \hat{\sigma}^{new2}) - E(y_i|\mathbf{z}, \mathbf{w}, \hat{\beta}^{new}, \hat{\mathbf{g}}^{cur}, \hat{\sigma}^{new2})^2\big]\Big\}.\end{aligned}$$

Again by the information inequality, maximizing $Q(\hat{\beta}^{new}, \mathbf{g}|\hat{\beta}^{new}, \hat{\mathbf{g}}^{cur})$ for $\mathbf{g}$ increases $\log f(\mathbf{z}; \mathbf{w}, \hat{\beta}^{new}, \mathbf{g})$. Hence, $\check{\mathbf{g}}^{new}$ and $\check{\mathbf{g}}'^{new}$ can be updated by the following E-step and M-step (set $\check{\mathbf{g}}^{cur} = \hat{\mathbf{g}}^{cur}$ and $\check{\mathbf{g}}'^{cur} = \hat{\mathbf{g}}'^{cur}$ before EM steps):

- E-step. Calculate

$$\mathbf{u2} = E(\mathbf{y}|\mathbf{z}, \mathbf{w}, \hat{\beta}^{new}, \check{\mathbf{g}}^{cur}, \hat{\sigma}^{new}),$$

and take derivative to β,

$$\mathbf{u3} = D(\mathbf{z}, \beta, \mathbf{g}, \mathbf{g}'), \sigma)|_{\beta=\hat{\beta}^{new}, \mathbf{g}=\check{\mathbf{g}}^{cur}, \mathbf{g}'=\check{\mathbf{g}}'^{cur}, \sigma=\hat{\sigma}^{cur}}.$$

- M-step. Update $\check{\mathbf{g}}^{\mathbf{new}}$ by

$$\check{\mathbf{g}}^{\mathbf{new}} = \mathbf{u2} - \mathbf{w}\hat{\beta}^{new}.$$

and update $\check{\mathbf{g}}'^{\mathbf{new}}$ by

$$\check{\mathbf{g}}'^{\mathbf{new}} = \mathbf{u3} - \mathbf{w}.$$

Smooth the estimators by

$$\check{g}^{new}(\mathbf{v}) = M\mathbf{u2} - (M\mathbf{w})\hat{\beta}^{new},$$

and

$$\check{g}'^{new}(\mathbf{v}) = M\mathbf{u3} - M\mathbf{w},$$

where

$$\begin{aligned} D\big(\mathbf{z}, \beta, \mathbf{g}, \mathbf{g}'), \sigma\big) &= \partial E\big(\mathbf{y}|\mathbf{z}, \mathbf{w}, \beta, \mathbf{g}, \sigma\big)/\partial\beta \\ &= \Big(\frac{\partial E\big(\mathbf{y}|\mathbf{z}, \mathbf{w}, \beta, \mathbf{g}, \sigma\big)}{\partial\beta_1}, \cdots, \frac{\partial E\big(\mathbf{y}|\mathbf{z}, \mathbf{w}, \beta, \mathbf{g}, \sigma\big)}{\partial\beta_q}\Big) \end{aligned}$$

is a function of β, σ, $\mathbf{g}$ and $\mathbf{g}'$. The projection matrix M is defined by $M = P(P^TP)^{-1}P^T$, where P is a $n \times s$ matrix, i^{th} row of which is s B-spline basis functions of v_i. The B-spline basis functions and the selection of s will be discussed in the simulation part. In the M-step, we can use other nonparametric smoothing methods to smooth the estimator of nonparametric part. However, by using projection matrix, we can connect our estimator to the efficient estimator without missing data given in Ahmad et al. (2005). The connection is stated in Theorem 3.

Repeating above two steps iteratively until converge gives $\hat{g}^{new}(\cdot)$ and $\hat{g}'^{new}(\cdot)$, which are the updated estimators of nonparametric function $g_\beta(\cdot)$ and its derivative to β, $g'_\beta(\cdot)$.

In conclusion, our estimation method iterates between the two modules and repeats the EM algorithm in each module until converge, that is, both $\left\|\hat{\beta}^{new} - \hat{\beta}^{cur}\right\|$ and $\|\hat{g}^{new} - \hat{g}^{cur}\|$ are very small.

2.3.2 Connection to the efficient estimators from complete data

There is a relationship between the estimators from the above iterative algorithm and the efficient estimator from the complete data, which is stated in Theorem 3.

Theorem 3 (Connection to the efficient estimator without missing data). The estimators, $\hat{\beta}$ and $\hat{g}(\cdot)$, from the above iterative algorithm have the following connection to the semiparametric efficient estimator without missing data. The estimators, $\hat{\beta}$ and $\hat{\mathbf{g}} = \hat{g}(\mathbf{v})$, are the solution of

$$\begin{cases} \beta = E[\hat{\beta}^*|\mathbf{z}, \mathbf{w}, \beta, \mathbf{g}] \\ \mathbf{g} = E[\hat{\mathbf{g}}^*|\mathbf{z}, \mathbf{w}, \beta, \mathbf{g}], \end{cases}$$

where $\hat{\beta}^*$ is the semiparametric efficient estimator given $(y_1, \cdots, y_n)$, and $\hat{\mathbf{g}}^*$ is the corresponding estimator of nonparametric component. By Ahmad et al. (2005), $\hat{\beta}^*$ and $\hat{\mathbf{g}}^*$ are given by

$$\begin{cases} \hat{\beta}^* = \big[(\mathbf{w} - M\mathbf{w})^T(\mathbf{w} - M\mathbf{w})\big]^-(\mathbf{w} - M\mathbf{w})^T(\mathbf{y} - M\mathbf{y}) \\ \hat{\mathbf{g}}^* = M(\mathbf{y} - \mathbf{w}\hat{\beta}^*). \end{cases}$$

The proof is given is Appendix 2.6.2.

2.3.3 Estimator of asymptotic variance of $\hat{\beta}$

Theorem 2 gives the estimator of the asymptotic variance of $\hat{\beta}$. However, in many missing data cases, direct calculation of the estimator given in Theorem 2 may be difficult and the following approach can be applied to obtain the estimator. Suppose the length of β is 1.

$$\begin{aligned} & -\frac{1}{m}\frac{d^2 l\big(\beta, g_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big)}{d\beta^2} \\ = \; & -\frac{1}{m}\Big\{E\big[\frac{d^2 \log f\big(\mathbf{y}; \mathbf{w}, \mathbf{v}, \beta, g_\beta(\cdot)\big)}{d\beta^2}|\mathbf{z}, \mathbf{w}, \mathbf{v}, \beta, g_\beta(\cdot)\big] \\ & \qquad -E\big[\frac{d^2 \log f\big(\mathbf{y}|\mathbf{z}, \mathbf{w}, \mathbf{v}, \beta, g_\beta(\cdot)\big)}{d\beta^2}|\mathbf{z}, \mathbf{w}, \mathbf{v}, \beta, g_\beta(\cdot)\big]\Big\} \\ = \; & -\frac{1}{m}\Big\{E\big[\frac{d^2 \log f\big(\mathbf{y}; \mathbf{w}, \mathbf{v}, \beta, g_\beta(\cdot)\big)}{d\beta^2}|\mathbf{z}, \mathbf{w}, \mathbf{v}, \beta, g_\beta(\cdot)\big] \\ & \qquad +Var\big[\frac{d \log f\big(\mathbf{y}; \mathbf{w}, \mathbf{v}, \beta, g_\beta(\cdot)\big)}{d\beta}|\mathbf{z}, \mathbf{w}, \mathbf{v}, \beta, g_\beta(\cdot)\big]\Big\} = -\frac{1}{m}(I1 + I2). \quad (2.4) \end{aligned}$$

