Matrix-Variate Regression with Measurement Error

by

Junhan Fang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics (Biostatistics)

Waterloo, Ontario, Canada, 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Richard Lockhart Professor (Simon Fraser University)
Supervisor:	Grace Y. Yi Professor (University of Western Ontario)
Internal Member:	Pengfei Li Professor
	Leilei Zeng Associate Professor
Internal-External Member:	Fue-Sang Lien Professor (Department of Mechanical and Mechatronics Engineering)

Author's Declaration

Abstract

Matrix-variate regression models are useful for featuring data with a matrix structure, such as brain imaging data. However, those methods do not apply to data with measurement error or misclasscification. While mismeasurement is an inevitable issue in the data collecting process, little research has been available to handle matrix-variate regression with mismeasurement. In this thesis, we explore several important problems concerning matrix-variate regression with error-contaminated data.

In Chapter 1, we provide a brief introduction for matrix-variate data and review relevant topics including logistic regression analysis, measurement error/misclassification mechanisms, regularization methods, and Bayesian inference procedures.

In Chapter 2, we discuss matrix-variate logistic regression for handling error-contaminated data. Measurement error in covariates has been extensively studied in many conventional regression settings where covariate information is typically expressed in a vector form. However, there has been little work on error-prone matrix-variate data which commonly arise from studies with imaging, spatial-temporal structures. We particularly focus on matrix-variate logistic measurement error models. We examine the biases induced from the naive analysis which ignores measurement error. Two measurement error correction methods are developed to adjust for measurement error effects. The proposed methods are justified both theoretically and empirically. We analyze a data set arising from a study examining electroencephalography(EEG) correlates of genetic predisposition to alcoholism with the proposed methods.

In Chapter 3, we consider a problem complement to that in Chapter 2. Instead of examining mismeasurement in covariates, here we study mismeasurement in binary responses. We particularly investigate the response misclassification effects on the matrix-variate logistic regression model. Matrix-variate logistic regression is useful in facilitating the relationship between the binary response and matrix-variates which arise commonly from medical imaging research. However, such a model is impaired by the presence of the response misclassification. It is imperative to account for misclassification effects when employing matrix-variate logistic regression to handle such data. In this chapter, we develop two inferential methods which account for misclassification effects. The first method is an imputation method which replaces the response variable with an observed and unbiased pseudo-response variable in the estimation procedure. The second method is derived from the likelihood function for the observed response surrogate. Our development is carried out for two settings where misclassification rates are either known or estimated from validation data. The proposed methods are justified both theoretically and empirically. We analyze the breast cancer Wisconsin prognostic data with the proposed methods.

Chapter 4 is a continuation and extension of Chapter 3. We consider regularized matrix-variate logistic regression with response misclassification, where matrix-variate data may assume a sparsity structure. With a limited sample size, the presence of a large number of redundant parameters entails the difficulty of estimation. In this chapter, we develop inferential methods which account for misclassification effects in combination with the inclusion of penalty functions to deal with the sparsity of matrix-variate data. We examine the biases induced from the naive analysis which ignores the response misclassification. Our development is carried out for two settings where misclassification rates are either known or estimated from validation data. The proposed methods are justified both theoretically and empirically. We analyze the breast cancer Wisconsin prognostic data with the proposed methods.

In Chapter 5, we shift our attention to the Bayesian framework. We consider applying Bayesian analysis to matrix-variate logistic regression. We propose a Bayesian algorithm to estimate the matrix-variate parameters element-wisely in combination with the use of horse-shore shrinkage prior. We investigate the influence on parameter estimation when ignoring the response misclassification and propose an algorithm to accommodate the effects of response misclassification. The performance of the proposed method is evaluated through numerical studies. We analyze the Lee Silverman voice treatment (LSVT) Companion data with the proposed method.

Finally, Chapter 6 summarizes the thesis work and presents some future work.

Acknowledgements

I wish to acknowledge the support of various individuals and organizations. Firstly, I would like to express my deepest gratitude to my supervisor, Dr. Grace Y. Yi for her patience, encouragement, and support during my Ph.D. studies. Her guidance and support helped me overcome all difficulties throughout my time as her student, and to achieve the goals beyond my imagination at the beginning. It was a real honor for me to share her profound scientific knowledge and rigorous academic attitude.

Besides my supervisor, I wish to show my gratitude to my committee members Drs. Pengfei Li, Fue-Sang Lien, Richard Lockhart and Leilei Zeng for reviewing my thesis and their constructive suggestions. I would like to pay my regards to Drs. Wenqing He, Joseph Sedransk and Mary Thompson for their insightful comments, advice and help to enrich my research knowledge and experience.

I would like to thank all the members of the GW-DS (Grace-Wenqing Data Science) Research Group who have contributed a lot to my personal and professional time at Waterloo. Special thanks to my academic sisters and brothers: Li-Pang Chen, Di Shu, Qihuang Zhang and Haoxin Zhuang for giving me all their supports and joy through good and tough times. I would also thank Ms. Mary Lou Dufton and Mr. Greg Preston for their administrative assistance through my Ph.D. program.

I would like to thank my sincere friends Yilin Chen, Danqiao Guo, Yidan Shi and Meng Yuan for all the protean and funny topics (can be called as encyclopedia) we have been sharing and enjoying during my Ph.D. studies. A special thank to my ten-years friend Lin Qin for her support and being there when needed. I would also like to thank Fangya Mao (who I first met at Waterloo) and Menglu Che for the days that we exchanged thoughts and discussions on the topics such as “the values of life”. Thanks go to Ying Lin, for giving me years of fun memory when we were talking. Thank you Roseanne Park for her encouragement and support. I also thank all my friends who are not be listed here for your support and friendship that accompany me through those happy and hard days. Thanks also go to the members of the University of Waterloo Kendo Club for their warm words and different Kendo practice nights. Last, but not least, a deep appreciation and “thank you” to my parents, Kun and Lieyi, for their love and support.

Dedication

This is dedicated to my parents. For their endless love, support, and encouragement.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this thesis, we focus on topics concerning matrix-variate regression with measurement error from both frequentist and Bayesian aspects. Matrix-variate regression is a useful modelling method for analyzing data with a matrix structure, such as brain imaging data. This model assumes that the elements from the same row or column share the same effect. Many researchers proposed various modeling methods based on generalized linear regression (e.g; Hung and Wang 2013; Li 2014) or Bayesian modelling methods (Carvalho and West 2007; Guhaniyogi et al. 2017). However, those methods do not apply to data with measurement error or misclasscification, an issue which is inevitable in the data collecting process. Li (2014) discussed some issues on this topic from the frequentist viewpoint. But it lacks a solid theoretical support. To fill in this incomplete research area, we investigate the influence of measurement error in matrix-variates and the response misclassification on parameter estimation procedures and propose valid inference models.

Another problem of our interest concerns variable selection with matrix-variate regression. The structure of the matrix-variate data is complex, and the sparsity assumption usually needs to be added to the data. Some existing works considered to include a penalty function to the matrix-variate regression model to conduct inferences (Hung and Wang 2013; Zhou et al. 2013), but the influence of measurement error or misclassification on inferential procedures has not been investigated. In this thesis, we consider the matrix-variate logistic regression model with penalty functions where the response misclassification is accounted for.

Besides the frequentist viewpoint, Bayesian methods can provide useful procedures to model matrix-variate data. With Bayesian methods (Wei and Ghosal 2020), the parameters of the matrix-variate can be obtained by dropping the assumption that each row and column share the same effect. A shrunk prior can be imposed on the row and column parameters element-wisely to conduct inferences. However, no work has been available to accommodate measurement error or misclassification effects under such settings. Motivated by this, we consider a Bayesian method based on logistic regression with matrix-variate data and response misclassification.

To better understand our development in the following chapters, in this chapter, we review relevant topics. The remainder is organized as follows. In Section 1.1, we introduce basic notation for describing matrix-variate data. In Section 1.2, we introduce logistic regression analysis from both the frequentist and Bayesian aspects. In Section 1.3, we explain the measurement error/misclassificaiton mechanisms and present the basics for correcting measurement error/misclassificaiton. In Sections 1.4 and 1.5, we discuss commonly used regularization methods for frequentist and Bayesian procedures.

1.1 Matrix-Variate Data

For $k = 1, ..., n$, let x_k be a $(p+1) \times q$ dimension matrix, where $x_{k,ij}$ is the ith *row* and the jth *column* element in x_k for $i = 1, ..., (p+1)$ and $j = 1, ..., q$. We name covariate data which has the structure like x_k as *matrix-variate* data. In applications, biomedical data, such as Electroencephalography (EEG) data and anatomical magnetic resonance imaging (MRI), exhibit a natural matrix structure. Traditional modelling methods, such as generalized linear regression (GLM), by vectorizing matrix data, may not be feasible for handling this kind of data due to the complex data structure and computation burdens. The assumption that each *row* or *column* shares the same effects is often imposed (Kolda and Bader 2009; Li et al. 2010) for dimension reduction.

To see this, considering the GLM, one may model the matrix-variate data as

$$g(\mu_k) = \gamma_0 + \big\langle x_k, \mathcal{B} \big\rangle + \gamma_1^{\intercal} z_k \tag{1.1}$$

where $\mu_k = P(Y_k = 1|x_k, z_k)$, $g(\cdot)$ is the link function, γ_0 is a scalar, γ_1 is a $p_z \times 1$ vector parameter, $\mathcal{B}$ is a $(p+1) \times q$ matrix, and $\big\langle x_k, \mathcal{B} \big\rangle = \langle \text{vec}(x_k), \text{vec}(\mathcal{B}) \rangle = \sum_{i,j} \mathcal{B}_{ij} x_{k,ij}$.

Using model (1.1), we have to estimate $(p+1) \times q + p_z + 1$ parameters, which are usually large relative to the usual sample size. The rank-1 matrix decomposition of $\mathcal{B}$, say, $\mathcal{B} = \alpha^{\intercal}\beta$, separates the matrix-variate coefficients into two vectors of covariates, where α is the $(p+1)$-dimensional *row* coefficients and β is the q-dimensional *column* coefficients. Under this rank-1 matrix decomposition, the number of matrix-variate parameters needed to be estimated decreases from $(p+1) \times q$ to $p+q+1$. The model by Hung and Wang (2013) used this idea to fit a logistic regression model with an additional constraint that one of the elements in α is set as 1 to overcome the nonidentifiability issue related to the rank-1 matrix decomposition. Zhou et al. (2013) proposed a more general case using the GLM with penalty functions based on a rank-R parafac decomposition where the covariate data x_k is a tensor, where the rank-R parafac decomposition of $\mathcal{B}$ is $\mathcal{B} = \sum_{r=1}^{R} \alpha^{(r)} \circ \beta^{(r)}$, $\alpha^{(r)}$ and $\beta^{(r)}$ for $r = 1, ..., R$ are column vectors, $\alpha^{(r)} \circ \beta^{(r)}$ is the outer product of $\alpha^{(r)}$ and $\beta^{(r)}$, and R is a positive integer. Zhou and Li (2014) formulated a spectral regularization for matrix-variate regression, which minimizes a function combining a function of the singular values of $\mathcal{B}$ and the loss function of the negative log-likelihood based on the GLM. Recently, Guhaniyogi et al. (2017) proposed a tensor regression method with Bayesian analysis under the rank-R parafac decomposition, where shrinkage priors were assigned to $\alpha^{(r)}$ and $\beta^{(r)}$ under the sparsity assumption on $\mathcal{B}$.

1.2 Logistic Regression Analysis

As claimed by Walker and Duncan (1967), the logistic regression model, initially proposed by Cox (1958) to estimate the probability of an event as a function of independent variables, has been widely used for binary responses related to disease classification, risk factor selection, and other aims. In this section, the model is reviewed from the frequentist and Bayesian aspects.

1.2.1 Logistic Regression Analysis

For $k = 1, ..., n$, let Y_k be the independent binary response labeled as 1 with an event occurring or 0 otherwise. Given the vector-covariates, z_k, the logistic model is

$$\text{logit}\{P(Y_k = 1|z_k)\} = \beta_0 + \beta_z^{\intercal} z_k \tag{1.2}$$

for $k = 1, ..., n$, where β_0 is the scalar parameter, and β_z is a $p_z \times 1$ vector of parameters. Let $\beta = (\beta_0, \beta_z^{\intercal})^{\intercal}$ and $\mathbf{z} = (z_1^{\intercal}, ..., z_n^{\intercal})^{\intercal}$.

An important concept related to the logistic regression model is the odds ratio which is easy to interpret. The odds of the event occurring is

$$\frac{P(Y_k = 1|z_k)}{1 - P(Y_k = 1|z_k)} = \exp(\beta_0 + \beta_z^{\intercal} z_k).$$

Then the odds ratio for a unit change in a specific covariate j, z_{kj}, with other covariates kept fixed, is $\exp(\beta_{zj})$ for $j = 1, ..., p_z$.

With high dimensional data, the penalized logistic regression model is often employed (Lokhorst 1999; Shevade and Keerthi 2003; Lukas Meier and Bühlmann 2008).

1.2.2 Bayesian Logistic Regression Analysis

Bayesian analysis is another useful tool in statistical analysis. As discussed by O'Brien and Dunson (2004), Bayesian approaches have two main advantages over quasi-likelihood and likelihood-based frequentist methods. First, based on the Markov chain Monte Carlo (MCMC) algorithms, the large MCMC iterations can overcome the small sample limitation by using the exact posterior. Secondly, Bayesian methods can impose additional information into estimation processes by using an informative prior distribution. With Bayesian analysis, a prior probability density function (pdf) is assigned to β:

$$\beta \sim \pi(\beta|I_0),$$

where I_0 denotes the initial information, and $\pi(\beta|I_0)$ can be non-informative or informative. Combining with the logistic regression model, the posterior distribution $p(\beta|D(I_0, \mathbb{Y}))$ is

$$p(\beta|D(I_0, \mathbb{Y})) = c\pi(\beta|I_0)\ell(\beta|\mathbb{Y}), \tag{1.3}$$

where $D(I_0, \mathbb{Y})$ contains the prior information as well as the sample information, $\ell(\beta|\mathbb{Y})$ is the likelihood function derived from the logistic regression model, with $\mathbb{Y} = (Y_1, ..., Y_n)^\intercal$, and c is the normalizing constant with the form

$$c^{-1} = \int \pi(\beta|I_0)\ell(\beta|\mathbb{Y})d\beta.$$

Under model (1.2), (1.3) generally does not have a closed form. Thus, Zellner and Rossi (1984) proposed to estimate model (1.2) with the help of a normal approximation to (1.3). To evaluate c, they used the importance sampling procedure. A variety of MCMC methods were developed for the Bayesian estimation of the logistic regression model, such as Gibbs sampling or independent Metropolis-Hastings (MH) sampling methods (Zeger and Karim 1991; Gamerman 1997; Rossi et al. 2005), in combination with an approximation to (1.3).

Meanwhile, the data augmentation methods that facilitate Gibbs sampling (Holmes and Held 2006; Gramacy and Polson 2012; Polson et al. 2013) can avoid the approximation procedure to (1.3). The data augmentation methods for the logistic regression model were extended from the simple latent-variable method of Albert and Chib (1993), who introduced n latent variables, $W = (w_1, ..., w_n)^\intercal$, where $w_k \sim N(\beta_0 + \beta_z^\intercal z_k, 1)$, such that $Y_k = 1$ if $w_k > 0$ and $Y_k = 0$ otherwise. Then the posterior density of β given W, $\mathbb{Y}$, and $\mathbf{z}$ is distributed as a multivariate normal distribution, where β can be sampled using the Gibbs sampler easily. The Bayesian inference for logistic models using Pólya–Gamma latent variables (Polson et al. 2013) is perhaps the most efficient method, compared to other data augmentation methods. The Pólya–Gamma approach is close to that of the independent MH, whereas the MH jumping distribution needs to be chosen carefully for simple logistic regression models with abundant data and no hierarchical structures.

1.3 Measurement Error/Misclassification

In practice, imprecise measurements, or mismeasurements, often exist in data collection procedures with different reasons (Yi and Cook 2005; Carroll et al. 2006). They usually generate new inference problems and need to be corrected for conducting valid inferences. In this section, we review some measurement error models for continuous covariates and misclassification models for univariate binary responses. For $k = 1, ..., n$, let x_k be the p_x-dimensional true continuous covariate, and let X_k^* be its the surrogate, or observed value. Let z_k be the p_z-dimensional true covariate which is precisely measured. We let Y_k denote the true binary response, taking value 0 or 1, and let Y_k^* denote its surrogate or observed response. Let $h(\cdot)$ and $h(\cdot|\cdot)$ denote the true marginal and conditional probability density or mass functions for the random variables indicated by the corresponding arguments, respectively; in the following development, we may loosely use upper case letters for some of their arguments though ideally, lower case letters should be used for clarity.

1.3.1 Modelling Measurement Error in Continuous Variables

First, we describe the measurement error/misclassification mechanisms. Given the true covariates $\{x_k, z_k\}$, if Y_k and X_k^* are conditionally independent, i.e.,

$$h(Y_k|X_k^*, x_k, z_k) = h(Y_k|x_k, z_k),$$

then we call the measurement error process a *nondifferential measurement error mechanism* or a *nondifferential misclassification mechanism* (if x_k is discrete). This mechanism implies that the surrogate X_k^* has no information on inference about the response process if the true covariates are given (Yi 2017, Section 2.4). To do inferences, we may factorize the joint distribution $h(Y_k, x_k, X_k^*, z_k)$ as

$$h(Y_k, x_k, X_k^*, z_k) = h(Y_k|x_k, X_k^*, z_k)h(x_k, X_K^*, z_k) = h(Y_k|x_k, z_k)h(x_k, X_k^*, z_k).$$

To describe the measurement error process, we can further factorize $h(x_k, X_k^*, z_k)$ as

$$h(x_k, X_k^*, z_k) = h(X_k^*|x_k, z_k)h(x_k, z_k)$$

or

$$h(x_k, X_k^*, z_k) = h(x_k|X_k^*, z_k)h(X_k^*, z_k).$$

In contrast to the *nondifferential error mechanism*, if

$$h(Y_k|X_k^*, x_k, z_k) \neq h(Y_k|x_k, z_k),$$

then the mechanism is called a *differential measurement error mechanism* or a *differential misclassification mechanism* (if x_k is discrete). This mechanism usually arises from retrospective studies, such as case-control studies. In this case, we may decompose $h(Y_k, x_k, X_k^*, z_k)$ as

$$h(Y_k, x_k, X_k^*, z_k) = h(X_k^*|Y_k, x_k, z_k)h(Y_k|x_k, z_k)h(x_k, z_k).$$

This decomposition allows us to express our interested $h(Y_k|x_k, z_k)$ explicitly, which can be modelled by standard modelling techniques.

In the following, we introduce two widely used measurement error models for scenarios with *nondifferential measurement error mechanism*, a mechanism that has been mostly considered in the literature of measurement error models (Fuller 1987; Carroll et al. 2006; Yi 2017):

- *Classical Additive Error Model*

  With the feature that the observed covariate X_k^* is more variable than the true covariate x_k, the model is

  $$X_k^* = x_k + e_k,$$

  where the error term e_k is assumed to be independent of x_k, and the e_k have mean zero and covariance matrix, say Σ_e.

- *Berkson Model*

Viewing that the true observation x_k as fluctuating around the surrogate X_k^*, we consider the model

$$x_k = X_k^* + e_k, \tag{1.4}$$

where the error term e_k is assumed to be independent of X_k^*, and the e_k have mean zero and covaraince matrix, say Σ_e.

Model (1.4) indicates that the true covariate x_k is more variable than the surrogate X_k^*. For example in radiation epidemiology, the radiation dose is prescribed for a patient but the actually absorbed dose by the patient is unknown and varies around the prescribed dose.

When the error covariance matrix Σ_e is unknown, replicates or validation samples are often needed to estimate the error covariance matrix. In the following, we introduce two kinds of data sets discussed by Yi (2017, Section 2.4).

- Validation Subsample

  We let $\mathcal{M}$ denote the index set of subjects who are in the main study. Let $\mathcal{D}$ be the data set that collects different types of measurements, say $\mathcal{D} = \{W_k : k \in \mathcal{V}\}$, where $\mathcal{V}$ is the set of subjects indices, and W_k may be $\{Y_k, x_k, X_k^*, z_k\}$ or $\{x_k, X_k^*, z_k\}$. When $\mathcal{V}$ is a subset of $\mathcal{M}$ with $W = \{Y_k, x_k, X_k^*, z_k\}$, $\mathcal{D}$ is called an *internal validation* subsample; when $\mathcal{V}$ and $\mathcal{M}$ are disjoint, $\mathcal{D}$ is called an *external validation* subsample where W_k may only contain $\{x_k, X_k^*, z_k\}$.

- Repeated Measurements

  In practice, the surrogate measurements may be measured a couple of times such that W_k may have a form $\{X_{kj}^*\}$ or $\{Y_k, X_{kj}^*\}$, where X_{kj}^* is the jth repeated measurement of x_k for $k = 1, ..., n_k$ and n_k is an integer larger than 1.

Investigating the measurement error effects has attracted attention long ago (Wald 1940; Madansky 1959). General strategies of handing measurement error include likelihood-based correction methods (Lindsay 1982; Stefanski and Carroll 1987; Yi et al. 2015), unbiased estimating functions methods (Prentice 1982; Wang and Pepe 2000; Freedman et al. 2004; Yi et al. 2012), and methods of correcting naive estimators (Stefanski and Carroll 1985; Cook and Stefanski 1994; Yi and Reid 2010). More references on different topics can be found in Fuller (1987), Carroll et al. (2006) and Yi (2017).

1.3.2 Modelling Misclassification in Univariate Binary Response

We assume that Y_k is modeled through a binary regression model within the class of generalized linear models. Let z_k be the $p_z \times 1$ dimensional precisely measured covariates. The relationship between the response and covariate variables can be featured by the conditional mean response given covariates, $\mu_k = E(Y_k|z_k)$, through various link functions (McCullagh and Nelder 1989, p.31).