Notice that

$$\begin{aligned}\frac{d\log f\big(\mathbf{y};\mathbf{w},\mathbf{v},\beta,g_\beta(\cdot)\big)}{d\beta} &= \frac{1}{\sigma^2}\big(\mathbf{y}-\mathbf{w}^T\beta-g_\beta(\mathbf{v})\big)^T\big(\mathbf{w}+g'_\beta(\mathbf{v})\big)\\ \frac{d^2\log f\big(\mathbf{y};\mathbf{w},\mathbf{v},\beta,g_\beta(\cdot)\big)}{d\beta^2} &= -\frac{1}{\sigma^2}\big(\mathbf{w}+g'_\beta(\mathbf{v})\big)^T\big(\mathbf{w}+g'_\beta(\mathbf{v})\big),\end{aligned}$$

therefore,

$$I1=-\frac{1}{\sigma^2}\big(\mathbf{w}+g'_\beta(\mathbf{v})\big)^T\big(\mathbf{w}+g'_\beta(\mathbf{v})\big)$$

and

$$\begin{aligned}I2=\frac{1}{\sigma^4}\sum_{i,j=1}^{n}\Big\{&\big[E\big(y_iy_j|\mathbf{z},\mathbf{w},\mathbf{v},\beta,g_\beta(\cdot)\big)-E\big(y_i|\mathbf{z},\mathbf{w},\mathbf{v},\beta,g_\beta(\cdot)\big)E\big(y_j|\mathbf{z},\mathbf{w},\mathbf{v},\beta,g_\beta(\cdot)\big)\big]\\ &\big(w_i+g'_\beta(v_i)\big)\big(w_j+g'_\beta(v_j)\big)\Big\}.\end{aligned}$$

Hence, if we can calculate $E\big(y_iy_j|\mathbf{z},\mathbf{w},\mathbf{v},\beta,g_\beta(\cdot)\big)$, then the estimator of asymptotic variance can be obtained by substituting the estimators for the true parameters in (2.4) with $I1$ and $I2$ given above.

2.4 SIMULATION STUDIES

In this section we use two simulation studies to examine the finite sample performance of the proposed estimation methodology. The two simulation studies use the same set of complete data, but with different missing data structures. In the first study, the observed likelihood function has a simple form, while in the other study the observed response variable has a complicated likelihood function. And we show that our algorithm works well in both cases.

Consider the following data generating process from a partially linear regression model:

$$y_i=w_i\beta+g(v_i)+\epsilon_i,\quad i=1,\cdots,n,$$

where $g(v_i)=1+6\sin(2\pi v_i)$ and $\beta=4$. The error ϵ_i are i.i.d normal random variables with mean 0 and standard deviation $\sigma=0.25$, $w_i=u_{1i}+2u_{2i}$ and $v_i=u_{2i}+u_{3i}$, where u_{ji}, $j=1,2,3$ are i.i.d from uniform U[0, 0.5]. The (y_i,w_i,v_i) for $i=1,\cdots,n$ are complete data set. Both of the two simulation studies use this complete data set, but

with different missing data structures. First, y_i are randomly grouped by size $k = 5$, and only the sum of each group is observed instead of individual y_i. Second, we assume only the maximum of each group is observed, where y_i are still randomly grouped by size $k = 5$. Under this missing structure, the distribution of the observed data has a complicated form and direct estimation from the likelihood function of observed data is hard to obtain. The group size k can be any value, as long as the number of groups has the same order as n. Without loss of generality, we take $k = 5$, and take n as a multiple of k for the simplicity of implementation. The estimator $\hat{\beta}$ from complete data (suppose that $\mathbf{y} = (y_1, \cdots, y_n)$ are available completely) using the estimation method given by Ahmad et al. (2005) are obtained for the purpose of comparison. The sample sizes are $n = 500$ and $n = 1000$ and we repeat 500 times for both simulations.

We use the proposed algorithm to fit the simulation data and obtain the point estimator of β, estimated mean squared error (MSE) of $\hat{\beta}$ defined by $MSE(\hat{\beta}) = \sum_{j=1}^{500} (\hat{\beta} - \beta)^2/500$, estimator of the asymptotic variance of $\hat{\beta}$ to measure the performance of parametric estimation and use estimated mean average squared error (MASE) of $\hat{g}(\cdot)$ defined by $MASE(\hat{g}(\cdot)) = \sum_{j=1}^{500} [\frac{1}{n} \sum_{i=1}^{n} (\hat{g}(v_i) - g(v_i))^2]/500$ to measure the performance of nonparametric estimation, where $\hat{\beta}$ and $\hat{g}(v_i)$ are the estimates of β and $g(v_i)$ from the jth replication respectively. We use a univariate cubic B-spline basis function defined by

$$B(v|t_0, \cdots, t_4) = \frac{1}{3} \sum_{j=0}^{4} (-1)^j \big[\max(0, v - t_j) \big]^3,$$

where $t_0, \cdots, t_4$ are the evenly-spaced design knots. In fitting the nonparametric part $g(\cdot)$, we need to select the number of interior knots r of the B-spline as in any nonparametric model fitting. In our simulation studies, r is selected by minimizing the generalized cross-validation criterion for missing data (GCVM) defined as following:

$$\begin{aligned} GCVM(r) &= m \times \frac{\text{residual sum of squares for observed data}}{(\text{equivalent degrees of freedom})^2} \\ &= m \times \frac{\sum_{j=1}^{m} (z_j - \hat{z}_j)^2}{\big(m - (r+5)\big)^2}, \end{aligned}$$

where $(z_1, \cdots, z_m)$ are observed data ($m = n/k$) and $\hat{z}_j$ is the estimator of z_j. The equivalent degrees of freedom is $m - (r + 5)$, because the univariate cubic B-spline basis

function with r interior knots has $r + 4$ free parameters and β is one dimension in our examples. Here GCVM is modified from the definition of GCV in Mao and Zhao (2003). In their model, they used free-knot polynomials, that is both the knot locations and the regression coefficients are considered to be unknowns and to be estimated.

In the pilot study of the first simulation setting for $n = 1000$ with 200 replications, we found that MASE and GCVM are minimized at almost the same number of knots and MSE is not affected much by the number of knots in 1:50 as shown in Figure 2.1.

Insert Figure 2.1 here.

Therefore, in the following simulation studies, we will use GCVM to select the number of the knots from {2(2)20} to reduce the computational burden. We use GCVM for our proposed method, then use the same number of nknots for the complete data observed case.

Table 2.1 represents the point estimator of β, $MSE(\hat{\beta})$, $AVAR(\hat{\beta})$, $MASE(\hat{g})$ and $\hat{\sigma}$ for sample sizes $n = 500$ and $n = 1000$ from the first simulation study. The 'proposed' is the results using our proposed algorithm with missing data and 'com. lik.' is the results from maximizing the complete likelihood function.

Insert Table 2.1 here.

From Table 2.1, we notice that our proposed method performs pretty well with only 1/5 response values. The point estimators of β with missing data are the same as the true value 4 with both sample sizes. The $MSE(\hat{\beta})$ and $MASE(\hat{g})$ from the proposed method are all very small. Furthermore, MSE and $MASE$ reduce to about half as sample size doubles. The estimators of σ are all close to the true value 0.25.

Figure 2.2 shows the box-plots of the estimators of β from group sum observed case for $n = 500$ (left panel) and $n = 1000$ (right panel).

Insert Figure 2.2 here.

In Figure 2.2, for sample sizes $n = 500$ and $n = 1000$, both the proposed method with missing data and complete data have the median of the estimators of β around the

true value. When sample size increases, the interquartile of the estimators decreases a lot for both the proposed method with missing data and the complete data. In addition, when $n = 1000$,the proposed method with missing data and the complete data have the similar interquartile.

Figure 2.3 shows the point-wise average of the estimators of $g(\cdot)$ for group sum observed case for $n = 500$ (left panel) and $n = 1000$ (right panel).

Insert Figure 2.3 here.

In Figure 2.3, blue dotted curve presents true $g(\cdot)$, red solid curve is the point-wise average of the estimators of $g(\cdot)$ by proposed algorithm and green dashed curve is the point-wise average of the estimators of $g(\cdot)$ for complete data. The average curve from proposed algorithm almost overlaps with the true curve in the whole support of v, so does the average curve from complete data.

Figure 2.4 shows the point-wise variance of the estimators of $g(\cdot)$ from proposed algorithm (red solid curve) and from complete data (green dashed curve) for $n = 500$ (left panel) and $n = 1000$ (right panel).

Insert Figure 2.4 here.

The point-wise variance is large in the margin of v, however, in the middle of the support of v, the variances are very small for proposed algorithm with missing data. The variances decrease as the sample size increases, especially for the margin of v.

Table 2.2 shows the point estimator of β, $MSE(\hat{\beta})$, $AVAR(\hat{\beta})$, $MASE(\hat{g})$ and $\hat{\sigma}$ for sample sizes $n = 500$ and $n = 1000$ from the second simulation study (the estimator of asymptotic variance is given in Appendix 2.6.3). The 'proposed' represents the result using our proposed algorithm with missing data and 'com. lik.' represents the result from maximizing the complete likelihood function.