To model the response misclassification process, we let

$$\tau_{01}(z_k) = P(Y_k^* = 1|Y_k = 0, z_k) \text{ and } \tau_{10}(z_k) = P(Y_k^* = 0|Y_k = 1, z_k)$$

be the conditional misclassification probabilities, given the covariates z_k. The *sensitivity* of the measurement Y_k^* is given by $1 - \tau_{10}(z_k)$, and the *specificity* of the measurement Y_k^* is $1 - \tau_{01}(z_k)$.

As discussed by Neuhaus (1999) and Yi (2017, Section 8.2), under the condition that the misclassification probabilities are constants, ignoring the response misclassification in the analysis has the same effects as misspecifiying the link function in the analysis for generalized linear models. When the misclassification probabilities are associated with the covariates, the model for $P(Y_k^* = 1|z_k)$ may not be in the family of the generalized linear model (Neuhaus 1999).

1.4 Regularization Methods

High dimensional data analysis is a challenging problem because of the computation burden and the complexity of data structures. Under the sparsity assumption that only a few important covariates are non-zeros in the model, various regularization methods have been proposed to overcome these difficulties. A general form of regularization methods under the likelihood method, based on the penalty function $p_\lambda(\cdot)$, is given by

$$\ell(\beta) + \sum_{j=1}^{q} p_\lambda(\beta_j) \tag{1.5}$$

where $\ell(\cdot)$ is the log-likelihood function derived from a model, β_j is the jth component of the $q \times 1$ unknown vector parameter β, and λ is the *tuning parameter*.

The following are commonly used penalty functions:

- Least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996):

$$p(\beta_j) = \lambda|\beta_j|.$$

- Smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001):

$$p'_{\lambda_n}(\zeta) = \lambda_n\Big\{I(|\zeta| \le \lambda_n) + \frac{(a\lambda_n - |\zeta|)_+}{(a-1)\lambda_n} I(|\zeta| > \lambda_n)\Big\}\text{sign}(\zeta),$$

  where the sign function $\text{sign}(\zeta) = -1, 0$ and 1 when $\zeta < 0, = 0$ and > 0, respectively; a is a constant larger than 2; and $I(\cdot)$ is the indicator function.

- Elastic net (Zou and Hastie 2005):

$$p(\beta_j) = \lambda_1|\beta_j| + \lambda_2\beta_j^2.$$

- Adaptive LASSO (Zou 2006):

$$p(\beta_j) = \lambda w_i|\beta_j|,$$

  where w_i is a weight.

The LASSO method with the ℓ_1 penalty function imposed on the regression coefficients does both continuous shrinkage and automatic variable selection simultaneously. However, as discussed by Zou and Hastie (2005), due to the nature of the convex optimization problem, the LASSO method can only select at most n variables if $p > n$. Moreover, the LASSO method tends to select one of a group of variables which are highly correlated with each other. To overcome these problems, they proposed the elastic net method which combines the ℓ_1 and ℓ_2 penalties together to select groups of correlated variables. Later, Zou (2006) proposed the adaptive LASSO method to fix a problem that in some scenarios, the LASSO selection cannot be consistent. Unlike the traditional regularization methods, the SCAD method (Fan and Li 2001) is based on the non-convex penalty functions and possesses the oracle properties.

The regularization methods can be naturally employed under matrix-variate or tensor regression models after different data decomposition processes are implemented. Zhou and Li (2014) proposed regularized matrix regression for the response in the exponential family by penalizing the spectrum of the matrix parameters. Hung and Wang (2013) and Zhou et al. (2013) added penalty functions to the models for matrix-variate data and tensor data, respectively, based on the rank-R decomposition of the parameters, which has the form

$$\ell(\gamma, \alpha, \beta) + \sum_{r=1}^{R} \sum_{i=1}^{p+1} p(\alpha_i^{(r)}) + \sum_{r=1}^{R} \sum_{j=1}^{q} p(\beta_j^{(r)}),$$

where matrix-variate $\mathcal{B} = \sum_{r=1}^{R} \alpha^{(r)} \circ \beta^{(r)}$, $\alpha_i^{(r)}$ is the ith component of $\alpha^{(r)}$ for $i = 1, ..., p+1$, $\beta_j^{(r)}$ is the jth component of $\beta^{(r)}$ for $j = 1, ..., q$, and R is a positive integer.

1.5 Bayesian Variable Selection Methods

Unlike regularization methods, Bayesian variable selection methods address the parameter selection procedure by assigning shrinkage prior to the parameters. These priors have the ability to shrink small coefficients towards zero while minimizing shrinkage of large coefficients. The first type of these priors is the point-mass prior which combines a probability at zero and a non-zero continuous distribution, such as the spike-and-slab prior (George and McCulloch 1993; Ishwaran and Rao 2005) which mixes two normal distributions with one highly concentrated at zero. For example, a popular version of the spike-and-slab model (George and McCulloch 1993) is

$$\beta_j | \psi_j \sim (1 - \psi_j) N(0, \delta_l^2) + \psi_j N(0, c_j^2 \delta_j^2),$$

where β_j is the jth component of a $q \times 1$ vector parameter β, c_j is a constant, δ_l^2 and δ_j^2 are hyper-parameters, ψ_j is a latent variable with value 0 or 1, and

$$P(\psi_j = 1) = 1 - P(\psi_j = 0) = p_j.$$

When $\psi_j = 0$ and δ_l^2 is assigned to be small, $\beta_j | \psi_j \sim N(0, \delta_l^2)$ and β_j can be estimated as zero. When $\psi_j = 1$ and c_j is assigned to be large, then a non-zero β_j can be selected in the final model.

Compared to the spike-and-slab prior, Bayesian LASSO (Park and Casella 2008), which uses a double exponential prior distribution on coefficients, and it has good performance for high dimensional models with the sparsity assumption. Park and Casella (2008) considered a conditional Laplace prior for β with the form

$$\pi(\beta | \sigma^2) = \prod_{j=1}^{q} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda |\beta_j| / \sqrt{\sigma^2}},$$

where λ is the tuning parameter, and σ^2 is the hyper-parameter.

Another type of Bayesian shrinkage priors uses continuous densities which have a good performance on parameters shrinkage, such as the horseshoe prior (Carvalho et al. 2010) and the Dirichlet-Laplace (DL) prior (Bhattacharya et al. 2015). This type of prior can be written as a global-local scale mixture of Gaussian distributions:

$$\beta_j \sim N(0, \lambda_j a), \quad \lambda_j \sim f, \quad a \sim g,$$

where for $j = 1, ..., q$, λ_j is the local scale which follows a distribution f, and a is the global scale which follows a distribution g.

A few papers address convergence results on Bayesian variable selection methods beyond the linear regression model, such as adaptive density regression (Shen and Ghosaĺ 2016) and logistic regression (Atchadé 2017; Wei and Ghosal 2020). Under the logistic regression model, the horseshoe prior has the best performance than the point-mass prior, the DL prior, Bayesian LASSO, and non-informative priors (Wei and Ghosal 2020). With tensor data, a multiway Dirichlet generalized double Pareto prior (Guhaniyogi et al. 2017) was recently proposed for the generalized linear regression setting, and it performed well under the Gaussian assumption for the response variable.

1.6 Thesis Topics and the Outline

While many inference methods have been developed to handle various problems concerning matrix-variate regression or measurement error models, interesting research problems remain unexplored. This thesis investigates several important problems which are described as follows. This thesis contains six chapters with the last chapter concluding the thesis and the appendix including additional materials for Chapters 2-5. The remaining chapters are organized as follows.

Chapter 2: Matrix-Variate Logistic Regression with Measurement Error

The logistic regression model has been widely used to handle data with binary responses, where the logit link function is used to feature the relationship between the response probability and a vector of associated covariates. With the advent of the new technology of collecting complex-featured data (e.g., electroencephalography(EEG) imaging data which involve both channel and temporal information), conventional logistic regression models become inadequate to facilitate the dependence of the binary outcome on covariates in a matrix form. Driven by this, matrix-variate logistic regression models were proposed to cover a broader scope of problems than that of the usual logistic regression model. Such models are useful for analyzing brain imaging data which commonly contain a matrix-variate or a tensor structure (e.g., Hung and Wang 2013; Zhou et al. 2013; Li et al. 2018).

While matrix-variate logistic regression models are useful for dealing with brain imaging data, its application hinges on the critical assumption that the variables are precisely measured. Such an assumption is commonly violated in pre-processed imaging data due to various reasons related to cardiac and respiratory activities. Even through scientists attempt to apply different methods to process the data, measurement error in the variables cannot be completely eliminated (Sobel and Lindquist 2014). It has been well understood that measurement error in the variables can seriously bias the inference results derived from the logistic regression model, and many methods have been developed to correct for the measurement error effects accordingly (e.g., Stefanski and Carroll 1985; Gleser 1996; Cook and Stefanski 1994; Buzas and Stefanski 1996). However, little work on matrix-variate logistic regression with measurement error has been available although in an unpublished PhD thesis, Li (2014) discussed some issues on this topic in an ad hoc way.

It is unclear how measurement error in the matrix-variate may affect inference results. In the presence of measurement error, it is imperative to develop valid inference procedures to accommodate measurement error effects in a rigorous manner. In Chapter 2, we target these problems and explore matrix-variate logistic regression models with covariate measurement error. We investigate the asymptotic bias induced from the naive analysis which ignores measurement error, and then develop two methods to correct for the biases of the naive analysis by making or not making a distribution assumption for the measurement error model. To the best of our knowledge, this is the first research which provides a rigorous study on matrix-variate logistic regression with covariate measurement error with the theoretical results carefully established. The work in this chapter has been wrapped up as a research article, Fang and Yi (2020b), and has been accepted by *Biometrika*.

Chapter 3: Imputation and Likelihood Methods for Matrix-Variate Logistic Regression with Response Misclassification

In contrast to the challenges presented by error-contaminated covariates discussed in Chapter 2, response misclassification impairs inference procedures derived from the matrix-variate logistic regression model as well. In the conventional regression context, bias analysis of response mismeasurements has attracted extensive attention, and many methods of accommodating mismeasurement effects have been developed (e.g., Stefanski and Carroll 1985; Albert et al. 1997; Neuhaus 1999; Neuhaus 2002; Chen et al. 2011; Li 2014; Yi 2017, Chapter 8) However, matrix-variate logistic regression with response error has not received much attention though real data do often possess such features.

Driven by the paucity of such research, in Chapter 3, we study matrix-variate logistic regression with response misclassification. We develop two inferential methods to account for misclassification effects. The first method is an imputation method which replaces the response variable with an unbiased pseudo-response variable, derived from the observed surrogate response measurement, in the estimation procedure. The second method is derived from the likelihood function for the observed response surrogate. Our development is carried out for two settings to address misclassification effects: misclassification rates are either known or estimated from the validation subsample information. The validity of our methods is justified by the establishment of theoretical results. The work in this chapter has been wrapped up as a research paper, Fang and Yi (2020a), that has been invited for a revision by *The Canadian Journal of Statistics*.

Chapter 4: Regularized Matrix-Variate Logistic Regression with Response Misclassification

As introduced in Chapter 3, usual logistic regression has been generalized to accommodate covariates with matrix structures which arise commonly from biomedical research concerning cancer classification and brain imaging analysis (e.g., Zhou et al. 2013; Hung and Wang 2013; Zhang et al. 2014). Meanwhile, matrix-variates usually haves the sparsity property and penalty terms are commonly added to estimation procedures for variable selection (e.g., Zhang et al. 2014).

Although we examine the effects of response misclassification on matrix-variate logistic regression and propose valid methods in Chapter 3 to correct for the biases of the naive analysis, the sparsity property is not considered there. Though penalized estimation procedures are commonly used in regression analysis, only a few settings incorporate measurement error (e.g., Ma and Li 2010; Yi et al. 2015). With misclassification in response variables, there has been no work to study error effects under the matrix-variate logistic regression model.

Motivated by the paucity of such research, in Chapter 4, we extend the work in Chapter 3 to further study regularized matrix-variate logistic regression with response misclassification. Our development is carried out for two settings where misclassification rates are either known or estimated from a validation subsample. The validity of our methods is justified by the establishment of theoretical results. This project has been wrapped up as a paper and submitted to a journal for publication (Fang and Yi 2020c).

Chapter 5: Bayesian Analysis for Matrix-Variate Logistic Regression with/without Response Misclassification

For statistical models such as linear regression, high dimensional data analysis is challenging due to the computational burden and intrinsic complex data structures. Bayesian variable selection procedures have the advantage of addressing the parameter selection uncertainty automatically by using a prior, such as the spike-and-slab prior (George and McCulloch 1993; Ishwaran and Rao 2005), horseshoe prior (Carvalho et al. 2010), and Dirichlet-Laplace (DL) prior (Bhattacharya et al. 2015). For logistic regression, the Bayesian inference has long been considered as a hard problem due to the lack of closed forms of posterior densities of the model parameters. One useful Bayesian inference method is based on the normal approximation to the posterior density of the parameter of interest (Zellner and Rossi 1984; Zeger and Karim 1991; Gamerman 1997; Rossi et al. 2005), which however, has much computational burden. Data augmentation methods that facilitate Gibbs sampling (Holmes and Held 2006; Gramacy and Polson 2012; Polson et al. 2013) offer an effective alternative.

As introduced in Chapter 4, matrix-variate data has a complex matrix structure and often contains many unimportant components. The Bayesian variable selection procedure is an efficient way to handle such problems. A multiway Dirichlet generalized double Pareto prior (Guhaniyogi et al. 2017) was recently proposed for tensor-variate data with the Gaussian assumption. An interesting problem is to investigate the influence on the parameter estimation of imprecisely measured binary responses with matrix-variate regression models under the Bayesian framework. Although there has some works on investigating the effects of mismeasured covariates (Richardson and Gilks 1993; Dellaportas and Stephens 1993; Gustafson 2003) or binary response misclassification under conventional binary regression settings (Paulino et al. 2003; Gustafson 2003; McInturff et al. 2004; Gerlach and Stamey 2007), Bayesian matrix-variate logistic regression with response misclassification has not been explored.

Motivated by this, in Chapter 5, we propose a Bayesian inference procedure using the horseshoe prior under matrix-variate logistic regression with the help of augmented data from the Pólya-Gamma distribution. We develop an algorithm to accommodate the influence of binary response misclassification on the Bayesian estimation procedure.

Numerical studies are conducted to evaluate the performance of the proposed method.

Chapter 2

Matrix-Variate Logistic Regression with Measurement Error

In this chapter, we investigate how measurement error in the matrix-variate affects the parameter inference, and explore matrix-variate logistic regression models with measurement error. The remainder is organized as follows. In Section 2.1, we introduce the matrix-variate logistic regression model and the estimation method for the error-free context. In Section 2.2, we conduct the bias analysis of the naive analysis which ignores measurement error present in matrix-variate logistic regression. In Section 2.3, we develop two inference methods to adjust for measurement error effects by capitalizing on the bias analysis in Section 2.2. In Section 2.4, we conduct simulation studies to assess the performance of the methods developed in Section 2.4 as well as to demonstrate the biased effects of the naive analysis. We also present an application to a EEG data set.

2.1 Notation and Framework

2.1.1 Matrix-Variate Logistic Regression Model

For subject k with $k = 1, ..., n$, let Y_k be the binary response variable with value 1 for having a disease and 0 otherwise, let $x_k = [x_{k,ij}]_{(p+1)\times q}$ be the associated $(p + 1) \times q$ covariate matrix where $x_{k,ij}$ is the observation at row i and column j for subject k, and let z_k be the associated $p_z \times 1$ covariate vector for subject k. In this paper for subject $k = 1, ..., n$, x_k and z_k are treated as fixed measurements in the sense that their distributions are unspecified.

Matrix-variate regression is useful for handling data with a matrix structure when the data in the same rows are perceived to share the same effects, and the data in same columns share the same effects. For instance, EEG data involve measurements associated with multiple channels and different time points. Using a matrix, say a $(p+1) \times q$ matrix x_k for subject k, is most natural and informative to represent EEG measurements for a subject. If we use the conventional regression to study the effects of a combination of a specific channel and a time point on a disease, we would first convert the matrix into a vector by stacking the columns of x_k from left to right to form a column vector $\text{vec}(x_k)$, and then fit a regression model:

$$\text{logit}\{P(Y_k = 1 \mid x_k)\} =< x_k, \mathcal{B} >, \tag{2.1}$$

where $\mathcal{B}$ is the matrix-structured coefficients, and $< x_k, \mathcal{B} >= \text{vec}(x_k)^\intercal \text{vec}(\mathcal{B}) = \sum_{i,j} \mathcal{B}_{ij} x_{k,ij}$, with $\text{vec}(\mathcal{B})$ representing the vectoring form of $\mathcal{B}$ and $\mathcal{B}_{ij}$ standing for element (i, j) of $\mathcal{B}$. As pointed out by Hung and Wang (2013), this modeling scheme introduces $(p + 1) \times q$ parameters which can be too large to handle. In addition, limited sample sizes in many problems hinder us from estimating a large number of parameters. Vectorization not only introduces the model a huge number of parameters to estimate but also destroys the natural matrix structure which can be quite informative.

To overcome these issues, we consider the matrix-variate logistic regression model:

$$\text{logit}\{P(Y_k = 1 \mid x_k, z_k)\} = \alpha^\intercal x_k \beta + \gamma^\intercal z_k, \tag{2.2}$$

where α is a $(p+1) \times 1$ parameter vector, β is a $q \times 1$ parameter vector, and γ is a $p_z \times 1$ parameter vector. To distinguish α and β, we call them the *row* parameter and the *column* parameter, respectively. Note that since no intercept is included in model (2.2), x_k can be understood as a centered matrix, given by $x_{ck} = x_k - \bar{x}$, where $\bar{x} = (1/n) \sum_{k=1}^{n} x_k$; or alternatively, we include 1 as the first element of z_k. In the following development, we take x_k to be a centered version x_{ck} when using model (2.2).

By rank-1 Canonical Polyadic Decomposition (CP-decomposition)(Kolda and Bader 2009), parameters α and β in model (2.2) are related to $\mathcal{B}$ in model (2.1): $\mathcal{B} = \alpha \circ \beta$, where $\circ$ denotes the outer product of two column vectors. We note that the CP-decomposition of $\mathcal{B}$ is not unique. In other words, $\mathcal{B}$ is not identifiable since for any constant $c \neq 0$, $\mathcal{B} = (c^{-1}\alpha) \circ (c\beta)$. To overcome nonidentifiability issues, constraints are often imposed on the parameter space so that certain values are inadmissible. A convention is to set the first element of α to be 1 (e.g., Hung and Wang 2013), which is also adopted in our development unless stated otherwise. However, for ease of exposition, we still use α to denote the subvector of the rest p-dimensional real parameters, and let $\theta = (\alpha^\intercal, \beta^\intercal, \gamma^\intercal)^\intercal$

denote the vector of parameters of interest with dimension $d = p + q + p_z$.

Our model (2.2) generalizes the matrix-variate logistic model (2.1) discussed by Hung and Wang (2013). Model (2.2) is more flexible in featuring the dependence of the binary outcome on covariates which include both a matrix form and a vector form. Parameters in model (2.2) are interpretive in terms of odds ratios. Let α_i be the ith element of α, for $i = 2, ..., p+1$; let β_j be the jth element of β for $j = 1, ..., q$; and let γ_l be the lth element of γ for $l = 1, ..., p_z$. Given i and j, let $\tilde{x}_k(i,j)$ represent the matrix identical to x_k except that the element (i,j) of $\tilde{x}_k(i,j)$ is set to be $x_{k,ij} + 1$. Then $\alpha_i \beta_j$ represents the log odds ratio, $\log[\text{odds}\{\tilde{x}_k(i,j)\}/\text{odds}(x_k)]$, where $\text{odds}(A) = pr(Y = 1 \mid A, z_k)/pr(Y = 0 \mid A, z_k)$ with $A = \tilde{x}_k(i,j)$ or x_k. Parameters γ_l can be interpreted in a similar manner.

2.1.2 Estimation of Model Parameters

Estimation of θ is carried out using the maximum likelihood method with the constraint on α discussed in Section 2.1 imposed. Typically, this can be done using the *block relaxing algorithm* described by Zhou et al. (2013). Let

$$\ell_n(\alpha, \beta, \gamma) = (1/n) \sum_{k=1}^{n} \big[Y_k(\alpha^\intercal x_k \beta + \gamma^\intercal z_k) - \log\{1 + \exp(\alpha^\intercal x_k \beta + \gamma^\intercal z_k)\} \big]. \tag{2.3}$$

be $1/n$ times log-likelihood function contributed from the sample which is derived from model (2.2).

Instead of maximizing (2.3) with respect to α, β, γ simultaneously, we take three steps, or called three blocks, to obtain the estimates of α, β, γ, separately. Shown in Table 2.1, in each block the likelihood function $\ell_n(\alpha, \beta, \gamma)$ is maximized with respect to one parameter with other two parameters fixed at the values of the previous iteration, where we use slightly different notation such as $\ell_n(\alpha|\beta^{(t)}, \gamma^{(t)})$ to emphasize that $\ell_n(\alpha, \beta, \gamma)$ is treated as a function of α with β and γ fixed at $\beta^{(t)}$ and $\gamma^{(t)}$, respectively. Any optimization procedure may be applied for this purpose. Zhou et al. (2013) commented that this algorithm works well for generalized linear models, including logistic models with canonical link functions. Multiple initial values for α and β may be tried to obtain the global maximum values. We suggest to randomly generate initial values of α and β from the uniform distribution $U(0, 1)$ and simply set the initial value of γ as 0.