Insert Table 2.2 here.

Table 2.2 shows that in the case of group maximum observed, our method still works well. The point estimators of β are close to the true value 4 for both sample

sizes. The MSE and $MASE$ are all very small and decrease significantly as the sample size increases. In addition, the estimators of σ are all close to the true value 0.25.

Figure 2.5 shows the box-plot of the estimators of β from group maximum observed case for $n = 500$ (left panel) and $n = 1000$ (right panel).

Insert Figure 2.5 here.

In 2.5, the proposed method with missing data and the complete data have the median of the estimators both around the true value 4. When sample size increases, the interquartile decreases and the proposed method with missing data has similar interquartile with the complete data when $n = 1000$.

Insert Figure 2.5 here.

Figure 2.6 shows the point-wise average of the estimators of $g(\cdot)$ from proposed algorithm (red solid curve) and from complete data (green dashed curve) for $n = 500$ (left panel) and $n = 1000$ (right panel).

Insert Figure 2.6 here.

Figure 2.7 displays the point-wise variance of the estimators of $g(\cdot)$ from proposed algorithm (red solid curve) and from complete data (green dashed curve) for $n = 500$ (left panel) and $n = 1000$ (right panel).

Insert Figure 2.7 here.

In the group maximum observed case, the point-wise average curves and the true curve almost overlap and the point-wise variance curves have similar trend with the group sum observed case.

In conclusion, simulation studies demonstrate that our proposed estimation algorithm perform well in these settings.

2.5 DISCUSSION

In this chapter, we discussed about estimation method and algorithm for partially linear regression model with missing response variables. The missing pattern we considered has a general meaning and includes the case when each response is missing by certain probability.

For the estimation method, we applied the approach to maximizing the generalized profile likelihood function and showed that the estimator of parametric part, which maximizes the generalized profile likelihood, is a consistent and semiparametric efficient estimator under some conditions. In addition, we proposed an iterative algorithm to obtain the estimators. The algorithm runs iteratively between two modules, one of which uses EM algorithm and estimating equation to get the estimator of parametric component; the other uses EM algorithm and smoothing methods to obtain the estimator of nonparametric component. Simulation studies were performed to illustrate the proposed methodology and the simulation results showed that our algorithm works well in finite sample cases.

2.6 APPENDICES

2.6.1 Assumptions

The conditions $C1-C10$ are given in the following. Some of the conditions are regularity conditions and some of them are conditions needed because of the potential dependency of the observed data and semiparametric model.

The conditions $C1-C7$ are similar to the efficiency conditions in Bar-Shalom (1971) for parametric models with dependent response variables and the condition $C8-C10$ are extra conditions for semiparametric model.

The joint probability density function of $\mathbf{z} = (z_1, \cdots, z_m)$ can be written as

$$\begin{aligned} & f\big(z_1, \cdots, z_m; \beta, g_\beta(\cdot), \mathbf{w}, v_1, \cdots, v_n\big) \\ = \; & f\big(z_1|\beta, g_\beta(\cdot), \mathbf{w}, t_1\big) f\big(z_2|z_1, \beta, g_\beta(\cdot), \mathbf{w}, t_2\big) \cdots f\big(z_m|z_{m-1}, \cdots, z_1, \beta, g_\beta(\cdot), \mathbf{w}, t_m\big) \\ = \; & \prod_{k=1}^{m} f_k(\beta, g_\beta(\cdot), t_k), \end{aligned}$$

where $f_k\big(\beta, g_\beta(\cdot), t_k\big) = f\big(z_k|z_{k-1}, \cdots, z_1, \beta, g_\beta(\cdot), \mathbf{w}, t_k\big)$. Here vector t_k is a subset of $(v_1, \cdots, v_n)$ and contains all $v_i's$ involved in the probability density function $f_k\big(\beta, g_\beta(\cdot), t_k\big)$. We also assume that any k, $k = 1, \cdots, m$, there exists a constant $K < \infty$, such that $c(k) < K$, where $c(k)$ is the length of vector t_k and K does not depend on m. For simplicity, we use the notation, $l\big(\beta, g_\beta(\cdot), t_k\big) = \log f_k\big(\beta, g_\beta(\cdot), t_k\big)$.

$C1$.

Assume that for all $r, u = 0, \cdots, 4$, $r + u \leq 4$, the derivative

$$\frac{\partial^{r+u} l\big(\beta, g_\beta(\cdot), t_k\big)}{\partial \beta^r \partial g_\beta(t_{ks})^u}, s = 1, \cdots, c(k),$$

exists for almost all $\mathbf{z} = (z_1, \cdots, z_m)$ and

$$E_0\Big\{ \sup_{\beta \in B} \sup_{g_\beta(t_{ks}) \in R} \left| \frac{\partial^{r+u} l\big(\beta, g_\beta(\cdot), t_k\big)}{\partial \beta^r \partial g_\beta(t_{k,s})^u} \right|^2 \Big\} < \infty.$$

$C2$.

$$E_0\Big[\frac{dl(\beta, g_\beta(\cdot), t_k)}{d\beta}\Big]\Big|_{\beta=\beta_0} = 0$$

$C3$.

$$i_k(\beta_0) = E_0\Big[\frac{dl(\beta, g_\beta(\cdot), t_k)}{d\beta}\Big]^2\Big|_{\beta=\beta_0} \leq C_1 < \infty,$$

where C_1 is independent of k and $i_k(\beta_0)$ is the information in $l(\beta, g_\beta(\cdot), t_k)$. In addition, $i_\beta \equiv \lim_{m\to\infty} \frac{1}{m} \sum_{k=1}^{m} i_k(\beta_0)$ exists and i_β^{-1} is the semiparametric efficiency bound.

$C4$.

$$E_0\Big[\frac{d^2 l\big(\beta, g_\beta(\cdot), t_k\big)}{d\beta^2}\Big]\Big|_{\beta=\beta_0} = -i_k(\beta_0)$$

$C5$.

There exists a (μ^k)-measurable function $H_k(z^k)$ such that

$$\left|\frac{d^3 l\big(\beta, g_\beta(\cdot), t_k\big)}{d\beta^3}\right| < H_k(z_k), \forall \beta \in B,$$

and $H_k(z^k)$ is finite except on a set of probability zero, i.e.

$$\forall \epsilon > 0, \exists A < \infty, P\{H_k > A\} < \epsilon,$$

where A is independent of β and k.

Conditions $C6$ and $C7$ are required because of the potential dependency of $(z_1, \cdots, z_m)$.

$C6$.

$$E\Big[\frac{dl\big(\beta, g_\beta(\cdot), t_j\big)}{d\beta}\frac{dl\big(\beta, g_\beta(\cdot), t_k\big)}{d\beta}\Big]\Big|_{\beta=\beta_0} = 0, \forall j \neq k.$$

$C7$.

$$Var\Big[\frac{d^2 l\big(\beta, g_\beta(\cdot), t_k\big)}{d\beta^2}\Big]\Big|_{\beta=\beta_0} \leq C_2 < \infty,$$

where C_2 is independent of k and

$$\lim_{|k-j|\to\infty} Cov\Big[\frac{d^2 l\big(\beta, g_\beta(\cdot), t_j\big)}{d\beta^2}, \frac{d^2 l\big(\beta, g_\beta(\cdot), t_k\big)}{d\beta^2}\Big]\Big|_{\beta=\beta_0} = 0.$$

$C8$.

The derivative $g_0'(\cdot)$ satisfies

$$\begin{aligned} & \lambda_k' E\Big[\frac{\partial l\big(\beta, g_\beta(\cdot), t_k\big)}{\partial g(t_k)^T}\frac{\partial l\big(\beta, g_\beta(\cdot), t_k\big)}{\partial g(t_k)}\Big|t_k\Big]\Big|_{\beta=\beta_0} \\ = \quad & -E\Big[\frac{\partial l\big(\beta, g_\beta(\cdot), t_k\big)}{\partial \beta}\frac{\partial l\big(\beta, g_\beta(\cdot), t_k\big)}{\partial g(t_k)}\Big|t_k\Big]\Big|_{\beta=\beta_0}, \end{aligned}$$

where

$$\lambda_k' = \big(g_0'(t_{k,1}), \cdots, g_0'(t_{k,c(k)})\big),$$

and

$$\frac{\partial l\big(\beta, g_\beta(\cdot), t_k\big)}{\partial g(t_k)} = \Big(\frac{\partial l\big(\beta, g_\beta(\cdot), t_k\big)}{\partial g(t_{k,1})}, \cdots, \frac{\partial l\big(\beta, g_\beta(\cdot), t_k\big)}{\partial g(t_{k,c(k)})}\Big).$$

Here $g_0'(\cdot)$ is the least favorable direction.

$C9$.

$$(i)\frac{1}{\sqrt{m}}\sum_{k=1}^{m}\frac{d}{d\beta}\Big[\sum_{s=1}^{c(k)}\frac{\partial l\big(\beta,g_\beta(\cdot),t_k\big)}{\partial g(t_{k,s})}\big|_{\beta=\beta_0}\big(\hat{g}_0(t_{k,s})-g_0(t_{k,s})\big)\Big]=o_p(1),$$

$$(ii)\frac{1}{\sqrt{m}}\sum_{k=1}^{m}\Big[\sum_{s=1}^{c(k)}\frac{\partial l\big(\beta,g_\beta(\cdot),t_k\big)}{\partial g(t_{k,s})}\big|_{\beta=\beta_0}\big(\hat{g}_0'(t_{k,s})-g_0'(t_{k,s})\big)\Big]=o_p(1).$$

When there is no missing data, that is $z_i=y_i$ for $i=1,\cdots,n$, condition $C9$ corresponds to Lemma 2 in Severini and Wong (1992), which has been proved under the regularity conditions. For the block missing structure, where the observed data has independent and identical distribution, condition $C9$ can be proved following the same arguments for Lemma 2 under the regularity conditions. With more general missing patterns, it is not easy to simplify this condition.

$C10$.

$$(i)\sum_{k=1}^{m}\Big[l\big(\beta,\hat{g}_\beta(\cdot),t_k\big)-l\big(\beta,g_\beta(\cdot),t_k\big)\Big]=r_m^{(1)}(\beta),$$

where

$$\sup_\beta\left|m^{-1}\frac{d^2r_m^{(1)}(\beta)}{d\beta^2}\right|=o_p(1).$$

$$\begin{aligned}(ii)\sum_{k=1}^{m}l\big(\beta,\hat{g}_\beta(\cdot),t_k\big) &= \sum_{k=1}^{m}l\big(\beta,g_\beta(\cdot),t_k\big)+\sum_{k=1}^{m}\Big[\sum_{s=1}^{c(k)}\frac{\partial l\big(\beta,g_\beta(\cdot),t_k\big)}{\partial g(t_{k,s})}\big(\hat{g}_\beta'(t_{k,s})-g_\beta'(t_{k,s})\big)\Big]\\ &\quad +r_m^{(2)}(\beta)\end{aligned}$$

where

$$m^{-1/2}\frac{dr_m^{(2)}(\beta)}{d\beta}|_{\beta=\beta_0}=o_p(1).$$

This condition is the Lemma 3 in Severini and Wong (1992) without missing data and has been proved under regularity conditions. For the block missing structure, this condition can be proved following the same arguments for Lemma 3. With more general missing patterns, it is not trivial to simplify either.

2.6.2 Proofs

Proof of Theorem 1. The proof follows the proof of Proposition 1 in Severini and Wong (1992). Under the regularity conditions, $l\big(\beta,g_\beta(\cdot);\mathbf{z},\mathbf{w},\mathbf{v}\big)$ is continuous in β

and a measurable function of $\mathbf{z}$; $v_1, \cdots, v_n$ for each β. Therefore, it follows that $\hat{\beta}$ is measurable.

By Conditions I,

$$m^{-1} l\big(\beta, g_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) \to_p l_0(\beta) \quad \text{for each } \beta \in B,$$

furthermore, for $\beta_1, \beta_2 \in B$,

$$\begin{aligned}
& m^{-1} \big| l\big(\beta_1, g_{\beta_1}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) - l\big(\beta_2, g_{\beta_2}(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) \big| \\
&= m^{-1} \big| \log f\big(\mathbf{z}; \mathbf{w}, \mathbf{v}, \beta_1, g_{\beta_1}(\cdot)\big) - \log f\big(\mathbf{z}; \mathbf{w}, \mathbf{v}, \beta_2, g_{\beta_2}(\cdot)\big) \big| \\
&\le A_m \left|\beta_1 - \beta_2\right| + B_m \left\| g_{\beta_1} - g_{\beta_2} \right\| \\
&\le A_m \left|\beta_1 - \beta_2\right| + B_m \sup_\beta \left\| g'_\beta \right\| \left|\beta_1 - \beta_2\right| \\
&\equiv C_m |\beta_1 - \beta_2|.
\end{aligned}$$

Since by Conditions C, C_m is bounded in probability, it follows that

$$\big\{ m^{-1} l\big(\beta, g_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) : \beta \in B \big\}$$

is tight and hence,

$$m^{-1} l\big(\beta, g_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) \to_{\mathscr{D}} l_0(\beta) \quad \text{in } C(B).$$

For each β, by Conditions C,

$$\begin{aligned}
& m^{-1} \big| l\big(\beta, \hat{g}_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) - l\big(\beta, g_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) \big| \\
&= m^{-1} \big| \log f\big(\mathbf{z}; \mathbf{w}, \mathbf{v}, \beta, \hat{g}_\beta(\cdot)\big) - \log f\big(\mathbf{z}; \mathbf{w}, \mathbf{v}, \beta, g_\beta(\cdot)\big) \big| \\
&\le B_m \sup_\beta \left\| \hat{g}_\beta - g_\beta \right\|.
\end{aligned}$$

Therefore, by Condition NP

$$\sup_\beta \frac{1}{m} \big| l\big(\beta, \hat{g}_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) - l\big(\beta, g_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) \big| \to_p 0 \quad \text{as } m \to \infty$$

and hence,

$$\sup_\beta \big| \frac{1}{m} l\big(\beta, \hat{g}_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) - l_0(\beta) \big| \to_p 0 \quad \text{as } m \to \infty.$$

Furthermore, since (suppose that $l_0(\beta)$ is unimodal)

$$\sup_{\beta} \frac{1}{m} l\big(\beta, \hat{g}_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v}\big) \to_p \sup_{\beta} l_0(\beta) = l_0(\beta_0),$$

it follows that

$$l_0(\hat{\beta}) \to_p l_0(\beta_0) \quad \text{as } m \to \infty.$$

For a given $\beta \in B$, there exists an $\epsilon > 0$ and an open neighborhood N_β of β such that

$$\inf_{\beta_1 \in N_\beta} |l_0(\beta_1) - l_0(\beta_0)| > \epsilon.$$

Therefore,

$$P_0(\hat{\beta} \in N_\beta) \leq P_0\Big(\Big| l_0(\hat{\beta}) - l_0(\beta_0) \Big| > \epsilon \Big) \to 0 \quad \text{as } m \to \infty.$$

Let N_0 denote an open neighborhood of β_0 and consider the compact set $B_0 = B \backslash N_0$. Let $\{N_\beta : \beta \in B, \beta \neq \beta_0\}$ denote the open cover of B_0 constructed by the preceding procedure. By compactness of B_0 there exists a finite subcover $\{N_{\beta_1}, \cdots, N_{\beta_k}\}$. Then

$$P_0(\hat{\beta} \notin N_0) = P_0(\hat{\beta} \in B_0) \leq \sum_{j=1}^{k} P_0(\hat{\beta} \in N_{\beta_j}) \to 0 \quad \text{as } n \to \infty.$$

Therefore,

$$\hat{\beta} \to_p \beta_0 \quad \text{as } m \to \infty.$$

Proof of Theorem 2. As given in the Section 2.6.1, the joint probability density function of $\mathbf{z} = (z_1, \cdots, z_m)$ can be written as

$$\begin{aligned} & f\big(z_1, \cdots, z_m | \beta, g_\beta(\cdot), \mathbf{w}, v_1, \cdots, v_n\big) \\ = & \prod_{k=1}^{m} f_k\big(\beta, g_\beta(\cdot), t_k\big), \end{aligned}$$

where $f_k\big(\beta, g_\beta(\cdot), t_k\big) = f\big(z_k | z_{k-1}, \cdots, z_1, \beta, g_\beta(\cdot), \mathbf{w}, t_k\big)$

and denote $l\big(\beta, g_\beta(\cdot), t_k\big) = \log f_k\big(\beta, g_\beta(\cdot), t_k\big)$.