Table 2.1: Block relaxation algorithm for maximizing $\ell_n(\alpha, \beta, \gamma)$

```
Initialize $\gamma^{(0)} = 0$, set $\alpha^{(0)}$ and $\beta^{(0)}$ as values generated from the uniform
distribution $U(0,1)$, and fix the first element of $\alpha^{(0)}$ as 1.
Repeat for $t = 0, 1, 2, ...$
    Block 1. $\alpha^{(t+1)} = \text{argmax}_{\alpha} \ell(\alpha | \beta^{(t)}, \gamma^{(t)})$
    Block 2. $\beta^{(t+1)} = \text{argmax}_{\beta} \ell(\beta | \alpha^{(t+1)}, \gamma^{(t)})$
    Block 3. $\gamma^{(t+1)} = \text{argmax}_{\gamma} \ell(\gamma | \alpha^{(t+1)}, \beta^{(t+1)})$
until
$|\ell_n(\alpha^{(t)}, \beta^{(t)}, \gamma^{(t)}) - \ell_n(\alpha^{(t+1)}, \beta^{(t+1)}, \gamma^{(t+1)})| < \epsilon$,
where $\epsilon$ is a pre-specified positive value showing the tolerance level.
```

2.2 Bias Analysis

2.2.1 Additive Matrix-Variates Measurement Error Model

In applications, measurements of variables are often subject to error. We consider settings where z_k is precisely measured but x_k is error-contaminated. Suppose that the precise measurement of x_k is unavailable but repeated surrogate measurements for x_k, X^*_{kr}, are observed for $r = 1, ..., m_k$, where m_k is a positive integer which may or may not depend on k. Assume that

$$X^*_{kr} = x_k + E_{kr}, \tag{2.4}$$

where E_{kr} is a $(p+1) \times q$ matrix of random noise with mean zero and is independent of Y_k and $\{x_k, z_k\}$. The independence of E_{kr} of $\{Y_k, x_k, z_k\}$ implies the nondifferential measurement error mechanism (Carroll et al. 2006, p.36; Yi 2017, p.50). That is, conditional on the true covariates $\{x_k, z_k\}$, X^*_{kr} is independent of Y_k, suggesting that the surrogate measurements X^*_{kr} have no predictive value for the outcome variable Y_k when x_k and z_k are controlled.

For $k = 1, ..., n$, and $r = 1, ..., m_k$, let $\text{vec}(E_{kr})$ represents a $(p+1)q$-dimensional vectorized version of E_{kr}, and let Ω_0 be the covariance matrix of $\text{vec}(E_{kr})$, i.e., $\Omega_0 = E\{\text{vec}(E_{kr})\text{vec}(E_{kr})^\intercal\}$. Let $\bar{X}^*_{k+} = (1/m_k)\sum_{r=1}^{m_k} X^*_{kr}$ and $\bar{E}_{k+} = (1/m_k)\sum_{r=1}^{m_k} E_{kr}$. Then $\bar{X}^*_{k+} = x_k + \bar{E}_{k+}$, and the mean and variance of $\text{vec}(\bar{X}^*_{k+})$ are $\text{vec}(x_k)$ and Ω_0/m_k, respectively. For $k = 1, ..., n$, define centered surrogate measurements X^*_k: $X^*_k = \bar{X}^*_{k+} - (1/n)\sum_{k=1}^{n} \bar{X}^*_{k+}$. Equivalently, let $\bar{U}_k = \bar{E}_{k+} - \bar{E}$ and $x_{ck} = x_k - \bar{x}$, where $\bar{E} = (1/n)\sum_{k=1}^{n} \bar{E}_{k+}$ and $\bar{x} = (1/n)\sum_{k=1}^{n} x_k$. Then

$$X^*_k = x_{ck} + \bar{U}_k, \tag{2.5}$$

where $\text{vec}(\bar{U}_k)$ has mean zero and covarariance matrix $\{(n-2)/(nm_k) + (1/n^2)\sum_{k=1}^{n} 1/m_k\}\Omega_0$.

For ease of exposition, we assume the number of replicates for each subject to be the same, i.e., $m_k = m$ for $k = 1, ..., n$, where m is a positive integer. Let $m_{\mathrm{c}} = mn/(n-1)$, then the covarariance matrix of $\mathrm{vec}(\bar{U}_k)$ is $E\{\mathrm{vec}(\bar{U}_k)\mathrm{vec}(\bar{U}_k)^{\intercal}\} = \Omega_0/m_{\mathrm{c}}$.

2.2.2 Naive Analysis

When matrix-variates are subject to measurement error, naively using the logistic regression model (2.2) with x_k replaced by X_k^* yields the model

$$\mathrm{logit}\{P(Y_k = 1 \mid X_k^*, z_k)\} = \alpha^{*\intercal} X_k^* \beta^* + \gamma^{*\intercal} z_k, \tag{2.6}$$

where α^*, β^*, and γ^* are the parameters which may differ from the corresponding parameter in (2.2). Let $\theta^* = (\alpha^{*\intercal}, \beta^{*\intercal}, \gamma^{*\intercal})^{\intercal}$.

Estimation of θ^* may proceed by mimicking the maximum likelihood method. That is, we maximize the log likelihood function derived from (2.6),

$$\ell_n^*(\theta^*) = (1/n) \sum_{k=1}^{n} Y_k(\alpha^{*\intercal} X_k^* \beta^* + \gamma^{*\intercal} z_k) - \log\{1 + \exp(\alpha^{*\intercal} X_k^* \beta^* + \gamma^{*\intercal} z_k)\} \tag{2.7}$$

with respect to θ^* and let $\hat{\theta}^* = (\hat{\alpha}^{*\intercal}, \hat{\beta}^{*\intercal}, \hat{\gamma}^{*\intercal})^{\intercal}$ denote the estimator of θ^*. While (2.7) is similar to (2.3) in the function form, the meaning of θ^* in (2.7) is not the same as that of θ in (2.3).

Under regularity conditions (e.g., White 1982), $\hat{\theta}^*$ solves

$$\mathrm{S}_n^*(\hat{\theta}^*) = 0, \tag{2.8}$$

where

$$\mathrm{S}_n^*(\theta^*) = \frac{\partial \ell_n^*}{\partial \theta^*} \triangleq \begin{pmatrix} \mathrm{S}_{\alpha,n}^*(\theta^*) \\ \mathrm{S}_{\beta,n}^*(\theta^*) \\ \mathrm{S}_{\gamma,n}^*(\theta^*) \end{pmatrix} = (1/n) \sum_{k=1}^{n} \begin{pmatrix} C_t^{\intercal} X_k^* \beta^* \\ X_k^{*\intercal} \alpha^* \\ z_k \end{pmatrix} \{Y_k - p_k(\theta^*; X_k^*)\}, \tag{2.9}$$

$C_t = [0_p, I_p]^{\intercal}$, 0_p is the $p \times 1$ vector of zeros, I_p is the $p \times p$ identity matrix, and $p_k(\theta^*; X_k^*) = P(Y_k = 1 \mid X_k^*, z_k)$ which equals, by (2.6),

$$p_k(\theta^*; X_k^*) = \frac{\exp(\alpha^{*\intercal} X_k^* \beta^* + \gamma^{*\intercal} z_k)}{1 + \exp(\alpha^{*\intercal} X_k^* \beta^* + \gamma^{*\intercal} z_k)}; \tag{2.10}$$

the dependence on z_k is suppressed in the notation $p_k(\theta^*; X_k^*)$ for ease of exposition. In Appendix A.3, we show the following result.

Theorem 2.1 *Assume that Conditions (C.1), (C.3), (C.4) in Appendix A.1 hold and that* $\min(m,n) \to \infty$. *Then*

$$\hat{\theta}^* - \theta = o_p(1).$$

While Theorem 2.1 shows that $\hat{\theta}^*$ is a consistent estimator of θ under certain situations, this result does not suggest that the naive method of ignoring measurement error is a valid and practical procedure in applications. The requirement $\min(m,n) \to \infty$ in Theorem 2.1 essentially says that measurement error in matrix-variates virtually becomes null because the covariance matrix for $\mathrm{vec}(\bar{U}_k)$ in (2.5) approaches a zero matrix. In such an instance, it is not surprising that the estimator $\hat{\theta}^*$ would be a consistent estimator for θ. As Theorem 2.1 establishes the asymptotic difference of $\hat{\theta}^* - \theta$ when both m and n approach infinity, to complement this result, it is interesting to examine for given m and n, what quantities would dominate the difference $\hat{\theta}^* - \theta$. Such an exploration allows us to develop estimators of correcting for measurement error effects for settings with a given m, and thus establish their asymptotic distributions if only n approaches infinity. In the next subsection, we explore this problem.

2.2.3 Refined Expressions for Bias

Let $v_{1,k}(\cdot) = p_k(\cdot)\{1 - p_k(\cdot)\}$ and $v_{2,k}(\cdot) = p_k(\cdot)\{1 - p_k(\cdot)\}\{1 - 2p_k(\cdot)\}$ where $p_k(\cdot)$ is defined by (2.10). Define $\mathrm{S}_n = n^{1/2}\partial \ell_n(\alpha, \beta, \gamma)/\partial\theta$, where $\ell_n(\alpha, \beta, \gamma)$ is given by (2.3). Motivated by Stefanski and Carroll (1985, p.1339), we consider the following terms, each corresponding to parameter α, β or γ:

$$\mathrm{J}_{\alpha,n,1} = -(1/2n)\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{c}k}\beta \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}(\Omega_0/m_c)v_{2,k}(\theta; x_{\mathrm{c}k}),$$

$$\mathrm{J}_{\beta,n,1} = -(1/2n)\sum_{k=1}^{n} x_{\mathrm{c}k}^{\intercal}\alpha \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}(\Omega_0/m_c)v_{2,k}(\theta; x_{\mathrm{c}k}),$$

$$\mathrm{J}_{\alpha,n,2} = -(1/n)\sum_{k=1}^{n} C_t^{\intercal}\Pi_\alpha \times (\Omega_0/m_c)v_{1,k}(\theta; x_{\mathrm{c}k}),$$

$$\mathrm{J}_{\beta,n,2} = -(1/n)\sum_{k=1}^{n}\Pi_{\beta}(\Omega_0/m_c)v_{1,k}(\theta; x_{\mathrm{c}k}),$$

$$\mathrm{J}_{\gamma,n} = -(1/2n) \times \sum_{k=1}^{n} z_k \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}(\Omega_0/m_c)v_{2,k}(\theta; x_{\mathrm{c}k}),$$

where $\Pi_{\alpha} = \begin{bmatrix}\beta_1 I_{(p+1)} & \beta_2 I_{(p+1)} & \cdots & \beta_q I_{(p+1)}\end{bmatrix}$ is a $(p+1)\times\{(p+1)q\}$ matrix, and Π_{β} is a $q \times \{(p+1)q\}$ block matrix with $\alpha^{\intercal}$ being the diagonal block vectors and zero elsewhere. Let $\mathrm{J}_{\alpha,n} = \mathrm{J}_{\alpha,n,1} + \mathrm{J}_{\alpha,n,2}$, $\mathrm{J}_{\beta,n} = \mathrm{J}_{\beta,n,1} + \mathrm{J}_{\beta,n,2}$, and $\mathrm{J}_n(\theta) = (\mathrm{J}^{\intercal}_{\alpha,n}, \mathrm{J}^{\intercal}_{\beta,n}, \mathrm{J}^{\intercal}_{\gamma,n})^{\intercal}$. Write $\mathrm{H}_n(\theta) = -\partial^2\ell_n(\theta)/\partial\theta\partial\theta^{\intercal}$.

Theorem 2.2 *Under Conditions (C.1)-(C.3), and (C.6) in Appendix A.1, we have that*

$$\hat{\theta}^* - \theta = (1/n^{1/2})\mathrm{H}_n^{-1}(\theta)\mathrm{S}_n(\theta) + \mathrm{H}_n^{-1}(\theta)\mathrm{J}_n(\theta)\mathrm{vec}(\alpha\beta^{\intercal}) + o_p\{\max(1/m, 1/n^{1/2})\}. \quad (2.11)$$

Expression (2.11) shows that the asymptotic bias of $\hat{\theta}^*$ involves the terms pertinent to $\mathrm{H}_n(\theta)$, $\mathrm{S}_n(\theta)$, $\mathrm{J}_n(\theta)$ as well as the values of n and m. Stefanski and Carroll (1985, Theorem 1) showed that under regularity conditions, $\mathrm{H}_n^{-1/2}(\theta)\mathrm{S}_n(\theta)$ asymptotically follows a normal distribution with mean zero and an identity covariance matrix, and hence, yielding that $\mathrm{H}_n^{-1}(\theta)\mathrm{S}_n(\theta)$ in (2.11) has an asymptotic normal distribution with mean zero and covariance matrix $I^{-1}(\theta)$, where $I(\theta) = \mathrm{E}\{\mathrm{H}_n(\theta)\}$. Consequently, (2.11) implies that with m of an order $O(\sqrt{n})$, $n^{1/2}\{\hat{\theta}^* - \theta - \mathrm{H}_n^{-1}(\theta)\mathrm{J}_n(\theta)\mathrm{vec}(\alpha\beta^{\intercal})\}$ has an asymptotic normal distribution with mean zero and covariance matrix $I^{-1}(\theta)$ as $n \to \infty$.

The proof of Theorem 2.2 begins with applying the first-order Taylor series expansion to $\mathrm{S}_n^*(\hat{\theta}^*) = 0$ around θ with X_k^* and z_k fixed:

$$\hat{\theta}^* = \theta + \mathrm{H}_n^{*-1}(\theta)\mathrm{S}_n^*(\theta) + o_p\{\max(1/m, 1/n^{1/2})\}, \quad (2.12)$$

where $\mathrm{S}_n^*(\theta)$ is determined by (2.9) with θ^* replaced by θ, and $\mathrm{H}_n^*(\theta) = -\partial\mathrm{S}_n^*(\theta)/\partial\theta^{\intercal}$. The details are given in Appendix A.6.

(2.12) expresses the relationship between $\hat{\theta}^*$ and θ using the surrogate observations X_k^* for $k = 1, \ldots n$, together with the covariate z_k and the response variable Y_k. To obtain (2.11) expressed in terms of the true covariate $x_{\mathrm{c}k}$, we need only to examine $\mathrm{S}_n^*(\theta)$ and $\mathrm{H}_n^*(\theta)$ using their counterparts based on the true covariates $x_{\mathrm{c}k}$ together with $\{z_k, Y_k\}$, which is summarized in the following lemmas whose proofs are placed in Appendices A.4 and A.5.

Lemma 1 *Let* $\mathrm{Z}_n(\theta) = (1/n^{1/2})\mathrm{S}_n(\theta) + \mathrm{J}_n(\theta)\mathrm{vec}(\alpha\beta^{\mathsf{T}})$ *which depends on the true covariates* $x_{\mathrm{c}k}$ *as well as* $\{z_k, Y_k\}$*. Under Conditions (C.1) and (C.3) in Appendix A.1, we have that*

$$\mathrm{S}_n^*(\theta) = \mathrm{Z}_n(\theta) + o_p\{\max(1/m, 1/n^{1/2})\}. \tag{2.13}$$

Lemma 2 *Under Conditions (C.1)-(C.3) in Appendix A.1, we have that as* $\min(m, n) \to \infty$*,*

$$\mathrm{H}_n^*(\theta) = \mathrm{H}_n(\theta) + o_p(1). \tag{2.14}$$

2.3 Corrections for Measurement Error Effects

In this section, we describe two methods of correcting for measurement error effects on parameter estimation.

2.3.1 Moment-Based Correction Method

Noticing that $(1/n^{1/2})\mathrm{H}_n^{-1}(\theta)\mathrm{S}_n(\theta)$ approaches 0 in probability as $n \to \infty$ as discussed earlier, we are motivated by Theorem 2.2 to consider

$$\hat{\theta}_c^* = \hat{\theta}^* - \hat{\mathrm{H}}_n^{-1}(\hat{\theta}^*)\hat{\mathrm{J}}_n(\hat{\theta}^*)\mathrm{vec}(\hat{\alpha}^*\hat{\beta}^{*\mathsf{T}}), \tag{2.15}$$

where $\hat{\mathrm{J}}_n(\hat{\theta}^*)$ and $\hat{\mathrm{H}}_n(\hat{\theta}^*)$ correspond to $\mathrm{J}_n(\theta)$ and $\mathrm{H}_n(\theta)$ with $x_{\mathrm{c}k}$, θ and Ω_0, respectively, replaced by X_k^*, $\hat{\theta}^*$ and $\hat{\Omega}$ for $k = 1, ..., n$, with $\hat{\Omega}$ representing an estimator of Ω_0. In Appendix A.7, we show the following asymptotic properties of $\hat{\theta}_c^*$.

Theorem 2.3 *Suppose that Conditions (C.1)-(C.3) and (C.5)-(C.6) in Appendix A.1 hold. Assume that as* $n \to \infty$*,*

$$n^{1/2}(\hat{\Omega} - \Omega_0) = O_p(1). \tag{2.16}$$

Then as $n \to \infty$*,*

(a) $\hat{\theta}_c^* \xrightarrow{p} \theta$;

(b) $n^{1/2}(\hat{\theta}_c^* - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$.

The consistency of $\hat{\theta}_c^*$ requires the consistency assumption (2.16) for the estimator $\hat{\Omega}$ of the covariance matrix Ω_0. With the replicates X_{kr}^* following model (2.4), a $\sqrt{n}$-consistent covariance estimator is given by:

$$\hat{\Omega} = \left[\sum_{k=1}^{n} \sum_{r=1}^{m} \{\mathrm{vec}(X_{kr}^*) - \mathrm{vec}(\bar{X}_{k+}^*)\}\{\mathrm{vec}(X_{kr}^*) - \mathrm{vec}(\bar{X}_{k+}^*)\}^\intercal \right] \Big/ \{n(m-1)\}; \tag{2.17}$$

the proof is presented in Appendix A.8.

2.3.2 Sufficient Statistic Correction Method

Except for the requirement of zero mean of E_{kr} and the additive structure (2.4), the construction of the estimator $\hat{\theta}_c^*$ has the advantage of not requiring the specification of the full distribution for E_{kr}. However, when p and/or q are large, the calculation of $\hat{\theta}_c^*$ may be time-consuming due to the involvement of the large dimensional covariance matrix Ω_0, and the resulting estimator may not be accurate due to a small sample size n relative to the dimension $\{(p+1)q\} \times \{(p+1)q\}$ of Ω_0. Driven by these issues, we explore another estimator which capitalizes on imposing the normality distributional form of E_{kr}.

Specifically, for $k = 1, ..., n$ and $r = 1, ..., m$, we assume that E_{kr} follows a matrix normal distribution with $E_{kr} \sim MN(0_{(p+1)\times q}, R, C)$, where $MN(\cdot,\cdot,\cdot)$ represents a matrix normal distribution, R represents the $(p+1) \times (p+1)$ *row* covariance matrix, and C stands for the $q \times q$ *column* covariance matrix, respectively (Hoff 2011). Equivalently, $\mathrm{vec}(E_{kr}) \sim N_{(p+1)q}(\mathrm{vec}(0_{(p+1)\times q}), \Omega_0)$, where $\Omega_0 = C \otimes R$, where $\otimes$ denotes the *Kronecker product* (Dutilleu 1999). By (2.5), the observed matrix-variate X_k^* follows a matrix normal distribution as well, i.e., $\mathrm{vec}(X_k^*) \sim N_{(p+1)q}(\mathrm{vec}(x_{\mathrm{c}k}), \Omega_0/m_c)$, where $k = 1, ..., n$.

Under the assumption that E_{kr} is independent of Y_k and $\{x_k, z_k\}$ for $k = 1, ...n$, we have that given $\{x_{\mathrm{c}k}, z_k\}$, the joint distribution of Y_k and X_k^* can be written as

$$\begin{aligned} f_{\mathrm{Y},\mathrm{X}^*}(Y_k, X_k^* \mid x_{\mathrm{c}k}, z_k, \theta^*) &= f_{\mathrm{Y}}(Y_k \mid X_k^*, x_{\mathrm{c}k}, z_k, \theta^*) \times f_{\mathrm{X}^*}(X_k^* \mid x_{\mathrm{c}k}, z_k) \\ &= f_{\mathrm{Y}}(Y_k \mid x_{\mathrm{c}k}, z_k, \theta) \times f_{\mathrm{X}^*}(X_k^* \mid x_{\mathrm{c}k}), \end{aligned} \tag{2.18}$$

where $f_{\mathrm{Y}}(Y_k \mid x_{\mathrm{c}k}, z_k, \theta)$ is determined by (2.2) and $f_{\mathrm{X}^*}(X_k^* \mid x_{\mathrm{c}k})$ is determined by (2.5).

With θ treated as a given constant and $x_{\mathrm{c}k}$ regarded as an unknown parameter, using the formulation (2.18), we derive sufficient statistics for the $x_{\mathrm{c}k}$, given by

$$\Delta_k = X_k^* + (Y_k - 1/2)R\alpha\beta^\intercal C/m_c; \tag{2.19}$$

the details are included in Appendix A.9. The availability of such sufficient statistics allows us to find a conditional probability to carry out inference about θ in the absence of the $x_{\mathrm{c}k}$. To be specific, given Δ_k and z_k, the conditional probability of Y_k is

$$P(Y_k = 1 \mid \Delta_k, z_k) = \exp[(\eta^*_{\Delta_k} + \gamma^\intercal z_k) - \log\{1 + \exp(\eta^*_{\Delta_k} + \gamma^\intercal z_k)\}], \tag{2.20}$$

where $\eta^*_{\Delta_k} = \alpha^\intercal \Delta_k \beta$. Working with the conditional distribution (2.20) yields the likelihood score equation

$$\sum_{k=1}^{n} \{Y_k - P(Y_k = 1 \mid \Delta_k, z_k)\} \begin{pmatrix} \tilde{\Delta}_k^\intercal(\theta) \\ z_k \end{pmatrix} = 0, \tag{2.21}$$

where $\tilde{\Delta}_k(\theta) = (\beta^\intercal \Delta_k^\intercal C_t, \alpha^\intercal \Delta_k)^\intercal$.

Although the derivation of (2.21) is conceptually straightforward with the availability of the conditional probability (2.20), equation (2.21) cannot be directly used for finding a consistent estimator since it may produce multiple solutions which are not necessarily all consistent, as pointed out by Stefanski and Carroll (1985, p.1341). As an alternative, we maximize (2.3) with x_k replaced by $\hat{\Delta}_k$ and obtain an estimator of θ, denoted as $\hat{\theta}^*_s$, where $\hat{\Delta}_k$ is determined by (2.19) with R, C, α, β replaced by $\hat{R}$, $\hat{C}$, $\hat{\alpha}^*$ and $\hat{\beta}^*$, respectively, with $\hat{R}$ and $\hat{C}$ being the estimators of R and C, i.e.,

$$\hat{\Delta}_k = X^*_k + g_k/m_c, \tag{2.22}$$

with $g_k = (Y_k - 1/2)\hat{R}\hat{\alpha}^*\hat{\beta}^{\intercal *}\hat{C}$ for $k = 1, ..., n$.

To obtain the estimator $\hat{\theta}^*_s$, we need to estimate the unknown *row* and *column* covariance matrices R and C, which can be done using the *flip-flop* algorithm (Dutilleu 1999). This algorithm basically applies maximum likelihood estimation to estimate R and C one at a time iteratively to yield $\sqrt{n}-$consistent estimators, where the matrix normality assumption is typically imposed on E_{kr} to allow for manageable computation. Furthermore, we comment that the normality assumption for E_{kr} is needed in the derivation of the estimator $\hat{\theta}^*_s$. This assumption enables us to work out sufficient statistics (2.19) for the x_{ck}, as shown in Appendix A.9. In Appendix A.10, we show that following theorem.

Theorem 2.4 *Suppose that Conditions (C.1)-(C.3) and (C.5)-(C.6) in Appendix A.1 hold. Assume that as $n \to \infty$, $n^{1/2}(\hat{C} \otimes \hat{R} - \Omega_0) = O_p(1)$ and $\sum_{k=1}^{n} \|g_k\|^2 = O_p(n)$. Then we have that*

$$\hat{\theta}^*_s = \theta + (1/n^{1/2})\mathrm{H}_n^{-1}(\theta)\mathrm{S}_n(\theta) + o_p\{\max(1/m, 1/n^{1/2})\}, \tag{2.23}$$

and hence, with m of order $O(\sqrt{n})$, the following results:

(a) $\hat{\theta}_s^* \xrightarrow{p} \theta$ *as* $n \to \infty$;

(b) $n^{1/2}(\hat{\theta}_s^* - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$ *as* $n \to \infty$.

Theorems 2.3 and 2.4 show that both $\hat{\theta}_c^*$ and $\hat{\theta}_s^*$ are consistent and have an asymptotic normal distribution. The results offer us two estimators of θ under different assumptions of the measurement error model. Estimator $\hat{\theta}_c^*$ is more robust than $\hat{\theta}_s^*$ since it does not require the normality assumption for the measurement error term E_{kr}. Theorem 2.4 also offers a rigorous justification of the validity of $\hat{\theta}_s^*$ which was heuristically derived by Li's PhD thesis work.

2.4 Numerical Studies

2.4.1 Simulation Design

We now evaluate the performance of the proposed methods under different settings via simulation studies where we consider settings with $p + 1 = q$, denoted as p_x for ease of exposition. We also demonstrate the impacts of the naive analysis which ignores measurement error. The sample size is set as $n = 1000$ when $p_x = 5$, 10 or 20, and $n = 2000$ when $p_x = 20$. We consider the case with $m = 2$, 5 or 10. Five hundred simulations are run for each setting.

For $k = 1, ..., n$, we simulate $p_x \times p_x$ matrix-variate data, x_k, from the matrix normal distribution $MN(0, \sigma_x^2 I_{p_x}, I_{p_x})$, where σ_x^2 is set as 1.0. The z_k covariates are independently generated from the standard normal distribution. For $k = 1, ..., n$, the binary response Y_k is randomly generated from the Bernoulli distribution with the probability $P(Y_k = 1 \mid x_{ck}, z_k; \alpha, \beta, \gamma) = \exp(\alpha^\intercal x_{ck}\beta + \gamma z_k)/\{1 + \exp(\alpha^\intercal x_{ck}\beta + \gamma z_k)\}$, where $\gamma = 0.5$. When $p_x = 5$, we set $\alpha = (0.5, 1, -1, -1, 1)^\intercal$ and $\beta = (1, 0.5, 1, -1, -1)^\intercal$; when $p_x = 10$, we take $\alpha = (0.5, 1, -1 \times 1_4^\intercal, 1 \times 1_4^\intercal)^\intercal$ and $\beta = (1, 0.5, 1, -1, -1, 1, 0.5, 1, -1, -1)^\intercal$; when $p_x = 20$, we take $\alpha = (0.5, 1, -0.5 \times 1_4^\intercal, 0.5 \times 1_6^\intercal, -0.5 \times 1_4^\intercal, 0.5 \times 1_4^\intercal)^\intercal$ and $\beta = (0.5 \times 1_3^\intercal, -0.5 \times 1_2^\intercal, 0.5 \times 1_3^\intercal, -0.5 \times 1_2^\intercal, 0.5 \times 1_3^\intercal, -0.5 \times 1_2^\intercal, 0.5 \times 1_3^\intercal, -0.5 \times 1_2^\intercal)^\intercal$, where 1_d represents a $d \times 1$ unit vector. For $k = 1, ..., n$, repeated surrogate measurements X_{kr}^* are generated from model (3.1), where the E_{kr} are independently generated from the matrix normal distribution, $MN(0, \sigma^2 I_{p_x}, I_{p_x})$ for $r = 1, ..., m$. We let $\sigma = 0.25, 0.5, 0.75$ to feature small, moderate and large measurement error, which lead to the signal-to-noise ratio σ_x^2/σ^2 for each covariate component to be 16, 4, and 1.778, respectively.

We estimate the model parameters α and β using different methods. The first analysis is to use the naive approach which fits the data with model (2.7) using the block relaxation algorithm indicated in Table 2.1. To correct for measurement error effects, we conduct two analyses, respectively, called Methods 1 and 2, by using (2.15) and the estimator based on (2.23), respectively.

To use these methods, the covariance matrix Ω_0 for the measurement error model needs to be estimated. Since the sample size is not large enough relative to the dimension of the covaraince matrix, the sample covariance matrix $\hat{\Omega}$ may be poorly estimated and may not be invertible. To obtain a stable covaraince estimator of Ω_0, we apply the method of Ledoit and Wolf (2004) which uses a linear combination of the sample covariance matrix $\hat{\Omega}$ and the identity matrix I_{p_x} to obtain an adjusted covariance matrix $\hat{\Omega}_L$.

Specifically, for a given $k = 1, .., n$ and a given $r = 1, ..., m$, we first vectorize the p_x matrix X^*_{kr} to create a $p_x^2 \times 1$ column vector $\text{vec}(X^*_{kr})$. Next, we define a $p_x^2 \times M$ matrix, X_M, by arranging the vectors $\text{vec}(X^*_{kr})$ as its columns according to the order from $\text{vec}(X^*_{11})$ to $\text{vec}(X^*_{nm})$, where $M = nm$. Then, we calculate the sample covariance matrix $\hat{\Omega} = M^{-1} X_M X_M^\intercal$ and $r_M = \text{tr}(\hat{\Omega} I_{p_x}^\intercal)/p_x^2$, where $\text{tr}(\cdot)$ is the trace of a matrix. Furthermore, we calculate $d_M^2 = \|\hat{\Omega} - r_M I_{p_x}\|^2$ and $\text{C}_M^2 = \min(\bar{b}_M^2, d_M^2)$, where $\bar{b}_M^2 = (1/M) \sum_{i=1}^M \|x_{.i}^M (x_{.i}^M)^\intercal - \hat{\Omega}\|^2$ with $\|\cdot\|$ being the Frobenius Norm, and $x_{.i}^M$ represents the ith column of X_M for $i = 1, ..., M$. Finally, we consider the linear combination $\hat{\Omega}_L = a\hat{\Omega} + (1-a)I_{p_x}$ of $\hat{\Omega}$ and the identity matrix I_{p_x} with a given by C_M^2/d_M^2. Such $\hat{\Omega}_L$ is a $\sqrt{n}$-consistent covariance estimator (Ledoit and Wolf 2004, Theorem 3.4); its calculation can be realized using a Matlab function available at http://www.econ.uzh.ch/en/people/faculty/wolf/publications.html#9.

2.4.2 Simulation Results

We summarize the simulation results in the terms of the finite sample relative biases in percent (bias%), empirical standard errors (ESE), model-based asymptotic standard errors (ASE), and mean squared errors (MSE) as well as the coverage rates in percent (CR%) for 95% confidence intervals. Here bias% is defined as the ratio of the difference between the true parameter value and the average of the estimates obtained from all simulation runs to the true parameter value; ESE is defined to be the sample standard error of the estimates obtained from all simulation runs; MSE represents the average of the squared differences between the estimates and the true parameter value obtained from all simulation runs; ASE is calculated as the square root of the average of all the estimates of the asymptotic variance in all the simulation runs; CR% is the coverage rate for 95% confidence intervals for all the simulations.

Table 2.2 includes the results for the *row* and *column* effects for the cases with matrix-variate with small, moderate and severe measurement error when $p_x = 5$ where only the results for α_1, α_4, β_1, β_5, and γ are included to save space. Complete results for this case are placed in Tables A.1-A.2 in Appendices. It is seen that measurement error effects on estimating the *row* parameters α are not as striking as those for estimating, the *column* parameters β and the covariate parameters, γ. The performance of the naive method is influenced dramatically by the degree of measurement error. The naive method produces noticeable finite sample biases and the bias increases as the degree of measurement error increases. On the other hand, Methods 1 and 2 significantly improve the performance of the naive method, and the improvement is clearly noticeable for cases with not severe measurement error or a good number of replicates. Mean squared errors of the naive estimators are higher than those of the proposed methods, especially when measurement errors is not minor. Not surprisingly, the performance of the proposed methods deteriorates as measurement error becomes substantial, especially in combination with decreasing the number of replicates. This phenomenon is clearly indicated by the coverage rates of 95% confidence intervals.

In Tables A.3-A.10 and A.13-A.14 in Appendices, we respectively report the simulation results for the cases with $p_x = 10$ and $p_x = 20$. We observe patterns similar to those for the case $p_x = 5$, but the magnitudes of the finite sample biases and standard errors are larger than those with $p_x = 5$. As p_x becomes larger, the performance of the three methods tends to be more sensitive to the increase of measurement error and the number of replicates. Unsurprisingly, with a given sample size, the performance of the three methods deteriorates as p_x increases. With a given p_x, the two proposed methods tend to produce more accurate results as the sample size increases, which is evident from the results in Table A.10 and Table A.14 for $p_x = 20$ and $n = 1000$ and 2000, respectively.

2.4.3 Sensitivity Analysis of the Proposed Methods

In Sections 2.4.1-2.4.2, we conduct simulations to (1) demonstrate that the naive analysis ignoring the feature of measurement error can lead to seriously biased results, and (2) confirm the good performance of the two proposed methods. Our assessment is carried out for the case where surrogate measurements are generated from model (2.4) with the error term E_{kr} assumed to be normal, an assumption that is required by Method 2. Now we further assess how sensitive the performance of Method 2 is to the violation of the normality assumption for E_{kr}. In comparison, we also report results obtained from Method 1.

Specifically, we conduct a simulation study where E_{kr} is generated from a matrix t-distribution, $E_{kr} \sim T(\nu, W, R, C)$. Here $T(\cdot,\cdot,\cdot,\cdot)$ represents a matrix t-distribution, W is the $p_x \times p_x$ location matrix, R and C respectively represent the $p_x \times p_x$ *row* scale and the $p_x \times p_x$ *column* scale matrices, and ν is the degrees of freedom. This matrix t-distribution yields that $\Omega_0 = (C \otimes R)/(\nu - 2)$ for $\nu > 2$. We consider the setting $W = 0_{p_x \times p_x}$, $C = \sigma^2 I_{p_x}$, and $R = I_{p_x}$ together with $p_x = 5$ and $\nu = 3$, where $0_{p_x \times p_x}$ is the $p_x \times p_x$ zero matrix. Other settings for simulating data are the same as those in Section 2.4.1. We apply the naive approach and the two proposed methods to analyze the simulated data.

The simulation results are reported in Tables A.11-A.12 in Appendices to save space. It is clear that the naive method still produces biased results with patterns similar to those observed in Section 2.4.2. Method 1 is not sensitive to the change of the distribution of measurement error and its performance under the current setting is similar to that in the setting of Section 2.4.1. However, the performance of Method 2 greatly decays. The estimates of the *row* parameters α have large finite sample biases when the number of replicates is small. For the *column* parameters β and vector covariate parameters γ, Method 2 provides a lot larger finite sample biases than Method 1, especially when measurement error is large with a small number of replicates. Such findings are not surprising, because Method 2 is derived based on the model assumption (2.4) with the measurement error following a matrix normal distribution.

In summary, the naive method yields biased results when measurement error is not mild. Thus, it is imperative to accommodate measurement error effects in order to carry out valid inferences. The simulation studies confirm that the proposed methods significantly improve the performance of the naive method, and their performance is reasonably satisfactory for various settings. As described in Section 4, Method 1 is more robust than Method 2 since it does not impose a distributional assumption on the error terms E_{kr}. In applications, Method 1 is generally recommended if we are not certain about the feasibility of a normally distributed measurement error assumption.

2.4.4 Data Analysis

We apply the two correction methods, in contrast to the naive approach, to analyze the EEG imaging data which are available at the UCI Machine Learning Repository website (`http://archive.ics.uci.edu/ml/datasets/EEG+Database`). The EEG data include the measurements of 122 subjects who were selected from those exposed to one stimulus experiment. During this experiment, the voltage values were recorded from 64 channels of electrodes at 256 time points (in one second). Those 122 subjects are differentiated by being in the alcoholic group with 77 patients or the control group with 45 patients. The research interest was to make classification between the alcoholic group and the control group based on voltage values which are subject to measurement error over times and channels.

For $k = 1, ..., 122$, let Y_k be the binary response variable for subject k, with 1 for being in the alcoholic group and 0 for being in the control group; the matrix-variate of subject k, denoted as X_k^{**}, is a 256×64 matrix with each entry representing the mean voltage value of r replicates for the corresponding time point and channel, where $r = 1, ..., m_k$, and m_k is the number of replicates for subject k, ranging from 7 to 60 with an average 45.

Without considering issues of measurement error, Hung and Wang (2013) applied the matrix-variate logistic regression model (2.2) to fit the EEG data set which includes $256 + 64 + 1 = 321$ parameters. This modeling greatly reduces the number of parameters which would be $256 \times 64 + 1 = 16285$ if using model (2.1).

While using model (2.2) can significantly reduce the dimension of parameters compared to using model (2.1), we still cannot directly employ model (2.2) to fit the data here because the sample size is 122, smaller than the dimension of the model parameters. As a result, we have to first reduce the dimension of the matrix-variate X_k^{**} before fitting the model (2.2).

Motivated by the simulation findings that the response model parameters can be well estimated when the sample size is 10 times larger than the number of parameters, here we reduce the initial 256×64 matrix-variate X_k^{**} to a 5×5 matrix-variate X_k^* for $k = 1, ..., 122$ using the two-directional two-dimensional principal component analysis ($(2\mathrm{D})^2$PCA) method of Zhang and Zhou (2005).

We assume that X_k^* is an observed version of the true matrix-variate X_k and they are linked by (2.4) with the measurement error covariance matrix Ω_0 estimated using the method of Ledoit and Wolf (2004), as described in Section 2.4.1, where p_x is taken as 5, and the sample variance matrix $\hat{\Omega}$ is obtained using $\mathrm{vec}(X_k^*)$ across all the subjects using the total $M = 5486$ replicates of $n = 122$ subjects; this is needed for obtaining the estimator (2.15). The flip-flop algorithm is applied to X_k^* to find the *row* and *column* matrices, $\hat{R}$ and $\hat{C}$ for the sufficient statistics correction method given by (2.22). Consistent with Hung and Wang (2013), we set the second *row* parameter as 1 because the second *row* of X_k^* has the highest correlation with the response.

Table 2.3 reports the estimation results for the EEG data by fitting the model (2.2) with the 5×5 matrix-variate X_k^* using the two methods described in Sections 2.3.1 and 2.3.2, together with the naive method which ignores measurement errors. The two correction methods output very similar results. While the estimates for the channel parameters (i.e., row parameters) produced from the naive analysis are noticeably different from those obtained from the two correction methods, the estimates for the time points (i.e., column parameters) yielded from the naive method are quite similar to those given by the two correction methods. All the three methods reveal the same evidence for the column and row parameters. For the *column* parameters, time points one, two, three and four are detected to be significant. For the row parameters, the third channel has a significant effect on distinguishing the alcoholic and nonalcoholic status. Finally, we report that the computation times for Methods 1 and 2 are 0.409 and 0.386 seconds, respectively, using a PC equipped with 2.6 GHz Intel Core i5 CPU and 16 GB RAM.

Table 2.2: Simulation results for the *row* parameters with $p_x = 5$, E_{kr} is generated from the matrix normal distribution, and $n = 1000$

	σ	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
Parameter			Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
α_1	0.25	2	1.192	0.062	0.063	94.4	0.004	1.232	0.063	0.062	93.8	0.004	1.206	0.063	0.061	89.6	0.004
		5	1.423	0.060	0.060	94.4	0.004	1.434	0.060	0.060	94.0	0.004	1.437	0.060	0.060	92.4	0.004
		10	1.442	0.060	0.060	93.8	0.004	1.448	0.060	0.060	93.8	0.004	1.442	0.060	0.059	93.0	0.004
	0.5	2	0.987	0.071	0.072	94.4	0.005	1.109	0.074	0.069	92.4	0.006	1.013	0.073	0.067	92.0	0.005
		5	1.398	0.064	0.064	95.2	0.004	1.424	0.064	0.063	94.2	0.004	1.444	0.064	0.062	94.6	0.004
		10	1.495	0.063	0.062	94.6	0.004	1.529	0.063	0.061	94.2	0.004	1.501	0.063	0.061	94.4	0.004
	0.75	2	0.864	0.083	0.083	94.2	0.007	0.969	0.088	0.077	91.0	0.008	0.849	0.086	0.074	90.8	0.007
		5	1.385	0.069	0.070	95.0	0.005	1.412	0.071	0.068	94.6	0.005	1.450	0.071	0.066	94.0	0.005
		10	1.576	0.066	0.065	94.6	0.004	1.669	0.067	0.064	94.4	0.005	1.599	0.067	0.063	94.0	0.005
α_4	0.25	2	0.500	0.079	0.079	95.4	0.006	0.517	0.080	0.078	94.6	0.006	0.497	0.080	0.077	91.6	0.006
		5	0.232	0.077	0.076	94.2	0.006	0.225	0.077	0.076	93.6	0.006	0.223	0.077	0.075	92.4	0.006
		10	0.330	0.077	0.075	94.8	0.006	0.329	0.077	0.075	94.6	0.006	0.326	0.077	0.075	91.6	0.006
	0.5	2	0.796	0.090	0.091	94.6	0.008	0.948	0.093	0.087	93.2	0.009	0.910	0.092	0.085	92.8	0.009
		5	0.216	0.082	0.081	94.2	0.007	0.177	0.083	0.080	93.4	0.007	0.180	0.083	0.078	92.6	0.007
		10	0.409	0.081	0.078	94.6	0.006	0.410	0.081	0.077	94.0	0.007	0.399	0.081	0.077	93.4	0.007
	0.75	2	1.105	0.106	0.106	95.6	0.011	1.318	0.112	0.098	92.0	0.013	1.337	0.109	0.095	91.6	0.012
		5	0.253	0.088	0.089	94.2	0.008	0.151	0.091	0.085	92.2	0.008	0.190	0.092	0.084	91.2	0.008
		10	0.511	0.085	0.083	95.4	0.007	0.518	0.087	0.081	94.4	0.008	0.501	0.087	0.079	93.6	0.008
β_1	0.25	2	-9.082	0.089	0.086	76.0	0.016	1.021	0.107	0.104	94.3	0.012	1.345	0.108	0.097	93.0	0.012
		5	-2.766	0.098	0.092	89.5	0.010	2.016	0.106	0.100	93.0	0.012	2.090	0.106	0.097	91.8	0.012
		10	-0.656	0.098	0.093	94.0	0.010	1.866	0.103	0.097	93.8	0.011	1.891	0.103	0.096	93.0	0.011
	0.5	2	-30.079	0.071	0.070	2.4	0.095	-8.581	0.106	0.108	83.8	0.019	-7.138	0.111	0.094	80.0	0.017
		5	-14.297	0.087	0.082	53.8	0.028	-0.299	0.112	0.107	93.6	0.013	0.305	0.114	0.098	90.0	0.013
		10	-7.178	0.093	0.088	80.6	0.014	1.318	0.108	0.102	94.8	0.012	1.553	0.109	0.097	93.4	0.012
	0.75	2	-48.331	0.057	0.056	0.0	0.237	-24.022	0.095	0.096	29.6	0.067	-22.216	0.102	0.081	27.8	0.060
		5	-27.863	0.075	0.071	4.0	0.083	-7.012	0.111	0.107	87.0	0.017	-5.632	0.116	0.094	82.6	0.017
		10	-16.210	0.085	0.080	49.8	0.033	-1.343	0.111	0.106	93.6	0.013	-0.605	0.114	0.096	90.4	0.013
β_5	0.25	2	-9.053	0.091	0.086	76.3	0.016	1.099	0.108	0.103	93.0	0.011	1.435	0.109	0.097	91.0	0.012
		5	-2.578	0.097	0.092	90.0	0.010	2.227	0.106	0.100	95.5	0.011	2.294	0.106	0.097	94.0	0.012
		10	-0.552	0.098	0.093	93.3	0.009	1.977	0.102	0.098	94.0	0.011	2.001	0.102	0.096	94.0	0.011
	0.5	2	-30.188	0.070	0.070	2.6	0.096	-8.621	0.105	0.108	83.4	0.018	-7.101	0.110	0.094	80.2	0.017
		5	-14.281	0.087	0.082	54.2	0.028	-0.239	0.112	0.107	93.4	0.013	0.369	0.114	0.098	91.2	0.013
		10	-7.259	0.091	0.088	81.2	0.014	1.254	0.106	0.102	94.8	0.011	1.481	0.107	0.097	92.6	0.012
	0.75	2	-48.523	0.056	0.056	0.0	0.058	-24.226	0.093	0.096	26.2	0.067	-22.379	0.100	0.081	24.2	0.060
		5	-27.888	0.076	0.071	4.4	0.084	-6.986	0.113	0.107	87.6	0.018	-5.598	0.118	0.094	83.8	0.017
		10	-16.300	0.084	0.081	44.2	0.034	-1.398	0.109	0.108	94.6	0.012	-0.680	0.111	0.097	91.6	0.012
γ	0.25	2	-5.747	0.110	0.109	94.3	0.013	1.877	0.121	0.121	95.8	0.015	2.285	0.122	0.120	95.0	0.015
		5	-1.171	0.114	0.113	94.3	0.013	2.454	0.119	0.118	93.8	0.014	2.555	0.119	0.117	93.5	0.014
		10	0.363	0.115	0.114	94.5	0.013	2.277	0.118	0.117	93.5	0.013	2.306	0.118	0.116	93.5	0.013
	0.5	2	-20.889	0.099	0.099	80.2	0.021	-5.330	0.124	0.127	95.6	0.016	-3.505	0.129	0.125	94.4	0.017
		5	-9.686	0.106	0.107	91.0	0.014	0.655	0.121	0.123	95.0	0.015	1.363	0.123	0.122	94.4	0.015
		10	-4.921	0.109	0.110	94.4	0.012	1.381	0.117	0.120	94.2	0.014	1.603	0.118	0.119	94.0	0.014
	0.75	2	-33.506	0.089	0.090	53.0	0.036	-16.864	0.118	0.127	91.0	0.021	-14.352	0.126	0.123	90.2	0.021
		5	-19.413	0.099	0.100	83.6	0.019	-4.445	0.123	0.126	94.4	0.016	-2.740	0.127	0.124	93.8	0.016
		10	-11.602	0.104	0.106	91.6	0.014	-0.771	0.120	0.124	94.6	0.014	-0.082	0.120	0.122	94.0	0.015