Using a Taylor's expansion,

$$\begin{aligned} 0 &= \frac{d \sum_{k=1}^{m} l\big(\beta, \hat{g}_\beta(\cdot), t_k\big)}{d\beta}\Big|_{\beta=\hat{\beta}} \\ &= \frac{d \sum_{k=1}^{m} l\big(\beta, \hat{g}_\beta(\cdot), t_k\big)}{d\beta}\Big|_{\beta=\beta_0} + \frac{d^2 \sum_{k=1}^{m} l\big(\beta, \hat{g}_\beta(\cdot), t_k\big)}{d\beta^2}\Big|_{\beta=\hat{\beta}^*} (\hat{\beta} - \beta_0), \end{aligned}$$

where $\hat{\beta}^*$ lies between β_0 and $\hat{\beta}$ and by Theorem 1, $\hat{\beta}^* \to_p \beta_0$. Hence,

$$\sqrt{m}(\hat{\beta}-\beta_0) = -\frac{(1/\sqrt{m})\Big(d\sum_{k=1}^m l\big(\beta,\hat{g}_\beta(\cdot),t_k\big)/d\beta|_{\beta=\beta_0}\Big)}{(1/m)\Big(d^2\sum_{k=1}^m l\big(\beta,\hat{g}_\beta(\cdot),t_k\big)/d\beta^2|_{\beta=\hat{\beta}^*}\Big)}. \tag{2.5}$$

Conditions $C9$ and $C10$ imply the following two equations:

$$\frac{1}{\sqrt{m}}\frac{d\sum_{k=1}^m l\big(\beta,\hat{g}_\beta(\cdot),t_k\big)}{d\beta}\Big|_{\beta=\beta_0} = \frac{1}{\sqrt{m}}\frac{d\sum_{k=1}^m l\big(\beta,g_\beta(\cdot),t_k\big)}{d\beta}\Big|_{\beta=\beta_0} + o_p(1), \tag{2.6}$$

and

$$\sup_\beta\left|\frac{1}{m}\frac{d^2\sum_{k=1}^m l\big(\beta,\hat{g}_\beta(\cdot),t_k\big)}{d\beta^2} - \frac{1}{m}\frac{d^2\sum_{k=1}^m l\big(\beta,g_\beta(\cdot),t_k\big)}{d\beta^2}\right| = o_p(1). \tag{2.7}$$

Then with equations (2.6) and (2.7) and by the conditions $C1-C8$, we have

$$E\big(\sqrt{m}(\hat{\beta}-\beta_0)\big)^2 = \Big(\frac{1}{m}\sum_{k=1}^m i_k(\beta_0)\Big)^{-1} \to i_\beta^{-1} \quad \text{as} \quad m\to\infty,$$

therefore, the estimator is semiparametric efficient.

The result

$$\hat{i}_\beta \to_p i_\beta \quad \text{as } m\to\infty,$$

follows from (2.7) and Theorem 1.

Proof of Theorem 3. By Ahmad et al. (2005), $\hat{\beta}^*$ and $\hat{\mathbf{g}}^*$ are given by

$$\begin{cases} \hat{\beta}^* = \big[(\mathbf{w}-M\mathbf{w})^T(\mathbf{w}-M\mathbf{w})\big]^-(\mathbf{w}-M\mathbf{w})^T(\mathbf{y}-M\mathbf{y}) \\ \hat{\mathbf{g}}^* = M(\mathbf{y}-\mathbf{w}\hat{\beta}^*). \end{cases}$$

For any fixed β, the corresponding estimator of nonparametric component from the iterative algorithm satisfies the following equation:

$$\begin{aligned} \hat{\mathbf{g}}_\beta &= ME(\mathbf{y}|\mathbf{z},\mathbf{w},\beta,\hat{\mathbf{g}}_\beta) - M\mathbf{w}\beta \\ &\equiv M\mathbf{E1} - M\mathbf{w}\beta, \end{aligned}$$

and its derivative to β is

$$\begin{aligned} \hat{\mathbf{g}}'_\beta &= M\frac{\partial E(\mathbf{y}|\mathbf{z},\mathbf{w},\beta,\hat{\mathbf{g}}_\beta)}{\partial\beta} - M\mathbf{w} \\ &\equiv M\mathbf{E2} - M\mathbf{w}, \end{aligned}$$

where $\mathbf{E1} = E(\mathbf{y}|\mathbf{z}, \mathbf{w}, \beta, \hat{\mathbf{g}}_\beta)$ and $\mathbf{E2} = \partial E(\mathbf{y}|\mathbf{z}, \mathbf{w}, \beta, \hat{\mathbf{g}}_\beta)/\partial\beta$. The estimator $\hat{\beta}$ is the solution of

$$(\mathbf{E1} - \mathbf{w}\beta - \hat{\mathbf{g}}_\beta)^T(\mathbf{w} + \hat{\mathbf{g}}'_\beta) = 0,$$

that is

$$\begin{aligned} 0 &= \big[(\mathbf{E1} - M\mathbf{E1}) - (\mathbf{w} - M\mathbf{w})\beta\big]^T(\mathbf{w} - M\mathbf{w} + M\mathbf{E2}) \\ &= (\mathbf{E1} - M\mathbf{E1})^T M\mathbf{E2} - \beta^T(\mathbf{w} - M\mathbf{w})^T M\mathbf{E2} \\ &\quad +(\mathbf{E1} - M\mathbf{E1})^T(\mathbf{w} - M\mathbf{w}) - \beta^T(\mathbf{w} - M\mathbf{w})^T(\mathbf{w} - M\mathbf{w}) \\ &= (\mathbf{E1} - M\mathbf{E1})^T(\mathbf{w} - M\mathbf{w}) - \beta^T(\mathbf{w} - M\mathbf{w})^T(\mathbf{w} - M\mathbf{w}). \end{aligned}$$

The third equation is from $M^T = M$ and $M^T M = M$. Therefore,

$$\begin{aligned} \beta &= \big[(\mathbf{w} - M\mathbf{w})^T(\mathbf{w} - M\mathbf{w})\big]^{-1}\big[(\mathbf{w} - M\mathbf{w})^T(\mathbf{E1} - M\mathbf{E1})\big] \\ &= E[\hat{\beta}^*|\mathbf{z}, \mathbf{w}, \beta, \hat{\mathbf{g}}_\beta], \end{aligned}$$

and

$$\begin{aligned} \hat{\mathbf{g}}_\beta &= M\mathbf{E1} - M\mathbf{w}\beta \\ &= E(M\mathbf{y}|\mathbf{z}, \mathbf{w}, \beta, \hat{\mathbf{g}}_\beta) - E(M\mathbf{w}\hat{\beta}^*|\mathbf{z}, \beta, \hat{\mathbf{g}}_\beta) \\ &= E[\hat{\mathbf{g}}^*|\mathbf{z}, \beta, \hat{\mathbf{g}}_\beta]. \end{aligned}$$

Hence, the estimators from iterative algorithm are the solution of

$$\begin{cases} \beta = E[\hat{\beta}^*|\mathbf{z}, \mathbf{w}, \beta, \mathbf{g}] \\ \mathbf{g} = E[\hat{\mathbf{g}}^*|\mathbf{z}, \mathbf{w}, \beta, \mathbf{g}]. \end{cases}$$

2.6.3 Estimator of asymptotic variance in group maximum observed case

In group maximum observed case, $E\big(y_i y_j|\mathbf{z}, \mathbf{w}, \mathbf{v}, \beta, g_\beta(\cdot)\big)$ is not easy to calculate. So we need to use another way to estimate the covariance matrix part. Suppose β is length