Table 2.3: Analysis results for the EEG data using the three methods

Parameter	Naive Method			Method 1			Method 2		
	Est.	SE	95% CI	Est.	SE	95% CI	Est.	SE	95% CI
Channel 1	-0.006	0.136	(-0.273, 0.261)	-0.018	0.157	(-0.327, 0.290)	-0.014	0.128	(-0.266, 0.237)
Channel 3	1.601	0.553	(0.518 , 2.684)	1.395	0.666	(0.090, 2.700)	1.389	0.500	(0.409, 2.369)
Channel 4	0.761	0.583	(-0.383 , 1.904)	0.817	0.671	(-0.498, 2.131)	0.821	0.556	(-0.269, 1.910)
Channel 5	0.163	0.560	(-0.934 , 1.260)	0.430	0.657	(-0.856, 1.717)	0.449	0.549	(-0.627, 1.525)
Time-point 1	0.004	0.002	(0.001 , 0.008)	0.005	0.002	(0.001, 0.009)	0.005	0.002	(0.001, 0.009)
Time-point 2	-0.013	0.003	(-0.018, -0.008)	-0.014	0.003	(-0.019, -0.008)	-0.014	0.003	(-0.019, -0.008)
Time-point 3	0.020	0.006	(0.009, 0.031)	0.023	0.007	(0.009, 0.036)	0.023	0.007	(0.010, 0.036)
Time-point 4	0.010	0.005	(0.001, 0.019)	0.011	0.005	(0.001, 0.020)	0.011	0.005	(0.001, 0.021)
Time-point 5	0.002	0.005	(-0.008, 0.012)	0.003	0.005	(-0.007, 0.013)	0.003	0.006	(-0.008, 0.014)

Chapter 3

Imputation and Likelihood Methods for Matrix-Variate Logistic Regression with Response Misclassification

In this chapter, we target on investigating how response misclassification in the matrix-variate logistic regression affects the parameter inference, and propose imputation and likelihood methods to reduce the response misclassification effects. The remainder is organized as follows. In Section 3.1, we present the response model and introduce the misclassification process for binary response. In Section 3.2, we propose an important method using an unbiased surrogate for the true response. In Section 3.3, we explore the likelihood method based on the observed data. In Section 3.4, we conduct simulation studies to assess the performance of the methods developed in Sections 3.3-3.4 as well as to demonstrate the misclassification effects on the naive analysis which ignores the response misclassification. The proposed methods are also applied to analyze a breast cancer data set.

3.1 Notation and Framework

For subject k with $k = 1, ..., n$, Y_k, x_k and z_k are defined in the same way as those in Section 2.1.1. Let $\mu_k = P(Y_k = 1|x_k, z_k)$. We consider the model

$$\text{logit } \mu_k = \gamma_0 + \alpha^{\intercal} x_k \beta + \gamma_1^{\intercal} z_k, \tag{3.1}$$

where γ_0 is the intercept term and γ_1 is a $p_z \times 1$ parameter vector. Write $\gamma = (\gamma_0, \gamma_1^\intercal)^\intercal$. Model (3.1) differs slightly from model (2.2) in that we explicitly spell out the intercept term here.

As discussed in Section 2.1.1, parameters α and β are not identifiable because they are pertinent to a rank-1 CP-decomposition which is not unique. To overcome nonidentifiability issues, we use the same way as in Chapter 2 to set the first element of α to be 1; but here we write $\alpha = (1, \alpha_2, ..., \alpha_{p+1})^\intercal = (1, \tilde{\alpha}^\intercal)^\intercal$. Let $\theta = (\tilde{\alpha}^\intercal, \beta^\intercal, \gamma^\intercal)^\intercal$, which is the vector of parameters of interest.

Estimation of θ can be carried out using the likelihood method. For $k = 1, ..., n$, the log-likelihood function contributed from subject k is

$$\ell(\theta; Y_k) = Y_k(\gamma_0 + \alpha^\intercal x_k \beta + \gamma_1^\intercal z_k) - \log\{1 + \exp(\gamma_0 + \alpha^\intercal x_k \beta + \gamma_1^\intercal z_k)\}, \tag{3.2}$$

where the dependence on z_k and x_k is suppressed in the notation $\ell(\theta; Y_k)$. Let $U(\theta; Y_k) = \partial\ell(\theta; Y_k)/\partial\theta$ be the score function, and we write $U(\theta; Y_k) = \{U_{1k}^\intercal(\theta), U_{2k}^\intercal(\theta), U_{3k}^\intercal(\theta)\}^\intercal$, where $U_{1k}(\theta) = \partial\ell(\theta; Y_k)/\partial\tilde{\alpha}$, $U_{2k}(\theta) = \partial\ell(\theta; Y_k)/\partial\beta$, and $U_{3k}(\theta) = \partial\ell(\theta; Y_k)/\partial\gamma$.

Under regularity conditions, a consistent estimator of θ can be obtained by solving

$$\sum_{k=1}^{n} U(\theta; Y_k) = 0 \tag{3.3}$$

for θ. Using the block relaxing algorithm in Table 2.1, we solve (3.3) iteratively for $\tilde{\alpha}$, β and γ while keeping other components fixed.

In applications, the response Y_k may be subject to misclassification, and a surrogate response, Y_k^*, is observed, where $k = 1, ..., n$. For $i, j = 0, 1$, let $\tau_{kij} = P(Y_k^* = j | Y_k = i, x_k, z_k)$ be the probability that the observed response is j when the true response is i, where the dependence on x_k and z_k is suppressed in the notation τ_{kij}.

To facilitate the dependence of the τ_{kij} on the covariates, we consider the logistic models

$$\text{logit } \tau_{k01} = L_k^\intercal \phi_0,$$

and

$$\text{logit } \tau_{k11} = L_k^\intercal \phi_1, \tag{3.4}$$

where ϕ_0 and ϕ_1 are the vectors of associated regression parameters, and L_k is a vector of covariates that reflects various misclassification mechanisms. Let $\phi = (\phi_0^\intercal, \phi_1^\intercal)^\intercal$. L_k may be specified as various forms to feature different misclassification processes. In some cases, L_k

is taken as the entire vector covariate z_k; in the extreme case, L_k is taken as the constant 1 to express that the misclassification is independent of the covariates: $\tau_{k01} = \text{expit}(\phi_0)$ and $\tau_{k11} = \text{expit}(\phi_1)$, where $\text{expit}(u) = \exp(u)/\{1 + \exp(u)\}$ and ϕ_0 and ϕ_1 are scalar.

3.2 Imputation Method

3.2.1 Estimating Equations with Known Misclassification Probabilities

Define

$$Y_k^c = \frac{Y_k^* - \tau_{k01}}{\tau_{k11} - \tau_{k01}}, \tag{3.5}$$

where $\tau_{k10} = 1 - \tau_{k11}$. It is easily seen that $E(Y_k^c | Y_k, x_k, z_k) = Y_k$, i.e., Y_k^c is an unbiased surrogate for Y_k, as called by Chen et al. (2014).

Let $U_{1k}^c(\theta) = \partial \ell(\theta; Y_k^c)/\partial \tilde{\alpha}$, $U_{2k}^c(\theta) = \partial \ell(\theta; Y_k^c)/\partial \beta$, and $U_{3k}^c(\theta) = \partial \ell(\theta; Y_k^c)/\partial \gamma$. Define $U^c(\theta; Y_k^c) = \{U_{1k}^{c\intercal}(\theta), U_{2k}^{c\intercal}(\theta), U_{3k}^{c\intercal}(\theta)\}^\intercal$. Then

$$E\{U^c(\theta; Y_k^c) | Y_k, x_k, z_k\} = U(\theta; Y_k),$$

suggesting that $U^c(\theta; Y_k^c)$ is an unbiased estimating function of θ. When ϕ is known, solving

$$\sum_{k=1}^{n} U^c(\theta; Y_k^c) = 0 \tag{3.6}$$

for θ gives a consistent estimator, say $\hat{\theta}_c$, for θ, provided regularity conditions.

Let

$$M_{1k}^c(\tilde{\alpha}|\beta, \gamma) = \frac{\partial U_{1k}^c(\tilde{\alpha}|\beta, \gamma)}{\partial \tilde{\alpha}^\intercal},$$
$$M_{2k}^c(\beta|\tilde{\alpha}, \gamma) = \frac{\partial U_{2k}^c(\beta|\tilde{\alpha}, \gamma)}{\partial \beta^\intercal},$$

and

$$M_{3k}^c(\gamma|\tilde{\alpha}, \beta) = \frac{\partial U_{3k}^c(\gamma|\tilde{\alpha}, \beta)}{\partial \gamma^\intercal}.$$

Using the block relaxation algorithm, we solve (3.6) via the Fisher Scoring algorithm. At iteration $(t+1)$, we iteratively update $\tilde{\alpha}$, β and γ in each block by

$$\tilde{\alpha}^{t+1,r+1} = \tilde{\alpha}^{t+1,r} - \Big\{\sum_{k=1}^{n} M_{1k}^{c}(\tilde{\alpha}^{t+1,r}|\beta^{t},\gamma^{t})\Big\}^{-1}\Big\{\sum_{k=1}^{n} U_{1k}^{c}(\tilde{\alpha}^{t+1,r}|\beta^{t},\gamma^{t})\Big\},$$
$$\beta^{t+1,r+1} = \beta^{t+1,r} - \Big\{\sum_{k=1}^{n} M_{2k}^{c}(\beta^{t+1,r}|\tilde{\alpha}^{t+1},\gamma^{t})\Big\}^{-1}\Big\{\sum_{k=1}^{n} U_{2k}^{c}(\beta^{t+1,r}|\tilde{\alpha}^{t+1},\gamma^{t})\Big\},$$
$$\gamma^{t+1,r+1} = \gamma^{t+1,r} - \Big\{\sum_{k=1}^{n} M_{3k}^{c}(\gamma^{t+1,r}|\tilde{\alpha}^{t+1},\beta^{t+1})\Big\}^{-1}\Big\{\sum_{k=1}^{n} U_{3k}^{c}(\gamma^{t+1,r}|\tilde{\alpha}^{t+1},\beta^{t+1})\Big\},$$

for $r = 0, 1, 2, ...,$ where $\tilde{\alpha}^{t}$, β^{t} and γ^{t} represent the estimates of $\tilde{\alpha}$, β and γ at iteration t, respectively. Let θ_0 be the true value of θ. In Appendix B.2, we show the following asymptotic result of $\hat{\theta}_c$.

Theorem 3.1 *Assume Conditions (C.1)-(C.2) in Appendix B.1. Then as $n \to \infty$,*

$$\sqrt{n}(\hat{\theta}_c - \theta_0) \xrightarrow{d} N(0, \Gamma_c^{-1}\Sigma_c[\Gamma_c^{-1}]^{\intercal}),$$

where

$$\Gamma_c = E\{\partial U^{c}(\theta_0; Y_k^{c})/\partial\theta^{\intercal}\} \text{ and } \Sigma_c = E\{U^{c}(\theta_0; Y_k^{c})U^{c\intercal}(\theta_0; Y_k^{c})\}.$$

To carry out inference such as constructing confidence intervals, we use the asymptotic distribution in Theorem 3.1 by replacing Γ_c and Σ_c with their consistent estimates

$$\hat{\Gamma}_c = \frac{1}{n}\sum_{k=1}^{n}\frac{\partial U^{c}(\theta; Y_k^{c})}{\partial\theta^{\intercal}}\Big|_{\theta=\hat{\theta}_c} \text{ and } \hat{\Sigma}_c = \frac{1}{n}\sum_{k=1}^{n} U^{c}(\hat{\theta}_c; Y_k^{c})U^{c\intercal}(\hat{\theta}_c; Y_k^{c})$$

respectively, thus, yielding a consistent estimator of the asymptotic covariance matrix of $\hat{\theta}_c$, given by $\hat{\Gamma}_c^{-1}\hat{\Sigma}_c[\hat{\Gamma}_c^{-1}]^{\intercal}$.

3.2.2 Estimating Equations with Unknown Misclassification Probabilities

In Section 3.2.1, we solve (3.6) by assuming that the misclassification parameter ϕ is known. However, the misclassification parameters are usually unknown in practice. In this case, a two-stage estimation procedure can be applied to estimate θ and ϕ, where an unbiased estimating function for ϕ is constructed in addition to (3.6). Often a validation subsample is needed for estimation of the misclassification parameters (Roy et al. 2005). Here, we describe an inferential procedure by incorporating estimation of the misclassification parameters when an internal validation subsample is available (Chen et al. 2011; Chen et al. 2014).

For $k = 1, ..., n$, let δ_k be the indicator variable of the kth subject such that when $\delta_k = 1$, the kth subject is included in validation subsample and $\delta_k = 0$ otherwise. Then $p_v = \sum_{k=1}^{n} \delta_k/n$ is the proportion of the subjects that are included in the validation subsample.

For $k = 1, ..., n$, let H_k be the indicator variable $I(Y_k^* \neq Y_k)$ for the kth subject, taking value 1 if $Y_k^* \neq Y_k$ and 0 otherwise. Thus, $H_k = 1$ is equivalent to either "$Y_k^* = 1$, $Y_k = 0$" or "$Y_k^* = 0$, $Y_k = 1$". For ease of notation, for $y_k^* = 0, 1$, we let

$$\ell_{k0}(y_k^*) = \log\{\tau_{k01}^{H_k} \times (1 - \tau_{k01})^{(1-H_k)}\}$$

denote the logarithm of the conditional probability $P(Y_k^* = y_k^* | Y_k = 0)$ and let

$$\ell_{k1}(y_k^*) = \log\{\tau_{k10}^{H_k} \times (1 - \tau_{k10})^{(1-H_k)}\}$$

denote the logarithm of the conditional probability $P(Y_k^* = y_k^* | Y_k = 1)$.

Define $S_k(\phi) = (\partial \ell_{k0}(y_k^*)/\partial\phi)^{1-y_k}(\partial \ell_{k1}(y_k^*)/\partial\phi)^{y_k}$, which can be used to estimate ϕ using the measurements in the validation subsample. Now we describe a two-stage estimation procedure for estimation of ϕ and θ.

Stage 1. Applying $S_k(\phi)$ to the validation subsample and solving

$$\sum_{k=1}^{n} \delta_k S_k(\phi) = 0 \tag{3.7}$$

for ϕ gives an estimate, say $\hat{\phi}_v$, of ϕ.

Stage 2. Replace ϕ with $\hat{\phi}_v$ in (3.6) and solve it for θ using the block relaxation algorithm.

This two-stage estimation procedure can be expressed as a single procedure for ease of establishing the asymptotic results of the resulting estimator. Let $\eta = (\phi^\intercal, \theta^\intercal)^\intercal$. Then solving

$$\sum_{k=1}^{n} \begin{pmatrix} U^c(\theta, \phi; Y_k^c) \\ \delta_k S_k(\phi) \end{pmatrix} = 0 \tag{3.8}$$

for η gives a consistent estimator, denoted $\hat{\eta}_v = (\hat{\phi}_v^\intercal, \hat{\theta}_v^\intercal)^\intercal$, for η, provided regularity conditions.

Applying the first-order Taylor series approximation to the estimating functions in (3.8) around η_0, the true value of η, we can establish Theorem 3.2 as follows. The details are included in Appendix B.3.

Theorem 3.2 *Assume that Conditions (C.1)-(C.4) in Appendix B.1 hold and that p_v approaches a positive constant as $n \to \infty$. Then as $n \to \infty$,*

$$\sqrt{n}(\hat{\theta}_v - \theta_0) \xrightarrow{d} N(0, \Gamma_c^{-1}\Sigma_\tau[\Gamma_c^{-1}]^\intercal),$$

where $\Sigma_\tau = E\{\Omega_k(\theta_0, \phi_0)\Omega_k^\intercal(\theta_0, \phi_0)\}$, and

$$\begin{aligned}\Omega_k(\theta_0, \phi_0) = {} & U^c(\theta_0, \phi_0; Y_k^c) - \mathrm{E}\Big\{\partial U^c(\theta_0, \phi_0; Y_k^c)/\partial\phi\Big\} \\ & \times \Big[\mathrm{E}\Big\{\delta_k \times \partial S_k(\phi_0)/\partial\phi\Big\}\Big]^{-1} \times \{\delta_k S_k(\phi_0)\}.\end{aligned}$$

As $n \to \infty$, the matrix Σ_τ can be consistently estimated by $\hat{\Sigma}_\tau = \frac{1}{n}\sum_{k=1}^{n} \hat{\Omega}_k(\hat{\theta}_v, \hat{\phi}_v)\hat{\Omega}_k^\intercal(\hat{\theta}_v, \hat{\phi}_v)\}$, where

$$\begin{aligned}\hat{\Omega}_k(\hat{\theta}_v, \hat{\phi}_v) = {} & U^c(\hat{\theta}_v, \hat{\phi}_v; Y_k^c) - \Big\{\frac{1}{n}\sum_{k=1}^{n} \frac{\partial U^c(\theta, \phi; Y_k^c)}{\partial\phi}\Big|_{\eta=\hat{\eta}_v}\Big\} \\ & \times \Big[\frac{1}{n}\Big\{\sum_{k=1}^{n} \frac{\partial \delta_k S_k(\hat{\phi}_v)}{\partial\phi}\Big|_{\phi=\hat{\phi}_v}\Big\}\Big]^{-1} \times \{\delta_k S_k(\hat{\phi}_v)\}.\end{aligned}$$

3.3 Likelihood Method

3.3.1 Inference Method with Known Misclassification Probabilities

The second method of estimation of the parameters is based on the observed likelihood function. Let $\mu_k^* = P(Y_k^* = 1|x_k, z_k)$ be the conditional mean for the surrogate response Y_k^*, given $\{x_k, z_k\}$. As discussed in Yi (2017, Chapter 8), the conditional probability μ_k^* of the observed measurement Y_k^*, given $\{x_k, z_k\}$, is linked with the conditional probability μ_k

of the true response Y_k, given $\{x_k, z_k\}$, through

$$\mu_k^* = \tau_{k01} + (1 - \tau_{k01} - \tau_{k10})\mu_k. \tag{3.9}$$

If the parameters for the misclassification probabilities are known, then the maximum likelihood estimator, say $\hat{\theta}$, of θ can be obtained by maximizing the log-likelihood for the observed data $\sum\limits_{k=1}^{n} \ell^o(\theta; Y_k^*)$ with respect to θ, where for $k = 1, ..., n$,

$$\ell^o(\theta; Y_k^*) = Y_k^* \log \mu_k^* + (1 - Y_k^*) \log(1 - \mu_k^*), \tag{3.10}$$

and μ_k^* is determined by (3.9) in combination with (3.1) and (3.4).

Under regularity conditions, $\hat{\theta}$ can be equivalently obtained by solving the

$$\sum_{k=1}^{n} U^o(\theta; Y_k^*) = 0, \tag{3.11}$$

where $U^o(\theta; Y_k^*) = \{U_{1k}^{o\mathsf{T}}(\theta), U_{2k}^{o\mathsf{T}}(\theta), U_{3k}^{o\mathsf{T}}(\theta)\}^{\mathsf{T}}$ with $U_{1k}^o(\theta) = \partial \ell^o(\theta; Y_k^*)/\partial \tilde{\alpha}$, $U_{2k}^o(\theta) = \partial \ell^o(\theta; Y_k^*)/\partial \beta$, and $U_{3k}^o(\theta) = \partial \ell^o(\theta; Y_k^*)/\partial \gamma$.

Likelihood theory shows that as $n \to \infty$,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma^{-1}),$$

provided regularity conditions, where $\Sigma = E\{-\partial U^o(\theta_0; Y_k^*)/\partial \theta^{\mathsf{T}}\}$, which can be consistently estimated by $n^{-1} \sum\limits_{k=1}^{n} U^o(\hat{\theta}; Y_k^*) U^{o\mathsf{T}}(\hat{\theta}; Y_k^*)$.

3.3.2 Inference Method with Unknown Misclassification Probabilities

In this subsection, we consider using an internal validation sample to estimate the parameter vector ϕ associated with the misclassification model (3.4). The inference about η, defined in Section 3.2.2, can be carried out based on the likelihood function for the observed data, given by

$$L_v^o(\eta) = \Big\{ \prod_{\delta_k = 1} f(y_k, y_k^* | x_k, z_k) \Big\} \Big\{ \prod_{\delta_k = 0} f(y_k^* | x_k, z_k) \Big\},$$

where the contributions of the subjects in the validation sample are reflected by the model $f(y_k, y_k^*|x_k, z_k)$ for the conditional distribution of $\{Y_k, Y_k^*\}$ given $\{x_k, z_k\}$, determined by (3.1), (3.4) and (3.9); and the subjects in the main study contribute via the model $f(y_k^*|x_k, z_k)$, determined by (3.9). More specifically, $L_v^o(\eta)$ is given by

$$\begin{aligned} L_v^o(\eta) = &\left\{ \prod_{\delta_k=1} \left[\mu_k^{y_k}(1-\mu_k)^{1-y_k} \{a_{k1}(y_k^*)\}^{y_k} \{a_{k0}(y_k^*)\}^{1-y_k} \right] \right\} \\ &\times \prod_{\delta_k=0} \{\mu_k^{*y_k^*}(1-\mu_k^*)^{1-y_k^*}\}, \end{aligned} \tag{3.12}$$

where for $l = 0$ and 1, $a_{kl}(y_k^*) = P(Y_k^* = y_k^*|Y_k = l, X_k, z_k)$, given by

$$a_{k0}(y_k^*) = \tau_{k01}^{y_k^*}(1-\tau_{k01})^{1-y_k^*} \text{ and } a_{k1}(y_k^*) = \tau_{k10}^{1-y_k^*}(1-\tau_{k10})^{y_k^*}.$$

Maximizing (3.12) with respect to η leads to the maximum likelihood estimator for η. Although directly maximizing (3.12) can provide a statistically efficient estimator for η, the procedure may be computationally difficult to implement. Alternatively, we describe a two-stage estimation procedure which is computationally easier to implement, especially under the matrix-variate setting.

The two-stage algorithm treats θ and ϕ different. At the first stage, we employ (3.7) to obtain the estimate of ϕ using a validation subsample. At the second stage, estimation of θ is carried out by solving $\sum_{k=1}^{n} U_v^o(\eta) = 0$, or equivalently,

$$\sum_{\delta_k=1} U_1^o(\eta; y_k) + \sum_{\delta_k=0} U_2^o(\eta; y_k^*) = 0 \tag{3.13}$$

for θ with ϕ replaced by the estimate obtained from the first stage, where $U_v^o(\eta) = \partial \log(L_v^o)/\partial\theta$, $U_1^o(\eta; Y_k) = \left\{\frac{Y_k - \mu_k}{\mu_i(1-\mu_i)}\right\}\left(\frac{\partial\mu_k}{\partial\theta^\intercal}\right)$, and $U_2^o(\eta; Y_k^*) = \left\{\frac{Y_k^* - \mu_k^*}{\mu_i^*(1-\mu_i^*)}\right\}\left(\frac{\partial\mu_k^*}{\partial\theta^\intercal}\right)$.

This two-stage estimation procedure can be expressed as a single procedure for ease of establishing the asymptotic results of the resulting estimator. Solving

$$\sum_{k=1}^{n} \begin{pmatrix} U_v^o(\eta; Y_k, Y_k^*) \\ \delta_k S_k(\phi) \end{pmatrix} = 0 \tag{3.14}$$

for η gives a consistent estimator, denoted $\hat{\eta}_{ov} = (\hat{\phi}_v^\intercal, \hat{\theta}_{ov}^\intercal)^\intercal$, for η, provided regularity conditions. The asymptotic property of $\hat{\theta}_{ov}$ can be established similarly to Theorem 3.2 and is presented in Theorem 3.3 as follows.

Theorem 3.3 *Assume that Conditions (C.1)-(C.4) in Appendix B.1 hold and that p_v approaches a positive constant as $n \to \infty$. Then*

$$\sqrt{n}(\hat{\theta}_{ov} - \theta_0) \xrightarrow{d} N(0, \Gamma_o^{-1}\Sigma_{\tau o}[\Gamma_o^{-1}]^{\intercal}) \text{ as } n \to \infty,$$

where $\Gamma_o = E\{\partial U_v^o(\eta; Y_k, Y_k^*)/\partial\theta^{\intercal}\}$, $\Sigma_{\tau o} = E\{\Omega_{ok}(\theta_0, \phi_0)\Omega_{ok}^{\intercal}(\theta_0, \phi_0)\}$, *and*

$$\begin{aligned}\Omega_{ok}(\theta_0, \phi_0) = U_v^o(\eta_0) - \mathrm{E}\Big\{\partial U_v^o(\eta_0)/\partial\phi\Big\} \\ \times \Big[\mathrm{E}\Big\{\delta_k \times \partial S_k(\phi_0)/\partial\phi\Big\}\Big]^{-1} \times \{\delta_k S_k(\phi_0)\}.\end{aligned}$$

As $n \to \infty$, $\Sigma_{\tau o}$ and Γ_o can be consistently estimated by

$$\hat{\Sigma}_{\tau o} = \frac{1}{n}\sum_{k=1}^{n}\hat{\Omega}_{ok}(\hat{\theta}_{ov}, \hat{\phi}_v)\hat{\Omega}_{ok}^{\intercal}(\hat{\theta}_{ov}, \hat{\phi}_v) \text{ and } \hat{\Gamma}_o = \frac{1}{n}\sum_{k=1}^{n}\frac{\partial U_v^o(\eta; Y_k, Y_k^*)}{\partial\theta^{\intercal}}\Big|_{\eta=\hat{\eta}_{ov}},$$

respectively, where

$$\begin{aligned}\hat{\Omega}_{ok}(\hat{\theta}_{ov}, \hat{\phi}_v) = U_v^o(\hat{\eta}_{ov}; Y_k, Y_k^*) - \Big\{\frac{1}{n}\sum_{k=1}^{n}\frac{\partial U_v^o(\eta; Y_k, Y_k^*)}{\partial\phi}\Big|_{\eta=\hat{\eta}_{ov}}\Big\} \\ \times \Big\{\frac{1}{n}\sum_{k=1}^{n}\frac{\partial\delta_k S_k(\phi)}{\partial\phi}\Big|_{\phi=\hat{\phi}_v}\Big\}^{-1} \times \{\delta_k S_k(\hat{\phi}_v)\}.\end{aligned}$$

3.4 Numerical Studies

3.4.1 Simulation Designs

In this subsection, different simulations are designed to evaluate the performance of the proposed methods as well as the impacts of small, moderate and large degrees of response misclassification on parameter estimation, where we consider settings with $p + 1 = q$, denoted p_x for ease of exposition, and the sample size $n = 1000$.

Specifically, $p_x \times p_x$ matrix-variate data, x_k, are simulated from the matrix-normal distribution $MN(0, I_{p_x}, I_{p_x})$ for $k = 1, ..., n$, where $p_x = 5$. For the vector-covariate z_k, we consider two cases: (1) the z_k are continuous and independently generated from the standard normal distribution; (2) the z_k are binary and independently generated from the

Bernoulli distribution with $P(z_k = 1) = 0.5$. To easily differentiate these two types of z_k, we use z_{1k} and z_{2k} to express the covariate in these two cases, repectively.

For $k = 1, ..., n$, the binary response Y_k is randomly generated from the Bernoulli distribution with the probability

$$P(Y_k = 1|x_k, z_k) = \frac{\exp(\gamma_0 + \alpha^\intercal x_k \beta + \gamma_1 z_k)}{1 + \exp(\gamma_0 + \alpha^\intercal x_k \beta + \gamma_1 z_k)}, \tag{3.15}$$

where $\gamma_0 = \log 2$, $\gamma_1 = 0.5$, $\alpha = (0, 1, 0, 0.5, 0.5)^\intercal$, and $\beta = (0.5, -0.5, 0, 0.5, 0)^\intercal$.

The misclassification rates are determined by (3.4) and we consider five settings. In the first four settings, we let $L_k = 1$ and $\tau_{k01} = \tau_{k10}$ for simplicity where τ_{k01} and τ_{k10} are set as 2.5%, 5%, 10% and 20%, respectively, to reflect increasing degrees of misclassification. For these settings, the response Y_k are generated from (3.15) with z_k set as z_{1k}. In the fifth setting, we take $L_k = (1, z_{2k})^\intercal$ together with $\phi_0 = (-3, 0.5)^\intercal$ and $\phi_1 = (3, 0.5)^\intercal$ in (3.4) to generate τ_{k01} and τ_{k10}. When $z_{2k} = 0$, τ_{k01} and τ_{k10} are roughly 5%; when $z_{2k} = 1$, τ_{k01} and τ_{k10} are roughly 7.5%.

For $k = 1, ..., n$, the observed response, Y_k^*, is independently obtained using (3.4) as specified as one of the five settings with the designed misclassification rates. To apply the proposed methods to fit the data, we consider two scenarios. In Scenario 1, we take the misclassification rates as known and fit the data using the methods described in Sections 3.2.1 and 3.3.1. In Scenario 2, we apply the methods in Sections 3.2.2 and 3.3.2 by taking the misclassification rates as unknown and estimated from an internal validation sample. To investigate the effect of different sizes of the internal validation data, we randomly take 30% or 60% of the data as an internal validation sample.

3.4.2 Simulation Results

Tables 3.1-3.5 present the results for the estimators of α, β and γ where finite sample biases in percent (bias%), empirical standard errors (ESE), model-based asymptotic standard errors (ASE), and coverage rates (CR) for 95% confidence intervals are reported.

For the row effects α, the imputation methods and the likelihood methods give similar estimate results to those obtained from the naive method, regardless of the degrees of misclassification or the size of internal validation data. However, for the column effects β and the vector-covariate effects γ, we observe that biases resulted from the naive method are much larger than those obtained from the proposed methods even when the misclassification degree is small. The performance of estimators from the naive method becomes worse as the degree of misclassification increases. On the other hand, the imputation methods and the likelihood methods significantly improve the performance of the naive method. Furthermore, the likelihood methods outperform the imputation methods, although the performance of the imputation methods is fairly satisfactory under various settings. The likelihood methods are more efficient than the imputation methods and tend to be less affected by the size of the validation sample or the degree of misclassification than the imputation methods.

In summary, the naive method produces considerably biased results when misclassification exists in the response variables, suggesting that it is imperative to account for the misclassification effects in statistical inference when facing misclassification problems. The simulation studies confirm that the proposed methods significantly improve the performance of the naive method and satisfactorily accommodate the effects induced from the response misclassification. The likelihood methods have better performance than the imputation methods. It also confirms that when misclassification rates are unknown, the more the internal validation data, the better the results.

Finally, we comment that to improve the accuracy of estimation results for a given dimension of x_k and z_k, increasing the sample size is typically helpful, as noticed by a referee. In our numerical explorations, we found that for the settings considered in this section, reducing the sample size to a small value (such as 200) can generate unstable results with more nonconverging estimates.

3.4.3 Sensitivity Study

To investigate the robustness of the proposed methods in Sections 3.2.1 and 3.3.1, we conduct the following two simulation studies. In Simulation 1, we generate the Y_k using (3.15) with z_k set as z_{1k} and the Y_k^* using (3.4) with $L_k = 1$, yielding that τ_{k01} and τ_{k10} are common for $k = 1, ..., n$; let τ_{01} and τ_{10} denote them, respectively. We consider one of the two settings to generate the surrogate responses: (1) $\tau_{01} = 5\%$ and $\tau_{10} = 10\%$; (2) $\tau_{01} = 10\%$ and $\tau_{10} = 5\%$. We apply the methods in Sections 3.2.1 and 3.3.1 by mis-taking τ_{01} and τ_{10} as $\tau_{01} = \tau_{10} = 2.5\%$, 7.5% or 10% to fit the data.

In Simulation 2, we generate the Y_k from (3.15) with z_k set as z_{2k} and the Y_k^* from (3.4) with $L_k = (1, z_{2k})^\intercal$, $\phi_0 = (-3, 0.5)^\intercal$, and $\phi_1 = (1.5, 0.5)^\intercal$. However, we fit the data using the methods in Sections 3.2.1 and 3.3.1 with τ_{k01} and τ_{k10} misspecified as one of the settings: (1) $\tau_{k01} = 10\%$ and $\tau_{k10} = 15\%$; (2) $\tau_{k01} = 5\%$ and $\tau_{k10} = 10\%$, for all $k = 1, ..., n$.

The results are reported in Tables 3.6-3.8. For estimation of the row and column effects, our proposed methods still perform better than the naive method under the misspecification of the misclassification rates we consider here. However, for estimation of the vector effect, γ, our methods may perform better than the naive method only when the misspecified misclassification rates are not severe.

3.4.4 Analysis of the Breast Cancer Wisconsin Prognostic Data

We apply the proposed methods, in contrast to the naive approach, to analyze the breast cancer Wisconsin prognostic imaging data which are available at the UCI Machine Learning Repository website (`https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic)`). The data set contains 198 breast cancer patients whose cases exhibit invasive breast cancer but no evidence of distant metastases at the time of diagnosis. Ten real features, *Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave Points, Symmetry* and *Fractal Dimension*, of the cell nucleus in the digitized image of a fine needle aspirate (FNA) of breast mass of each subject were taken as the row effects. For each feature, the *Mean, Standard Error,* and *Worst (mean of the three smallest values)* were computed for the cell nucleus in each breast mass image and treated as the column effects. Besides those measurements, the tumor size for each subject is available.

Those subjects are divided into the recurrent group of 47 patients and the nonrecurrent group of 151 patients. A patient is classified to be in the recurrent group if the disease is observed at some subsequent time to the tumor excision; and the nonrecurrent group includes patients whose cancer has not observed to recur, or may never recur. There is a possibility that the patients may be misclassified due to incorrect diagnosis for the recurrent group or the unknown recurring time for the nonrecurrent group patients (Mangasarian et al. 1995). Here we are interested in using the observed but error-prone data to study how risk factors may be associated with the true status of being in the recurrent or nonrecurrent group, which is postulated by model (3.1).

For $k = 1, .., 198$, let Y_k^* be the observed binary response variable for subject k, with value 1 for being in the recurrent group and 0 for being in the nonrecurrent group. The matrix-variate of the subject k, denoted as x_k, is a 10×3 matrix with entry (i, j) representing the value of the jth characteristic of the ith feature, where $i = 1, ..., 10$ and $j = 1, 2, 3$. The breast tumor size of the subject k is denoted as z_k. Consistent with the notation in Section 3.1, we let τ_{10} denote the rate of misclassifying a subject who actually is in the recurrent group into the observed nonrecurrent group as τ_{10}, and let τ_{01} denote the rate of misclassifying a subject who actually is in the nonrecurrent group into the observed recurrent group.

Since our proposed methods require the knowledge of the misclassification mechanism but there are no validation data available, we conduct sensitivity analysis by examining the impacts of different misclassification probabilities on the estimation of the model parameters. In particular, we consider two possible scenarios. In the first scenario, we take $\tau_{01} = 0$, reflecting no misclassification in the recurrent group, and set $\tau_{10} = 1\%$, 3%, or 5% to feature increasing misclassification cases. In the second scenario, we set $\tau_{01} = 1\%$ and let $\tau_{10} = 1\%$, 3%, or 5%.

Tables 3.9-3.11 report the estimation results for the breast cancer Wisconsin data obtained from the naive analysis by using (3.3) with Y_k replaced by Y_k^*, the imputation method (3.6), and the likelihood method (3.11). For the row effects $\tilde{\alpha}$, all analyses show that *Radius* has the highest negative effect and *Perimeter* has the highest positive effect. For the column effects, all the methods show that *Mean* has the highest positive effect and *Worst* has the highest negative effect. However, only the intercept term is statistically significant under 5% significant level. As the misclassification rate increases, the size of the effect as well as the standard errors from the proposed methods increases.

To conclude, we point out that caution should be taken when interpreting the results here. As noted in Section 3.4.2, a small sample size does not ensure reliable estimation results as the asymptotic results do not come into the play. The analysis here can be more regarded as an illustration of the utility of the proposed methods than taken as a sound revealing of new scientific findings for such a study.

Table 3.1: Simulation study for the three methods: $\tau_{01} = \tau_{10} = 2.5\%$ with z_{1k}

Parameters	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%
					Imputation Methods											
									Internal Validation							
	Naïve Method				Known misclassification rates				60% internal validation				30% internal validation			
α_4	-0.912	0.105	0.102	94.6	-0.926	0.105	0.102	94.8	-0.920	0.105	0.102	94.6	-0.916	0.105	0.102	94.8
α_5	-0.284	0.098	0.102	96.7	-0.282	0.098	0.102	96.7	-0.283	0.098	0.102	96.7	-0.284	0.098	0.102	96.7
β_1	-5.800	0.068	0.068	93.0	3.151	0.076	0.076	95.3	3.110	0.076	0.076	94.8	3.366	0.078	0.082	94.6
β_2	-6.564	0.071	0.068	89.7	2.324	0.080	0.076	93.8	2.286	0.079	0.076	94.0	2.495	0.081	0.081	93.2
β_4	-6.395	0.072	0.068	88.2	2.511	0.080	0.076	94.2	2.470	0.080	0.076	94.2	2.709	0.082	0.081	93.6
γ_0	-6.435	0.076	0.076	91.3	2.031	0.085	0.084	94.8	2.057	0.084	0.084	95.7	2.404	0.095	0.123	92.2
γ_1	-6.627	0.079	0.077	90.7	2.252	0.088	0.085	94.8	2.212	0.088	0.086	95.3	2.437	0.089	0.092	95.3
					Likelihood Methods											
									Internal Validation							
					Known misclassification rates				60% internal validation				30% internal validation			
α_4					-0.942	0.106	0.116	96.1	-0.738	0.103	0.102	94.7	-0.813	0.103	0.108	95.3
α_5					-0.023	0.098	0.115	98.0	-0.030	0.094	0.102	96.3	-0.194	0.096	0.107	96.5
β_1					2.979	0.076	0.086	96.7	2.682	0.071	0.075	95.3	3.029	0.075	0.081	96.1
β_2					2.450	0.079	0.085	96.3	2.513	0.075	0.075	95.1	2.715	0.079	0.080	95.5
β_4					2.074	0.079	0.085	95.7	1.859	0.073	0.075	95.3	2.059	0.077	0.080	96.3
γ_0					1.763	0.084	0.084	94.7	1.730	0.084	0.082	94.5	1.929	0.092	0.086	94.1
γ_1					1.944	0.088	0.096	95.3	1.387	0.084	0.084	93.9	1.769	0.087	0.090	94.5

Table 3.2: Simulation study for the three methods: $\tau_{01} = \tau_{10} = 5\%$ with z_{1k}

Parameters	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%
					Imputation Methods											
									Internal Validation							
	Naïve Method				Known misclassification rates				60% internal validation				30% internal validation			
α_4	-0.173	0.115	0.110	94.2	-0.192	0.115	0.110	94.0	-0.195	0.115	0.110	94.0	-0.190	0.115	0.110	94.2
α_5	-0.332	0.107	0.110	95.8	-0.332	0.107	0.110	95.6	-0.336	0.107	0.110	95.6	-0.329	0.107	0.110	95.6
β_1	-13.922	0.068	0.067	79.0	3.131	0.084	0.083	95.4	3.205	0.084	0.083	95.6	3.454	0.088	0.084	94.4
β_2	-14.806	0.069	0.067	76.8	2.081	0.086	0.083	95.4	2.157	0.087	0.083	95.0	2.436	0.092	0.084	93.4
β_4	-14.547	0.068	0.067	77.0	2.396	0.085	0.083	94.0	2.471	0.086	0.083	94.0	2.756	0.090	0.084	93.0
γ_0	-13.991	0.072	0.074	73.3	2.208	0.089	0.092	95.4	2.187	0.088	0.092	96.2	2.378	0.107	0.093	90.8
γ_1	-14.858	0.078	0.075	82.0	2.022	0.096	0.093	94.2	2.049	0.095	0.093	94.0	2.292	0.098	0.093	94.4
					Likelihood Methods											
									Internal Validation							
					Known misclassification rates				60% internal validation				30% internal validation			
α_4					-0.950	0.107	0.118	96.1	-0.882	0.103	0.103	94.3	-0.800	0.104	0.109	95.7
α_5					-0.082	0.099	0.117	97.8	-0.007	0.094	0.103	96.6	-0.285	0.096	0.108	97.0
β_1					3.145	0.076	0.088	97.0	2.735	0.072	0.076	95.5	3.113	0.075	0.082	96.1
β_2					2.336	0.080	0.086	96.6	2.475	0.075	0.075	95.5	2.624	0.080	0.080	94.7
β_4					2.074	0.080	0.087	95.7	1.883	0.073	0.075	95.5	2.062	0.077	0.081	96.6
γ_0					1.869	0.085	0.086	94.7	1.752	0.085	0.082	94.5	1.819	0.096	0.087	93.5
γ_1					1.950	0.090	0.098	95.7	1.310	0.085	0.084	93.7	1.696	0.087	0.091	94.3

Table 3.3: Simulation study for the three methods: $\tau_{01} = \tau_{10} = 10\%$ with z_{1k}

Parameters	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%
					Imputation Methods											
									Internal Validation							
	Naïve Method				Known misclassification rates				60% internal validation				30% internal validation			
α_4	0.147	0.136	0.128	93.4	0.110	0.136	0.128	93.6	0.105	0.136	0.128	93.6	0.124	0.136	0.128	93.6
α_5	-0.263	0.125	0.128	96.0	-0.251	0.125	0.128	95.6	-0.249	0.125	0.128	95.4	-0.230	0.125	0.128	95.6
β_1	-27.302	0.065	0.064	42.0	4.645	0.101	0.099	94.6	4.645	0.100	0.100	94.6	5.319	0.108	0.101	94.8
β_2	-28.582	0.068	0.064	40.2	2.828	0.106	0.099	92.6	2.771	0.103	0.099	93.8	3.516	0.112	0.101	93.2
β_4	-27.880	0.066	0.064	41.6	3.818	0.102	0.099	94.2	3.824	0.101	0.099	95.4	4.533	0.109	0.101	95.4
γ_0	-27.171	0.070	0.071	24.8	3.172	0.108	0.109	95.6	2.952	0.098	0.110	98.0	3.663	0.136	0.112	91.4
γ_1	-28.288	0.073	0.072	49.2	3.246	0.112	0.110	94.2	3.171	0.109	0.111	95.0	3.741	0.114	0.112	93.6
					Likelihood Methods											
									Internal Validation							
					Known misclassification rates				60% internal validation				30% internal validation			
α_4					-0.783	0.115	0.124	96.1	-0.761	0.105	0.104	93.7	-0.571	0.109	0.112	95.1
α_5					-0.580	0.101	0.123	97.4	-0.255	0.096	0.104	96.3	-0.678	0.099	0.111	96.3
β_1					3.476	0.080	0.094	97.0	2.596	0.073	0.076	95.1	3.000	0.077	0.083	95.7
β_2					2.354	0.085	0.092	96.1	2.413	0.075	0.076	95.5	2.411	0.081	0.082	95.1
β_4					2.741	0.083	0.093	95.7	2.157	0.074	0.076	95.3	2.324	0.078	0.082	96.6
γ_0					2.189	0.089	0.090	94.9	1.737	0.086	0.083	93.9	1.610	0.099	0.089	92.7
γ_1					2.384	0.092	0.104	96.1	1.371	0.085	0.085	93.9	1.673	0.088	0.093	95.1