1. Notice that, with $m \to \infty$,

$$\begin{aligned}
&\frac{1}{m}Var\Big[\frac{dl\big(\beta, g_\beta(\cdot); \mathbf{y}, \mathbf{w}, \mathbf{v}\big)}{d\beta}\Big|\mathbf{z}\Big] \to E\Big\{\frac{1}{m}Var\Big[\frac{dl\big(\beta, g_\beta(\cdot); \mathbf{y}, \mathbf{w}, \mathbf{v}\big)}{d\beta}\Big|\mathbf{z}\Big]\Big\} \\
&= \frac{1}{m}\Big\{Var\Big[\frac{dl\big(\beta, g_\beta(\cdot)|\mathbf{y}, \mathbf{w}, \mathbf{v}\big)}{d\beta}\Big] - Var\Big[E\Big(\frac{dl\big(\beta, g_\beta(\cdot)|\mathbf{y}, \mathbf{w}, \mathbf{v}\big)}{d\beta}|\mathbf{z}\Big)\Big]\Big\} \\
&= \frac{1}{m}(II1 - II2),
\end{aligned}$$

where

$$\begin{aligned}
II1 &= Var\big[\frac{1}{\sigma^2}\big(\mathbf{y} - \mathbf{w}\beta - g_\beta(\mathbf{v})\big)^T\big(\mathbf{w} + g'_\beta(\mathbf{v})\big)\big] \\
&= \frac{1}{\sigma^2}\big(\mathbf{w} + g'_\beta(\mathbf{v})\big)^T\big(\mathbf{w} + g'_\beta(\mathbf{v})\big),
\end{aligned}$$

and

$$\begin{aligned}
II2 &= Var\Big[E\Big(\frac{\big(\mathbf{y} - \mathbf{w}\beta - g_\beta(\mathbf{v})\big)^T\big(\mathbf{w} + g'_\beta(\mathbf{v})\big)}{\sigma^2}\Big|\mathbf{z}\Big)\Big] \\
&= \frac{1}{\sigma^4}Var\Big[E\Big(\sum_{i=1}^{n}\big(y_i - w_i^T\beta - g(v_i)\big)\big(w_i + g'(v_i)\big)|\mathbf{z}\Big) \\
&\qquad , E\Big(\sum_{i=1}^{n}\big(y_i - w_i^T\beta - g(v_i)\big)\big(w_i + g'(v_i)\big)|\mathbf{z}\Big)\Big].
\end{aligned}$$

In group maximum observed case, $II2$ can be written as

$$\begin{aligned}
II2 &= \frac{1}{\sigma^4}Var\Big[\sum_{j=1}^{m}E(s_j|\mathbf{z}), \sum_{j=1}^{m}E(s_j|\mathbf{z})\Big] = \frac{1}{\sigma^4}\sum_{i=1}^{m}Var\big[E(s_j|\mathbf{z}), E(s_j|\mathbf{z})\big] \\
&= \frac{1}{\sigma^4}\sum_{i=1}^{m}E\big[E(s_j|\mathbf{z})E(s_j|\mathbf{z})\big] \qquad (2.8) \\
&\approx \frac{1}{\sigma^4}\sum_{i=1}^{m}E(s_j|\mathbf{z})E(s_j|\mathbf{z}), \qquad (2.9)
\end{aligned}$$

in first equation

$$\begin{aligned}
E(s_j|\mathbf{z}) &= E\Big(\sum_{i\in\text{group } j}\big(y_i - w_i^T\beta - g(v_i)\big)\big(w_i + g'_\beta(v_i)\big)|\mathbf{z}, \mathbf{w}, \mathbf{v}\Big) \\
&= \sum_{i\in\text{group } j}\Big(E(y_i|\mathbf{z}, \mathbf{w}, \mathbf{v}) - w_i^T\beta - g(v_i)\Big)\big(w_i + g'_\beta(v_i)\big),
\end{aligned}$$

depends on the group j only and is independent by groups, so we have second equation. Third equation is from $E\big[E(s_j|\mathbf{z})\big] = 0$, for $j = 1, \cdots, m$. As $m \to \infty$, (2.8) can be

approximated by (2.9). The negative information matrix is

$$-\frac{1}{m}\frac{d^2 l(\beta, g_\beta(\cdot); \mathbf{z}, \mathbf{w}, \mathbf{v})}{d\beta^2} \approx \frac{1}{m} II2$$
$$\approx \frac{1}{m\sigma^4}\sum_{i=1}^{m}\big[E(s_j|\mathbf{z})E(s_j|\mathbf{z})\big]. \tag{2.10}$$

Then the asymptotic variance of $\hat{\beta}$ can be estimated by inverse of estimator of (2.10), where β, g, g' and σ are replaced by $\hat{\beta}$, $\hat{g}_{\hat{\beta}}$, $\hat{g}'_{\hat{\beta}}$ and $\hat{\sigma}$ respectively.

Figure 2.1: Pilot study for $n = 1000$ with 200 replications. Left panel is the GCVM values for the number of nknots in 1:50; right panel is the $MSE(\hat{\beta})$ (red solid curve) and $MASE(\hat{g})$ (green dashed curve).

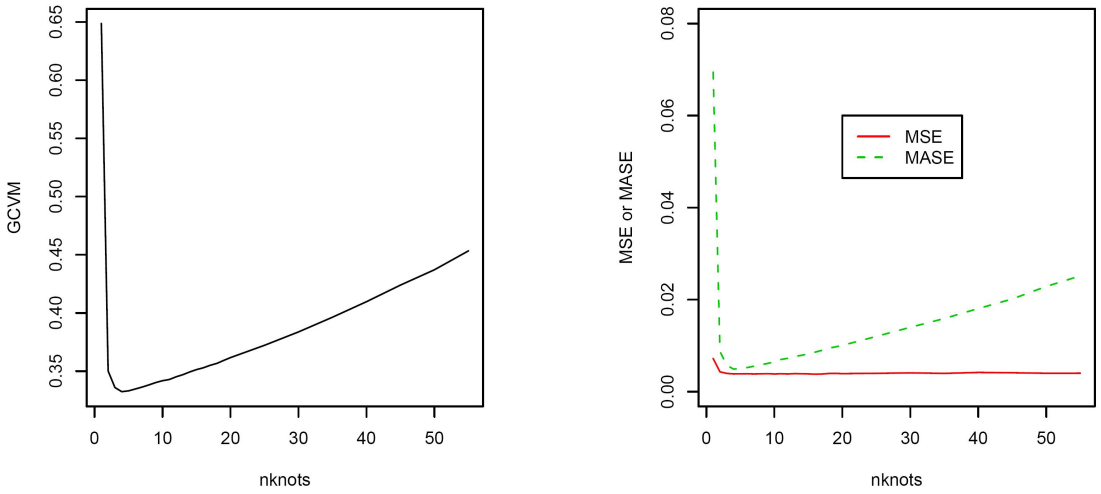


Table 2.1: Simulation results for estimators from group sum observed case based on 500 replications.

		$\hat{\beta}$ $(\beta = 4)$			$\hat{g}(\cdot)$	$\hat{\sigma}$
		$Mean$	MSE	$\hat{AVAR}$	$MASE$	(0.25)
$n = 500$	proposed	4.0	0.003	0.33	0.014	0.24
	com. lik.	4.0	0.001	0.34	0.002	0.25
$n = 1000$	proposed	4.0	0.001	0.33	0.008	0.24
	com. lik.	4.0	0.0003	0.34	0.001	0.25

Figure 2.2: Simulation results for group sum observed case: box-plot for estimators of β. Left panel is for $n = 500$ and right panel is for $n = 1000$.

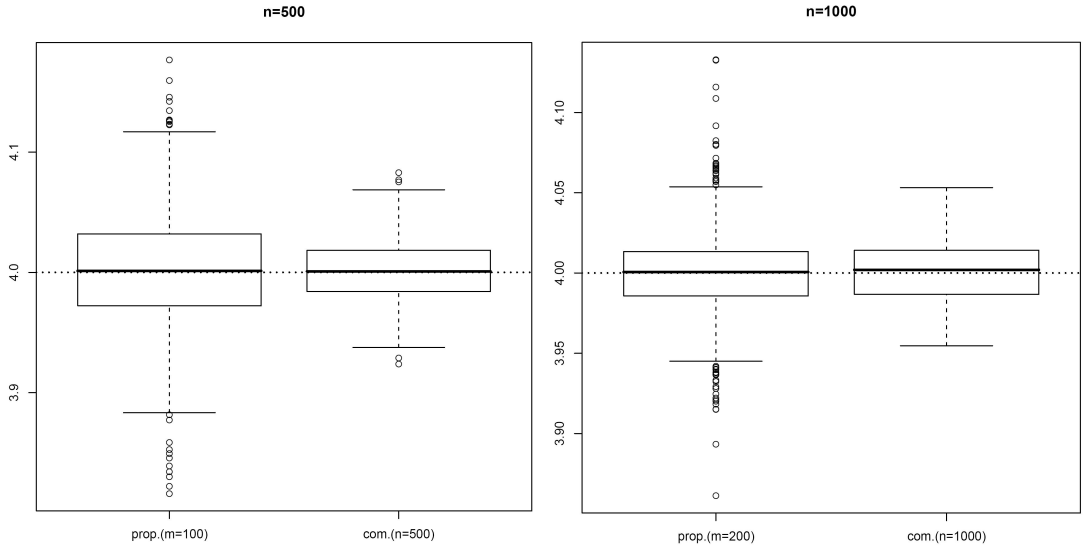


Figure 2.3: Simulation results for group sum observed case. The blue dotted curve presents true $g(\cdot)$, red solid curve is the point-wise average of the estimators of $g(\cdot)$ by proposed algorithm with missing data and green dashed curve is the point-wise average of the estimators of $g(\cdot)$ for complete data; left panel is for $n = 500$ and right panel is for $n = 1000$.