Table 3.4: Simulation study for the three methods: $\tau_{01} = \tau_{10} = 20\%$ with z_{1k}

Parameters	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%
					Imputation Methods											
									Internal Validation							
	Naïve Method				Known misclassification rates				60% internal validation				30% internal validation			
α_4	-0.790	0.189	0.181	94.2	-0.861	0.189	0.181	94.0	-0.873	0.189	0.181	94.2	-0.841	0.189	0.181	94.4
α_5	-0.201	0.194	0.182	95.0	-0.194	0.193	0.182	94.8	-0.164	0.193	0.182	95.0	-0.185	0.193	0.182	94.8
β_1	-50.040	0.061	0.060	3.8	7.392	0.151	0.149	95.8	7.731	0.147	0.151	96.2	11.902	0.177	0.167	96.0
β_2	-50.803	0.063	0.060	2.6	5.751	0.153	0.148	94.6	6.294	0.152	0.150	95.2	10.629	0.191	0.168	94.2
β_4	-50.417	0.060	0.060	2.4	6.667	0.148	0.148	96.2	7.197	0.148	0.150	95.6	11.398	0.181	0.167	95.2
γ_0	-49.381	0.065	0.067	0.2	5.574	0.159	0.164	96.2	6.081	0.145	0.167	98.8	10.788	0.256	0.190	87.4
γ_1	-50.987	0.068	0.068	3.2	5.367	0.160	0.164	97.0	5.872	0.161	0.167	96.4	9.821	0.188	0.182	96.4
					Likelihood Methods											
									Internal Validation							
					Known misclassification rates				60% internal validation				30% internal validation			
α_4					-1.221	0.187	0.544	97.6	-1.410	0.118	0.115	95.0	-1.211	0.147	0.145	94.6
α_5					-0.132	0.187	0.669	97.6	0.609	0.108	0.115	96.6	0.001	0.137	0.147	97.4
β_1					6.807	0.144	0.947	98.6	3.177	0.084	0.085	94.8	3.977	0.104	0.110	95.2
β_2					6.138	0.148	0.363	98.6	2.627	0.085	0.085	95.4	3.342	0.110	0.110	93.6
β_4					6.339	0.145	0.645	97.4	3.242	0.085	0.085	95.4	3.984	0.108	0.111	95.8
γ_0					5.482	0.155	0.386	96.6	3.269	0.094	0.094	94.2	4.300	0.142	0.118	89.0
γ_1					5.159	0.160	2.105	97.4	1.604	0.095	0.096	95.2	1.764	0.115	0.124	95.0

Table 3.5: Simulation study for the three methods: $\phi_0 = (3, 0.5)^\intercal$ and $\phi_1 = (-3, 0.5)^\intercal$ with z_{2k}

Parameters	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%
					Imputation Methods											
									Internal Validation							
	Naïve Method				Known misclassification rates				60% internal validation				30% internal validation			
α_4	1.313	0.118	0.112	92.6	1.301	0.118	0.112	92.6	1.326	0.118	0.112	92.6	1.326	0.118	0.112	92.6
α_5	-0.327	0.115	0.111	93.8	-0.321	0.115	0.111	94.0	-0.335	0.115	0.111	94.2	-0.325	0.115	0.111	94.0
β_1	-12.614	0.072	0.068	81.4	2.728	0.088	0.083	94.8	2.763	0.088	0.083	94.2	3.082	0.092	0.084	92.8
β_2	-13.788	0.076	0.068	78.2	1.348	0.092	0.083	93.2	1.383	0.092	0.083	92.6	1.596	0.092	0.084	92.0
β_4	-12.984	0.068	0.068	81.2	2.298	0.083	0.083	95.0	2.344	0.084	0.083	95.6	2.619	0.086	0.084	94.4
γ_0	-12.794	0.101	0.102	84.4	1.803	0.122	0.122	95.8	1.851	0.122	0.123	95.8	2.444	0.142	0.124	93.0
γ_1	9.044	0.143	0.149	95.2	1.139	0.168	0.175	95.8	2.249	0.172	0.176	94.2	1.358	0.203	0.177	91.8
					Likelihood Methods											
									Internal Validation							
					Known misclassification rates				60% internal validation				30% internal validation			
α_4					1.245	0.118	0.125	93.8	1.121	0.110	0.108	93.4	1.323	0.113	0.116	93.8
α_5					-0.199	0.115	0.125	96.0	-0.250	0.105	0.107	94.4	-0.296	0.109	0.115	95.4
β_1					2.527	0.088	0.094	96.4	1.590	0.080	0.083	94.8	1.932	0.083	0.102	97.0
β_2					1.254	0.092	0.093	94.4	0.666	0.082	0.084	94.4	0.887	0.087	0.104	95.0
β_4					2.058	0.083	0.094	96.8	1.325	0.077	0.084	96.0	1.702	0.080	0.104	97.0
γ_0					1.646	0.121	0.128	97.2	0.796	0.117	0.127	95.8	1.232	0.122	0.161	98.6
γ_1					1.072	0.167	0.193	97.4	1.411	0.161	0.170	95.8	1.469	0.165	0.189	96.6

Table 3.6: Simulation study for the robustness of the three methods: Simulation 1 with Setting 1 ($\tau_{01} = 5\%$ and $\tau_{10} = 10\%$) by mis-taking τ_{01} and τ_{10} as 2.5%, 7.5% or 10% when fitting the data

Paramaters	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%
									Imputation Methods							
	Naïve Method				Mis-taking $\tau_{01} = \tau_{10}$ as 2.5%				Mis-taking $\tau_{01} = \tau_{10}$ as 7.5%				Mis-taking $\tau_{01} = \tau_{10}$ as 10%			
α_4	-0.040	0.128	0.120	94.6	-0.048	0.128	0.120	94.6	-0.067	0.128	0.120	94.6	-0.077	0.128	0.120	94.4
α_5	-0.280	0.118	0.120	95.8	-0.279	0.118	0.120	95.6	-0.276	0.118	0.120	95.8	-0.274	0.118	0.120	95.8
β_1	-22.753	0.066	0.065	55.4	-16.621	0.072	0.070	75.0	-0.168	0.090	0.088	94.8	11.387	0.103	0.101	93.8
β_2	-24.123	0.068	0.064	52.6	-18.100	0.074	0.070	70.4	-1.935	0.092	0.087	94.2	9.420	0.106	0.101	92.6
β_4	-23.481	0.065	0.064	53.4	-17.408	0.072	0.070	73.8	-1.112	0.089	0.087	93.8	10.334	0.103	0.101	93.0
γ_0	-13.038	0.106	0.071	64.4	-34.320	0.076	0.077	13.8	-21.787	0.093	0.095	63.4	-13.038	0.106	0.108	85.0
γ_1	10.099	0.115	0.073	75.4	-17.598	0.081	0.079	78.8	-1.330	0.101	0.098	93.4	10.099	0.115	0.112	93.2
									Likelihood Methods							
					Mis-taking $\tau_{01} = \tau_{10}$ as 2.5%				Mis-taking $\tau_{01} = \tau_{10}$ as 7.5%				Mis-taking $\tau_{01} = \tau_{10}$ as 10%			
α_4					-0.089	0.127	0.133	95.6	-0.223	0.127	0.139	96.4	-0.320	0.127	0.145	96.6
α_5					-0.281	0.117	0.132	96.4	-0.293	0.115	0.138	96.6	-0.298	0.115	0.144	96.8
β_1					-16.100	0.072	0.077	78.6	0.383	0.089	0.100	96.8	10.833	0.099	0.118	96.0
β_2					-17.508	0.074	0.075	74.2	-1.025	0.091	0.097	95.8	9.490	0.102	0.115	95.2
β_4					-16.855	0.072	0.075	78.2	-0.415	0.088	0.098	95.2	10.016	0.099	0.116	95.2
γ_0					-33.852	0.076	0.077	15.0	-20.843	0.092	0.094	65.2	-12.490	0.104	0.106	86.2
γ_1					-17.081	0.082	0.086	82.2	-0.723	0.101	0.111	94.8	9.680	0.114	0.131	94.8

Table 3.7: Simulation study for the robustness of the three methods: Simulation 1 with Setting 2 ($\tau_{01} = 10\%$ and $\tau_{10} = 5\%$) by mis-taking τ_{01} and τ_{10} as 2.5%, 7.5% or 10% when fitting the data

Paramaters	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%
					Imputation Methods											
	Naïve Method				Mis-taking $\tau_{01} = \tau_{10}$ as 2.5%				Mis-taking $\tau_{01} = \tau_{10}$ as 7.5%				Mis-taking $\tau_{01} = \tau_{10}$ as 10%			
α_4	0.038	0.122	0.117	94.2	0.029	0.122	0.117	94.0	0.009	0.123	0.117	94.0	-0.001	0.123	0.117	93.8
α_5	-0.292	0.113	0.117	95.8	-0.291	0.113	0.117	95.8	-0.286	0.113	0.117	95.8	-0.282	0.113	0.117	95.8
β_1	-18.346	0.067	0.067	69.6	-11.071	0.074	0.074	87.0	9.421	0.095	0.095	93.6	24.836	0.114	0.114	86.4
β_2	-19.150	0.069	0.067	67.8	-11.943	0.076	0.074	83.8	8.362	0.098	0.095	93.8	23.645	0.117	0.114	84.8
β_4	-18.831	0.069	0.066	68.0	-11.593	0.076	0.073	83.4	8.802	0.099	0.095	94.2	24.154	0.118	0.114	84.6
γ_0	46.250	0.129	0.074	6.4	6.433	0.080	0.082	93.2	29.270	0.106	0.108	55.2	46.250	0.129	0.132	26.8
γ_1	23.355	0.127	0.075	61.8	-12.162	0.084	0.083	86.8	8.100	0.108	0.106	93.2	23.355	0.127	0.126	87.6
					Likelihood Methods											
					Mis-taking $\tau_{01} = \tau_{10}$ as 2.5%				Mis-taking $\tau_{01} = \tau_{10}$ as 7.5%				Mis-taking $\tau_{01} = \tau_{10}$ as 10%			
α_4					-0.026	0.122	0.130	95.4	-0.182	0.122	0.137	96.0	-0.264	0.123	0.146	97.2
α_5					-0.263	0.112	0.129	97.4	-0.230	0.111	0.137	97.6	-0.212	0.112	0.146	97.6
β_1					-10.847	0.074	0.082	90.4	7.404	0.092	0.113	97.0	18.830	0.104	0.140	95.8
β_2					-11.644	0.075	0.081	85.8	6.758	0.093	0.111	97.2	18.334	0.104	0.140	96.0
β_4					-11.286	0.077	0.082	86.8	7.126	0.096	0.113	97.4	18.681	0.109	0.141	94.8
γ_0					6.528	0.080	0.083	93.6	27.142	0.100	0.106	58.8	40.205	0.114	0.127	37.4
γ_1					-11.921	0.084	0.092	89.8	6.313	0.105	0.125	95.6	17.795	0.119	0.153	95.4

Table 3.8: Simulation study for the robustness of the three methods: Simulation 2 $\{\phi_0 = (-3, 0.5)^\intercal$ and $\phi_1 = (1.5, 0.5)^\intercal\}$ by mis-taking $\tau_{01} = 5\%$ and $\tau_{10} = 10\%$, or $\tau_{01} = 10\%$ and $\tau_{10} = 5\%$ when fitting the data

Paramaters	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%
					Imputation Methods							
	Naïve Method				Mis-taking $\tau_{01} = 5\% and \tau_{10} = 10\%$				Mis-taking $\tau_{01} = 10\% and \tau_{10} = 5\%$			
α_4	1.798	0.164	0.157	93.4	1.750	0.164	0.157	93.4	1.773	0.164	0.157	93.4
α_5	0.472	0.175	0.156	92.6	0.428	0.175	0.156	92.6	0.449	0.175	0.156	92.6
β_1	-40.211	0.065	0.063	11.2	9.409	0.134	0.129	95.0	3.032	0.124	0.120	93.8
β_2	-41.085	0.067	0.062	11.8	7.788	0.135	0.129	95.2	1.478	0.125	0.119	94.8
β_4	-41.185	0.064	0.062	11.2	7.688	0.131	0.129	94.8	1.402	0.122	0.119	93.8
γ_0	-58.648	0.094	0.095	1.4	-30.072	0.165	0.168	73.4	-10.903	0.156	0.159	91.0
γ_1	17.681	0.134	0.139	91.8	113.292	0.263	0.274	44.0	101.626	0.247	0.257	48.0
					Likelihood Methods							
					Mis-taking $\tau_{01} = 5\% and \tau_{10} = 10\%$				Mis-taking $\tau_{01} = 10\% and \tau_{10} = 5\%$			
α_4					1.552	0.142	0.168	95.8	1.613	0.143	0.156	94.8
α_5					0.287	0.148	0.169	96.2	0.105	0.148	0.157	95.6
β_1					0.588	0.104	0.119	96.0	-17.415	0.082	0.087	77.8
β_2					-0.628	0.111	0.115	93.2	-18.565	0.088	0.085	74.2
β_4					-1.173	0.104	0.115	95.4	-18.907	0.082	0.085	77.2
γ_0					-84.104	0.138	0.156	2.4	-87.906	0.114	0.125	0.0
γ_1					66.741	0.201	0.250	78.8	38.743	0.166	0.191	85.8

Table 3.9: Analysis of the breast cancer Wisconsin prognostic data without accounting for misclassification effects

Parameter	Est.	SE	95% CI
Radius	-4.790	3.778	(-12.195, 2.614)
Texture	-0.122	0.115	(-0.348, 0.104)
Perimeter	3.909	3.951	(-3.835, 11.654)
Smoothness	0.242	0.233	(-0.214, 0.698)
Compactness	0.073	0.273	(-0.463, 0.609)
Concavity	-0.372	0.293	(-0.945, 0.202)
Concave Points	0.117	0.251	(-0.376, 0.610)
Symmetry	-0.041	0.119	(-0.276, 0.193)
Fractal Dimension	-0.411	0.266	(-0.933, 0.111)
Mean	3.089	1.791	(-0.422, 6.599)
SE	0.276	0.388	(-0.484, 1.037)
Worst	-0.773	0.660	(-2.066, 0.520)
Tumor Size	0.267	0.167	(-0.060, 0.594)
Intercept	-1.371	0.205	(-1.772, -0.969)

Table 3.10: Sensitivity analyses of the first scenario ($\tau_{01} = 0$) of the breast cancer Wisconsin prognostic data with different degrees of τ_{10} accommodated

Parameter	Imputation Method			Likelihood Method		
	Est.	SE	95% CI	Est.	SE	95% CI
			Scenario (i): assuming $\tau_{10} = 1\%$			
Radius	-4.740	3.728	(-12.048, 2.567)	-5.128	4.273	(-13.504, 3.247)
Texture	-0.121	0.114	(-0.344, 0.102)	-0.137	0.133	(-0.397, 0.124)
Perimeter	3.855	3.898	(-3.785, 11.496)	4.273	4.454	(-4.456, 13.003)
Smoothness	0.237	0.229	(-0.212, 0.687)	0.268	0.271	(-0.263, 0.798)
Compactness	0.074	0.227	(-0.457, 0.603)	0.064	0.286	(-0.496, 0.624)
Concavity	-0.368	0.289	(-0.934, 0.199)	-0.395	0.342	(-1.066, 0.276)
Concave Points	0.119	0.249	(-0.370, 0.608)	0.111	0.290	(-0.457, 0.678)
Symmetry	-0.041	0.118	(-0.273, 0.190)	-0.041	0.122	(-0.280, 0.197)
Fractal Dimension	-0.407	0.263	(-0.923, 0.109)	-0.432	0.333	(-1.086, 0.220)
Mean	3.139	1.817	(-0.422, 6.701)	3.025	2.153	(-1.196, 7.245)
SE	0.280	0.392	(-0.490, 1.049)	0.286	0.428	(-0.553, 1.125)
Worst	-0.788	0.671	(-2.103, 0.528)	-0.748	0.742	(-2.203, 0.707)
Tumor Size	0.268	0.168	(-0.061, 0.597)	0.277	0.171	(-0.062, 0.609)
Intercept	-1.357	0.206	(-1.760, -0.954)	-1.442	0.215	(-1.863, -1.021)
			Scenario (ii): assuming $\tau_{10} = 3\%$			
Radius	-4.636	3.627	(-11.745, 2.472)	-5.993	5.394	(-16.565, 4.579)
Texture	-0.118	0.111	(-0.335, 0.099)	-0.179	0.180	(-0.532, 0.174)
Perimeter	3.743	3.791	(-3.686, 11.173)	5.207	5.650	(-5.867, 16.280)
Smoothness	0.229	0.223	(-0.208, 0.666)	0.338	0.358	(-0.364, 1.040)
Compactness	0.075	0.264	(-0.443, 0.593)	0.048	0.335	(-0.608, 0.704)
Concavity	-0.360	0.282	(-0.913, 0.194)	-0.459	0.421	(-1.285, 0.367)
Concave Points	0.122	0.245	(-0.358, 0.603)	0.094	0.322	(-0.538, 0.7250)
Symmetry	-0.041	0.115	(-0.267, 0.185)	-0.042	0.145	(-0.327, 0.242)
Fractal Dimension	-0.399	0.256	(-0.902, 0.103)	-0.425	0.430	(-1.331, 0.336)
Mean	3.248	1.873	(-0.423, 6.920)	2.862	2.273	(-1.593, 7.317)
SE	0.286	0.403	(-0.504, 1.076)	0.304	0.446	(-0.570, 1.177)
Worst	-0.820	0.696	(-2.183, 0.544)	-0.689	0.728	(-2.117, 0.739)
Tumor Size	0.271	0.171	(-0.063, 0.606)	0.292	0.265	(-0.066, 0.650)
Intercept	-1.329	0.207	(-1.735, -0.923)	-1.617	0.183	(-2.135, -1.098)
			Scenario (iii): assuming $\tau_{10} = 5\%$			
Radius	-4.528	3.523	(-11.433,2.376)	-7.410	7.631	(-22.368, 7.548)
Texture	-0.116	0.108	(-0.327, 0.095)	-0.255	0.283	(-0.810, 0.230)
Perimeter	3.627	3.679	(-3.854, 10.839)	6.734	8.047	(-9.038, 22.507)
Smoothness	0.220	0.216	(-0.204, 0.644)	0.455	0.537	(-0.598, 1.508)
Compactness	0.077	0.258	(-0.429, 0.582)	0.031	0.405	(-0.763, 0.826)
Concavity	-0.351	0.276	(-0.891, 0.189)	-0.570	0.590	(-1.726, 0.586)
Concave Points	0.126	0.241	(-0.346, 0.597)	0.074	0.380	(-0.671, 0.820)
Symmetry	-0.040	0.112	(-0.260, 0.180)	-0.048	0.1876	(-0.414, 0.318)
Fractal Dimension	-0.250	0.545	(-0.881, 0.098)	-0.617	0.623	(-1.839, 0.605)
Mean	3.367	1.936	(-0.426, 7.161)	2.617	2.463	(-2.211, 7.445)
SE	0.293	0.414	(-0.519, 1.106)	0.310	0.460	(-0.593, 1.212)
Worst	-0.855	0.723	(-2.272, 0.562)	-0.609	0.716	(-2.012, 0.794)
Tumor Size	0.275	0.173	(-0.065, 0.614)	0.320	0.198	(-0.068, 0.708)
Intercept	-1.301	0.209	(-1.709, -0.892)	-1.866	0.363	(-2.579, -1.154)

Table 3.11: Sensitivity analyses of the second scenario ($\tau_{01} = 1\%$) of the breast cancer Wisconsin prognostic data with different degrees of τ_{10} accommodated