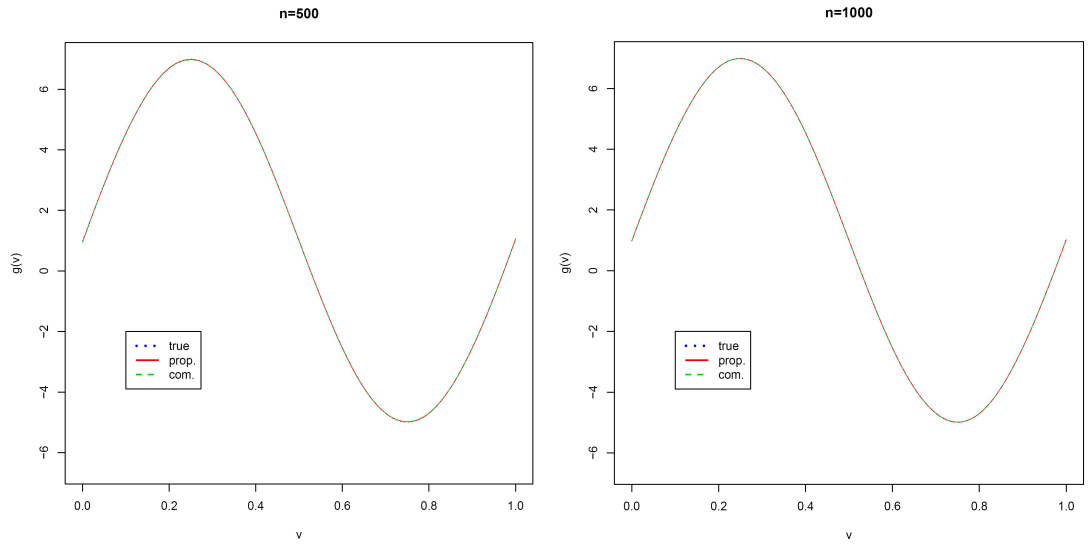


Figure 2.4: Simulation results for group sum observed case. The red solid curve is the point-wise variance of the estimators of $g(\cdot)$ by proposed algorithm with missing data and green dashed curve is the point-wise variance of the estimators of $g(\cdot)$ for complete data; left panel is for $n = 500$ and right panel is for $n = 1000$.

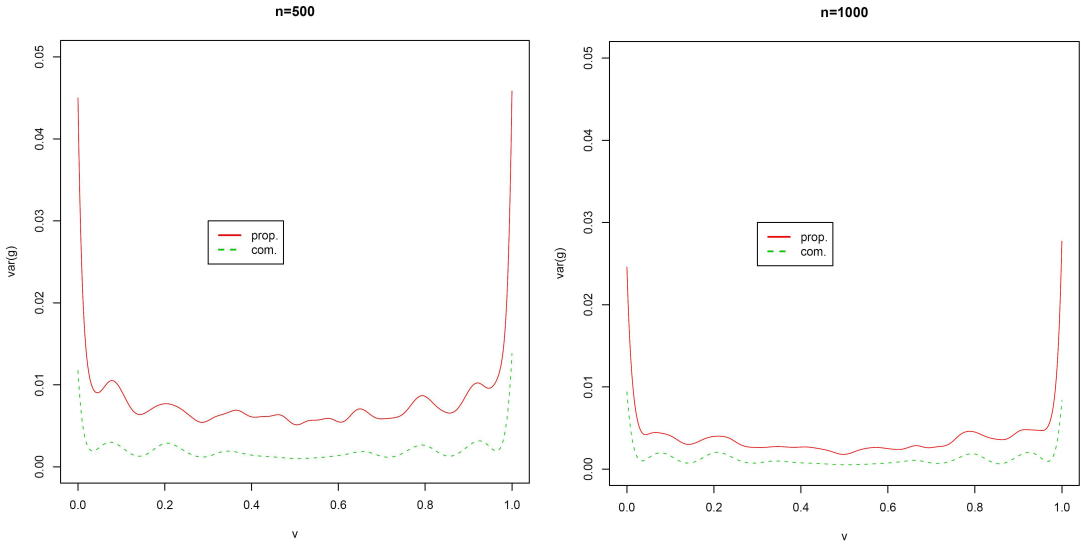


Table 2.2: Simulation results for estimators from group maximum observed case based on 500 replications.

		$\hat{\beta}$ $(\beta = 4)$			$\hat{g}(\cdot)$	$\hat{\sigma}$
		$Mean$	SE	$\hat{AVAR}$	$MASE$	(0.25)
$n = 500$	proposed	4.0	0.003	0.38	0.017	0.24
	com. lik.	4.0	0.001	0.34	0.003	0.25
$n = 1000$	proposed	4.0	0.001	0.36	0.012	0.25
	com. lik.	4.0	0.0004	0.34	0.002	0.25

Figure 2.5: Simulation results for group maximum observed case: box-plot for estimators of β. Left panel is for $n = 500$ and right panel is for $n = 1000$.

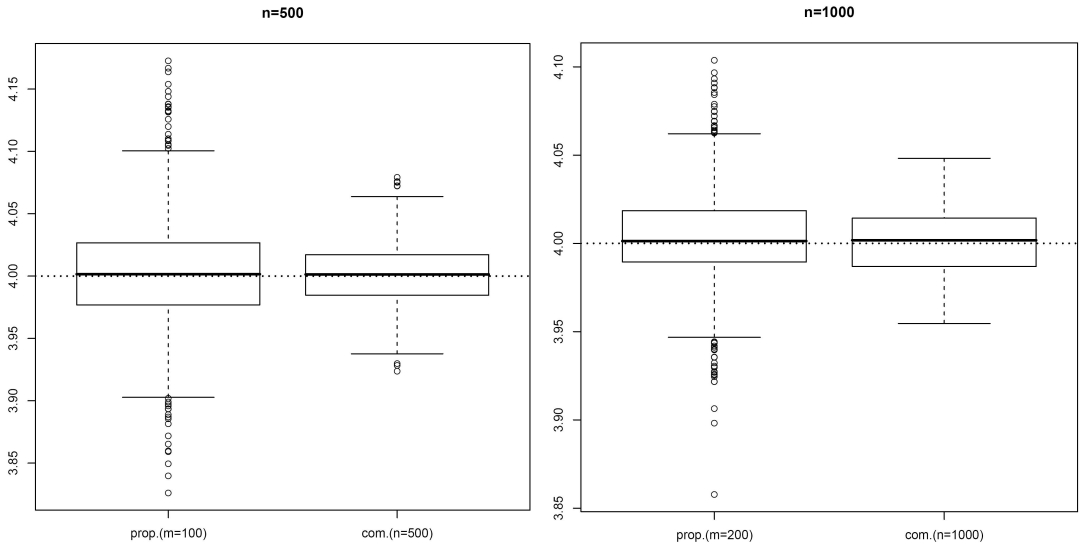


Figure 2.6: Simulation results for group maximum observed case. The blue dotted curve presents true $g(\cdot)$, red solid curve is the point-wise average of the estimators of $g(\cdot)$ by proposed algorithm and green dashed curve is the point-wise average of the estimators of $g(\cdot)$ for complete data; left panel is for $n = 500$ and right panel is for $n = 1000$.

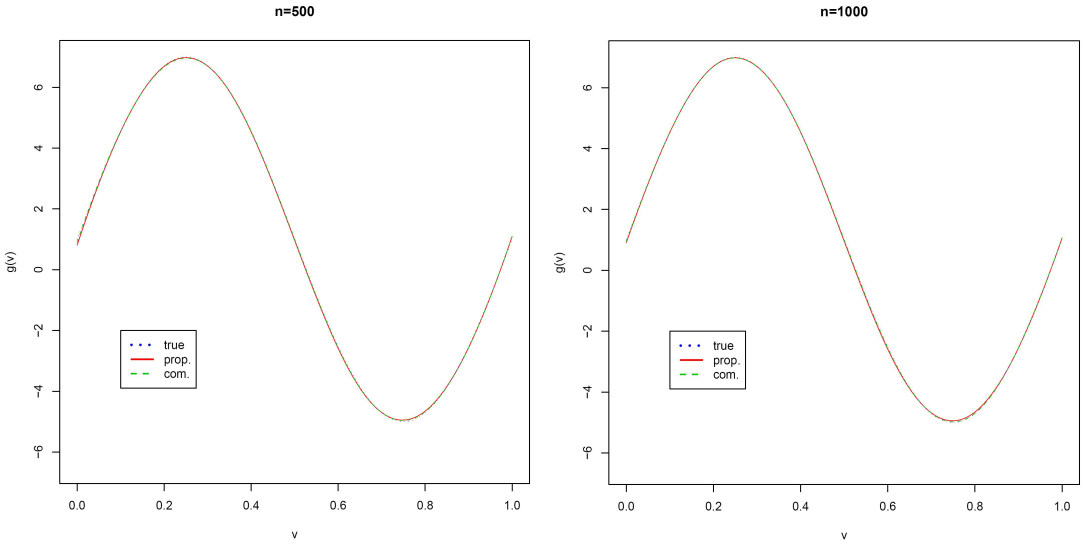
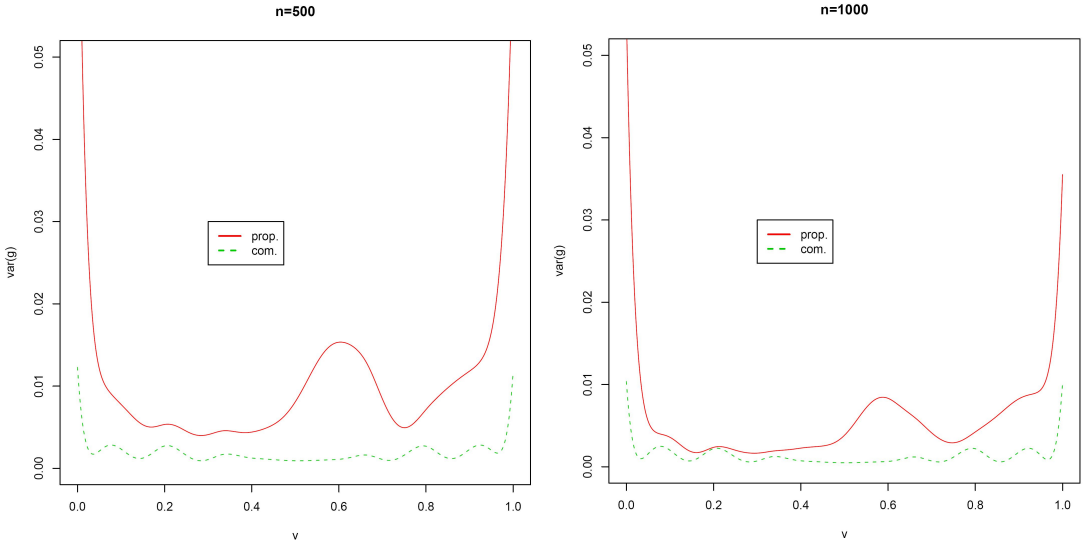


Figure 2.7: Simulation results for group maximum observed case. The red solid curve is the point-wise variance of the estimators of $g(\cdot)$ by proposed algorithm and green dashed curve is the point-wise variance of the estimators of $g(\cdot)$ for complete data; left panel is for $n = 500$ and right panel is for $n = 1000$.



References

[1] Ahmad, I., Leelahanon, S. and Li, Q. (2005), "Efficient Estimation of a Semiparametric Partially Linear Varying Coefficient Model," *The Annals of Statistics*, 33, 258-283.

[2] Bar-Shalom (1971), "On the Asymptotic Properties of the Maximum-Likelihood Estimate Obtained from Dependent Observations," *Journal of the Royal Statistical Society. Series B*, 33 No. 1, 72-77.

[3] Boente, G.,He, X. and Zhou, J. (2006), "Robust Estimates in Generalized Partially Linear Models," *The Annals of Statistics*, 34 No. 6, 2856-2878.

[4] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 No. 1, 1-38.

[5] Engle, R.F.,Granger, C.W.J, Rice, J. and Weiss, A. (1986), "Semiparametric Estimates of the Relation between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 310-320.

[6] Green, P.J. (1990), "On Use of the EM for Penalized Likelihood Estimation," *Journal of the Royal Statistical Society B*, 52 No.3, 443-452.

[7] Lam, C., Fan, J. (2008), "Profile-Kernel Likelihood Inference With Diverging Number of Parameters," *The Annals of Statistics*, 36 No.5, 2232-2260.

[8] Little, R.J.A. and Rubin, D.B (2002), *Statistical Analysis with missing data*, J.Wiley. New York, 2nd ed.

[9] Ma, Y., Chiou, J.-M., Wang, N. (2006), "Efficient semiparametric estimator heteroscedastic partially linear models," *Biometrika*, 93 No.1, 75-84.

[10] Mao, W., Zhao, L.H. (2003), "Free-knot polynomial splines with confidence intervals," *J. R. Statist. Soc. B* 65, Part 4, 901-919.

[11] Newey, W.K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99-135.

[12] Severini, T.A. and Wong, W.H. (1992), "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20, 1768-1802.

[13] Silverman, B.W., Jones, M.C., Wilson, J.D. and Nychka, D.W. (1990), "A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion)," *J.R.Statist.Soc.* B, 52, 271-324.

[14] Stein, C. (1956), "Efficient nonparametric testing and estimation," *Proc. Third Berkeley Symp. Math. Statist. Probab*, 1, 187-195.

[15] Wang, Q., Linton, O. and Härdle, W. (2004) "Semiparametric Regression Analysis With Missing Response at Random," *Journal of the American Statistical Association*, 99 No. 466, 334-345.

[16] Wu, C.F.J (1983), "On the convergence properties of the EM algorithm," *The Annals of Statistics*, 11 No.1, 95-103.

[17] Xie, M. (2001), "Regression Analysis of Group Testing Samples," *Statistics in Medicine*, 20, 1957-1969.

[18] Xie, M., Simpson, D.G. and Carroll, R.J (2008), "Semiparametric analysis of heterogeneous data using varying scale glm," *Journal of the American Statistical Association*, 103 No. 482, 650-660.

Vita

Mingyu Li

- Education

  2004-2010 Ph.D. in Statistics, Rutgers University

  2000-2004 B.S in Statistics, University of Science and Technology of China

- Professional Experiences

  10/2008-08/2009 Intern, sanofi-aventis, Bridgewater, NJ

  06/2007-08/2007 Intern, Eisai Medical Research Inc., Ridgefield Park, NJ

  2004-2008 Fellowship and Assistantship, Rutgers University

- Publications

  - Quan, H., **Li, M.**, Shih, W. and Jiang, K. (2009) Comparisons of procedures for two-stage adaptive design in clinical trials. *Sanofi-aventis Technical Report #028.*
  - The consistency sub-stream of PhRMA MRCT cross-functional KIT. (2009) Assessment of consistency of treatment effects in multi-regional clinical trials. *Sanofi-aventis Technical Report #034 and Submitted to DIJ.*
  - Quan, H., **Li, M.**, Zhao, P., Cho, M., Zhang, J. and Wu, Y. (2009) Considerations for design and data analysis of adaptive superiority/non-inferiority cardiovascular trials. *Sanofi-aventis Technical Report #037 and Submitted to Statistics in Medicine.*
  - Zhu, Y., **Li, M.**, Young, C.M., Xie, M. and Elsayed, E.A. (2009) Impact of measurement error on container inspection policies at port-of-entry. *Annals of Operations Research, Tentatively accepted.*
  - Young, C.M, **Li, M.**, Zhu, Y., Xie, M., Elsayed, E.A. and Asamov, T. (2009) Multiobjective optimization of a port-of-entry inspection policy. *IEEE Transactions-ASE, Accepted.*