Parameter	Imputation Method			Likelihood Method		
	Est.	SE	95% CI	Est.	SE	95% CI
			Scenario (i): assuming $\tau_{10} = 1\%$			
Radius	-4.963	3.973	(-12.751, 2.825)	-4.796	3.884	(-12.409, 2.817)
Texture	-0.126	0.120	(-0.361, 0.109)	-0.122	0.118	(-0.354, 0.110)
Perimeter	4.094	4.158	(-4.056, 12.243)	3.918	4.041	(-4.003, 11.839)
Smoothness	0.255	0.244	(-0.224, 0.733)	0.244	0.243	(-0.233, 0.721)
Compactness	0.068	0.282	(-0.486, 0.621)	0.071	0.267	(-0.451, 0.594)
Concavity	-0.385	0.303	(-0.979, 0.210)	-0.370	0.317	(-0.991, 0.251)
Concave Points	0.112	0.254	(-0.386, 0.610)	0.113	0.278	(-0.432, 0.657)
Symmetry	-0.041	0.123	(-0.282, 0.201)	-0.041	0.113	(-0.263, 0.182)
Fractal Dimension	-0.421	0.279	(-0.967, 0.125)	-0.410	0.305	(-1.008, 0.188)
Mean	3.167	1.882	(-0.523, 6.856)	3.097	2.118	(-1.055, 7.249)
SE	0.290	0.408	(-0.509, 1.089)	0.279	0.423	(-0.551, 1.109)
Worst	-0.787	0.681	(-2.122, 0.548)	-0.775	0.753	(-2.251, 0.701)
Tumor Size	0.277	0.173	(-0.062, 0.616)	0.270	0.169	(-0.061, 0.600)
Intercept	-1.426	0.222	(-1.860, -0.991)	-1.357	0.199	(-1.748, -0.966)
			Scenario (ii): assuming $\tau_{10} = 3\%$			
Radius	-4.856	3.867	(-12.435, 2.722)	-4.806	3.887	(-12.424, 2.812)
Texture	-0.123	0.117	(-0.352, 0.105)	-0.120	0.118	(-0.352, 0.112)
Perimeter	3.980	4.044	(-3.947, 11.906)	3.933	4.046	(-3.997, 11.864)
Smoothness	0.246	0.237	(-0.219, 0.710)	0.248	0.246	(-0.234, 0.729)
Compactness	0.069	0.276	(-0.472, 0.610)	0.068	0.268	(-0.458, 0.594)
Concavity	-0.376	0.296	(-0.956, 0.204)	-0.367	0.317	(-0.987, 0.254)
Concave Points	0.115	0.250	(-0.374, 0.605)	0.104	0.278	(-0.442, 0.649)
Symmetry	-0.040	0.120	(-0.276, 0.196)	-0.039	0.113	(-0.260, 0.183)
Fractal Dimension	-0.413	0.271	(-0.945, 0.119)	-0.409	0.306	(-1.009, 0.191)
Mean	3.276	1.940	(-0.527, 7.080)	3.117	2.140	(-1.077, 7.311)
SE	0.297	0.418	(-0.523, 1.117)	0.284	0.429	(-0.558, 1.125)
Worst	-0.818	0.706	(-2.201, 0.565)	-0.778	0.759	(-2.267, 0.710)
Tumor Size	0.280	0.176	(-0.065, 0.624)	0.276	0.172	(-0.062, 0.613)
Intercept	-1.398	0.223	(-1.835, -0.961)	-1.329	0.200	(-1.722, -0.937)
			Scenario (iii): assuming $\tau_{10} = 5\%$			
Radius	-4.746	3.757	(-12.111, 2.618)	-4.814	3.885	(-12.430, 2.801)
Texture	-0.121	0.113	(-0.343, 0.102)	-0.118	0.118	(-0.349, 0.113)
Perimeter	3.861	3.928	(-3.837, 11.559)	3.947	4.047	(-3.985, 11.879)
Smoothness	0.236	0.230	(-0.215, 0.687)	0.251	0.247	(-0.234, 0.736)
Compactness	0.071	0.269	(-0.457, 0.598)	0.065	0.270	(-0.465, 0.594)
Concavity	-0.367	0.289	(-0.933, 0.199)	-0.363	0.316	(-0.983, 0.257)
Concave Points	0.119	0.245	(-0.361, 0.599)	0.094	0.279	(-0.453, 0.641)
Symmetry	-0.039	0.117	(-0.269, 0.190)	-0.037	0.113	(-0.258, 0.184
Fractal Dimension	-0.405	0.264	(-0.923, 0.113)	-0.408	0.307	(-1.009, 0.193)
Mean	3.396	2.005	(-0.534, 7.326)	3.141	2.164	(-1.101, 7.383)
SE	0.305	0.430	(-0.538, 1.148)	0.289	0.436	(-0.565, 1.143)
Worst	-0.853	0.733	(-2.289, 0.583)	-0.783	0.767	(-2.285, 0.720)
Tumor Size	0.283	0.179	(-0.067, 0.633)	0.283	0.176	(-0.063, 0.628)
Intercept	-1.370	0.225	(-1.810, -0.929)	-1.301	0.202	(-1.696, -0.906)

Chapter 4

Regularized Matrix-Variate Logistic Regression with Response Misclassification

Chapter 4 is a continuation and extension of Chapter 3. In this chapter, we consider regularized matrix-variate logistic regression with response misclassification. The remainder is organized as follows. In Section 4.1, we propose the first set of methods based on regularized unbiased estimating functions, and establish the asymptotic results for the resulting estimators. In Section 4.2, we develop the second set of methods which employ regularized observed likelihood functions. In Section 4.3, we conduct simulation studies to assess the performance of the proposed methods. We also present an application to the breast cancer Wisconsin prognostic data discussed in Section 3.4.4.

Specifically, the notation and model setup are the same as those in Chapter 3. The only difference is that the covariates contained in x_k may be unimportant in explaining the mean response. We are interested in carrying out variable selection to exclude those irrelevant covariates in inferential procedures.

4.1 Regularized Estimation Equations Method

4.1.1 With Known Misclassification Probabilities

Refer to Chapter 3, the method based in model (3.1) applies when the dimensions of x and z are small or when x and z do not include unimportant covariates. In settings with unimportant covariates, it is imperative to perform variable selection when estimating the model parameters. Let $p_{\lambda_n}(\cdot)$ denote the penalty function with the tuning parameter λ_n which often depends on n, where the argument of $p_{\lambda_n}(\cdot)$ is a scalar. For ease of exposition, we let $p_{\lambda_n}(\tilde{\alpha})$, $p_{\lambda_n}(\beta)$ and $p_{\lambda_n}(\gamma)$ represent the vectors $\{p_{\lambda_n}(\alpha_2), ..., p_{\lambda_n}(\alpha_{p+1})\}^{\mathsf{T}}$, $\{p_{\lambda_n}(\beta_1), ..., p_{\lambda_n}(\beta_q)\}^{\mathsf{T}}$ and $\{p_{\lambda_n}(\gamma_0), p_{\lambda_n}(\gamma_1), ..., p_{\lambda_n}(\gamma_{p_z})\}^{\mathsf{T}}$, respectively, by using a vector as the argument of $p_{\lambda_n}(\cdot)$ to avoid possible confusion with $p_{\lambda_n}(\cdot)$ having a scalar argument. Here we propose the penalized estimating equations method by solving a modified version of (3.6) using the unbiased surrogate Y_k^c defined in (3.5):

$$\sum_{k=1}^{n} \begin{pmatrix} U_{1k}^c(\theta; Y_k^c) - p'_{\lambda_n}(\tilde{\alpha}) \\ U_{2k}^c(\theta; Y_k^c) - p'_{\lambda_n}(\beta) \\ U_{3k}^c(\theta; Y_k^c) - p'_{\lambda_n}(\gamma) \end{pmatrix} = 0, \tag{4.1}$$

where $p'_{\lambda_n}(\tilde{\alpha})$, $p'_{\lambda_n}(\beta)$ and $p'_{\lambda_n}(\gamma)$ represent the first derivative of $p_{\lambda_n}(\tilde{\alpha})$, $p_{\lambda_n}(\beta)$ and $p_{\lambda_n}(\gamma)$, respectively.

Following Ma and Li (2010), we choose the SCAD penalty as the penalty function with the derivative function

$$p'_{\lambda_n}(\zeta) = \lambda_n \Big\{ I(|\zeta| \leq \lambda_n) + \frac{(a\lambda_n - |\zeta|)_+}{(a-1)\lambda_n} I(|\zeta| > \lambda_n) \Big\} \text{sign}(\zeta), \tag{4.2}$$

where $I(\cdot)$ is the indicator function, $\text{sign}(\zeta) = -1, 0$ and 1 when $\zeta < 0, = 0$ and > 0, respectively, and a is a constant larger than 2 with a recommended value $a = 3.7$.

To establish the asymptotic results for the resulting estimators, we let

$$a_n = \max\{|p'_{\lambda_n}(|\theta_{j0}|)| : \theta_{j0} \neq 0\} \tag{4.3}$$

and

$$b_n = \max\{|p''{}_{\lambda_n}(|\theta_{j0}|)| : \theta_{j0} \neq 0\}, \tag{4.4}$$

where θ_{j0} is the jth component of θ_0. In Appendix C.2 we show the following theorem.

Theorem 4.1 *Assume Conditions (C.1)-(C.3) in Appendix C.1. If a_n and b_n tend to 0 as $n \to \infty$, then, there exists a solution to (4.1), $\hat{\theta}_c$, such that*

$$\|\hat{\theta}_c - \theta_0\| = O_p\Big(\frac{1}{\sqrt{n}} + a_n\Big),$$

where $\|A\|$ denotes the Euclidean norm if A is a vector.

This result shows the dependence of the convergence rate of $\hat{\theta}_c$ on the tuning parameter λ_n as well as the penalty function. To obtain a $\sqrt{n}-$consistent estimator, it suffices to take a small tuning parameter so that $a_n = O(\frac{1}{\sqrt{n}})$.

Next, we discuss the *oracle* property for $\hat{\theta}_c$ whose proof is included in Appendix C.4. Write $\tilde{\alpha}_0 = (\tilde{\alpha}_{\mathrm{I}0}^{\mathsf{T}}, \tilde{\alpha}_{\mathrm{II}0}^{\mathsf{T}})^{\mathsf{T}}$, $\beta_0 = (\beta_{\mathrm{I}0}^{\mathsf{T}}, \beta_{\mathrm{II}0}^{\mathsf{T}})^{\mathsf{T}}$, and $\gamma_0 = (\gamma_{\mathrm{I}0}^{\mathsf{T}}, \gamma_{\mathrm{II}0}^{\mathsf{T}})^{\mathsf{T}}$ so that elements in $\tilde{\alpha}_{\mathrm{I}0}^{\mathsf{T}}$, $\beta_{\mathrm{I}0}^{\mathsf{T}}$ and $\gamma_{\mathrm{I}0}^{\mathsf{T}}$ are all not zero, and elements of $\tilde{\alpha}_{\mathrm{II}0}^{\mathsf{T}}$, $\beta_{\mathrm{II}0}^{\mathsf{T}}$ and $\gamma_{\mathrm{II}0}^{\mathsf{T}}$ are all zero. Write $\theta_0 = (\theta_{\mathrm{I}0}^{\mathsf{T}}, \theta_{\mathrm{II}0}^{\mathsf{T}})^{\mathsf{T}}$, where $\theta_{\mathrm{I}0} = (\tilde{\alpha}_{\mathrm{I}0}^{\mathsf{T}}, \beta_{I0}^{\mathsf{T}}, \gamma_{\mathrm{I}0}^{T})^{\mathsf{T}}$ and $\theta_{\mathrm{II}0} = (\tilde{\alpha}_{\mathrm{II}0}^{\mathsf{T}}, \beta_{\mathrm{II}0}^{\mathsf{T}}, \gamma_{\mathrm{II}0}^{\mathsf{T}})^{\mathsf{T}}$. Similar notation is defined for $\theta = (\theta_{\mathrm{I}}^{\mathsf{T}}, \theta_{\mathrm{II}}^{\mathsf{T}})^{\mathsf{T}}$ with $\theta_{\mathrm{I}} = (\tilde{\alpha}_{\mathrm{I}}^{\mathsf{T}}, \beta_{\mathrm{I}}^{\mathsf{T}}, \gamma_{\mathrm{I}}^{T})^{\mathsf{T}}$ and $\theta_{\mathrm{II}} = (\tilde{\alpha}_{\mathrm{II}}^{\mathsf{T}}, \beta_{\mathrm{II}}^{\mathsf{T}}, \gamma_{\mathrm{II}}^{\mathsf{T}})^{\mathsf{T}}$. Denote the dimension of $\tilde{\alpha}_{\mathrm{I}0}^{\mathsf{T}}$, $\beta_{\mathrm{I}0}^{\mathsf{T}}$ and $\gamma_{\mathrm{I}0}^{\mathsf{T}}$ as $d_{1\alpha}$, $d_{1\beta}$ and $d_{1\gamma}$, respectively, and the dimension of $\tilde{\alpha}_{\mathrm{II}0}^{\mathsf{T}}$, $\beta_{\mathrm{II}0}^{\mathsf{T}}$ and $\gamma_{\mathrm{II}0}^{\mathsf{T}}$ as $d_{2\alpha}$, $d_{2\beta}$ and $d_{2\gamma}$ respectively. Let $d_\alpha = d_{1\alpha} + d_{2\alpha}$, $d_\beta = d_{1\beta} + d_{2\beta}$, and $d_\gamma = d_{1\gamma} + d_{2\gamma}$, which are all assumed to be fixed. Let $U_{k\alpha,\mathrm{I}}^c(\theta; Y_k^c)$ denote the first $d_{1\alpha}$ components of $U_{1k}^c(\theta; Y_k^c)$, let $U_{k\beta,\mathrm{I}}^c(\theta; Y_k^c)$ denote the first $d_{1\beta}$ components of $U_{2k}^c(\theta; Y_k^c)$, and let $U_{k\gamma,\mathrm{I}}^c(\theta; Y_k^c)$ denote the first $d_{1\gamma}$ components of $U_{3k}^c(\theta; Y_k^c)$. Let $U_{k\alpha,\mathrm{II}}^c(\theta; Y_k^c)$ denote the last $d_{2\alpha}$ components of $U_{1k}^c(\theta; Y_k^c)$, let $U_{k\beta,\mathrm{II}}^c(\theta; Y_k^c)$ denote the last $d_{2\beta}$ components of $U_{2k}^c(\theta; Y_k^c)$, and let $U_{k\gamma,\mathrm{II}}^c(\theta; Y_k^c)$ denote the last $d_{2\gamma}$ components of $U_{3k}^c(\theta; Y_k^c)$. Write $U_{k,\mathrm{I}}^c(\theta; Y_k^c) = \{U_{k\alpha,\mathrm{I}}^{c\mathsf{T}}(; Y_k^c), U_{k\beta,\mathrm{I}}^{c\mathsf{T}}(\theta; Y_k^c), U_{k\gamma,\mathrm{I}}^{c\mathsf{T}}(\theta; Y_k^c)\}^{\mathsf{T}}$ and $U_{k,\mathrm{II}}^c(\theta; Y_k^c) = \{U_{k\alpha,\mathrm{II}}^{c\mathsf{T}}(; Y_k^c), U_{k\beta,\mathrm{II}}^{c\mathsf{T}}(\theta; Y_k^c), U_{k\gamma,\mathrm{II}}^{c\mathsf{T}}(\theta; Y_k^c)\}^{\mathsf{T}}$.

Let

$$g_\alpha = \{p'_{\lambda_n}(\tilde{\alpha}_{10}), ..., p'_{\lambda_n}(\tilde{\alpha}_{d_{1\alpha}0})\}^{\mathsf{T}},$$

$$g_\beta = \{p'_{\lambda_n}(\beta_{10}), ..., p'_{\lambda_n}(\beta_{d_{1\beta}0})\}^{\mathsf{T}},$$

$$g_\gamma = \{p'_{\lambda_n}(\gamma_{10}), ..., p'_{\lambda_n}(\gamma_{d_{1\gamma}0})\}^{\mathsf{T}},$$

$$\Sigma_\alpha = \mathrm{diag}\{p''{}_{\lambda_n}(\tilde{\alpha}_{10}), ..., p''{}_{\lambda_n}(\tilde{\alpha}_{d_{1\alpha}0})\},$$

$$\Sigma_\beta = \mathrm{diag}\{p''{}_{\lambda_n}(\beta_{10}), ..., p''{}_{\lambda_n}(\beta_{d_{1\beta}0})\},$$

and

$$\Sigma_\gamma = \mathrm{diag}\{p''{}_{\lambda_n}(\gamma_{10}), ..., p''{}_{\lambda_n}(\gamma_{d_{1\gamma}0})\}.$$

Write $g_\theta = (g_\alpha^{\mathsf{T}}, g_\beta^{\mathsf{T}}, g_\gamma^{\mathsf{T}})$ and $\Sigma_\theta = \mathrm{diag}(\Sigma_\alpha, \Sigma_\beta, \Sigma_\gamma)$. With the SCAD penalty, g_α, g_β, g_γ, Σ_α, Σ_β and Σ_γ become zero when λ_n is sufficiently small.

Theorem 4.2 *Let* $\hat{\theta}_c = (\hat{\theta}^\intercal_{c,\mathrm{I}}, \hat{\theta}^\intercal_{c,\mathrm{II}})^\intercal$ *denote a* $\sqrt{n}-$*consistent solution of* (4.1)*, where* $\hat{\theta}_{c,\mathrm{I}} = (\hat{\tilde{\alpha}}^\intercal_{c,\mathrm{I}}, \hat{\beta}^\intercal_{c,\mathrm{I}}, \hat{\gamma}^\intercal_{c,\mathrm{I}},)^\intercal$ *and* $\hat{\theta}_{c,\mathrm{II}} = (\hat{\tilde{\alpha}}^\intercal_{c,\mathrm{II}}, \hat{\beta}^\intercal_{c,II}, \hat{\gamma}^\intercal_{c,\mathrm{II}})^\intercal$ *which correspond to the subvectors of* $\theta_{\mathrm{I}0}$ *and* $\theta_{\mathrm{II}0}$*, respectively. Under Conditions (C.1)-(C.4) in Appendix C.1, if*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0^+} \sqrt{n} p'_{\lambda_n}(\theta) = \infty, \tag{4.5}$$

then with the probability tending to one, the following results hold:

(a) $\hat{\theta}_{c,\mathrm{II}} = 0$;

(b) as $n \to \infty$,

$$\sqrt{n}\left(\hat{\theta}_{c,\mathrm{I}} - \theta_{\mathrm{I}0} - \Gamma_{\mathrm{U}}(\theta_{\mathrm{I}0})^{-1} g_\theta\right) \xrightarrow{d} N\left(0, \Gamma_{\mathrm{U}}(\theta_{\mathrm{I}0})^{-1} \Sigma_{\mathrm{U}}(\theta_{\mathrm{I}0}) \Gamma_{\mathrm{U}}(\theta_{\mathrm{I}0})^{-1\intercal}\right),$$

where $\Gamma_{\mathrm{U}}(\theta_{\mathrm{I}0}) = E\left\{\frac{\partial U^c_{k,\mathrm{I}}(\theta_{\mathrm{I}0}; Y^c_k)}{\partial \theta^\intercal}\right\} - \Sigma_\theta$ *and* $\Sigma_{\mathrm{U}}(\theta_{\mathrm{I}0}) = E\{U^c_{k,\mathrm{I}}(\theta_{\mathrm{I}0}; Y^c_k) U^{c\intercal}_{k,\mathrm{I}}(\theta_{\mathrm{I}0}; Y^c_k)\}$.

Theorem 4.2(a) shows that the proposed method can correctly identify the significant row and column parameters as well as the vector covariate effects with the unimportant parameters excluded. That is, the resulting estimator possesses the oracle property. Theorem 4.2(b) establishes the asymptotic distribution for the estimators of the parameters corresponding to the important covariates, which offers the basis for performing inferences.

Finally, we comment that solving (4.1) can be implemented by modifying the block relaxation algorithm, by adding the penalty functions to (3.6) (e.g., Zhang et al. 2014). In implementing (4.1), it is critical to select a suitable value of the tuning parameter λ_n. We now describe an algorithm for selecting an optimal tuning parameter within a given set of candidates. Let

$$I_{\mathrm{F}} = E\Big\{ - \frac{\partial \ell^2_k(\theta; Y_k)}{\partial \theta \partial \theta^\intercal} \Big\}$$

be the Fisher information matrix of the likelihood function (3.2). Define the degree of freedom for the selected model to be

$$\mathrm{DF}_\lambda = \mathrm{trace}\{I_{\mathrm{F}}(I_{\mathrm{F}} + \Sigma_\theta)^{-1}\}.$$

To emphasize the dependence of λ_n, we let $\hat{\theta}_c(\lambda_n)$ denote the estimate of θ. Since Y_k is unavailable, we approximate I_{F} by

$$\hat{I}_{\mathrm{F}} = -\frac{1}{n} \sum_{k=1}^{n} \frac{\partial \ell^2_k(\theta; Y^c_k)}{\partial \theta \partial \theta^\intercal}\Big|_{\theta = \hat{\theta}_c(\lambda_n)}$$

and approximate DF_λ by $\widehat{\mathrm{DF}}_\lambda = \mathrm{trace}\{\hat{I}_\mathrm{F}(\hat{I}_\mathrm{F} + \hat{\Sigma}_\theta)^{-1}\}$, where $\hat{\Sigma}_\theta$ is the estimate of Σ_θ with θ given by $\hat{\theta}_c(\lambda_n)$. Let $\hat{\mu}_{k\lambda}$ denote the value of $\mu_{k\lambda}$ with θ specified as $\hat{\theta}_c(\lambda_n)$ and $\hat{\sigma}^2_\lambda = \frac{1}{n}\sum\limits_{k=1}^{n} |Y_k^c - \hat{\mu}_{k\lambda}|^2$. Similar to Wang et al. (2007), we define an objective function

$$\mathrm{BIC}(\lambda_n) = \mathrm{log}\hat{\sigma}^2_\lambda + \widehat{\mathrm{DF}}_\lambda \mathrm{log}(n)/n. \tag{4.6}$$

Then the optimal tuning parameter, denoted λ_n^*, is selected as the one that minimizes $\mathrm{BIC}(\lambda_n)$, and the corresponding estimate $\hat{\theta}_c(\lambda_n^*)$, denoted $\hat{\theta}_c$, is taken as the estimate of parameter θ.

4.1.2 With Unknown Misclassification Probabilities

The procedure described in Section 4.1.1 applies if the misclassification parameter ϕ is known. In applications, the values of the misclassification parameters are usually unknown and they need to be estimated from additional data sources. In this section, we consider the case with an internal validation subsample available (e.g., Chen et al. 2011, 2014) and describe the inferential procedure by incorporating estimation of the misclassification parameters. Thus, we write $U_k^c(\theta; Y_k^c)$ in Section 4.1.1 as $U_k^c(\theta, \phi; Y_k^c)$ to emphasis that ϕ is unknown parameter in these estimation equations. We apply the two-stage estimation procedure described in Section 3.2.2, $\hat{\phi}_v$ is obtained by solving (3.7).

Next, solve (4.1) for θ with ϕ replaced by $\hat{\phi}_v$ using the block relaxation algorithm. Let $\hat{\theta}_v$ denote the resulting estimator for θ. Analogous to the estimator described in Section 4.1.1, the estimator $\hat{\theta}_v$ is consistent and possesses the oracle property, shown as follows.

Theorem 4.3 *Assume Conditions (C.1)-(C.6) in Appendix C.1 hold and that p_v approaches a positive constant as $n \to \infty$. If a_n and b_n tend to 0 as $n \to \infty$, then, there exists a solution of* (4.1) *combined with* (3.7)*, $\hat{\theta}_v$, such that*

$$\|\hat{\theta}_v - \theta_0\| = O_p\Big(\frac{1}{\sqrt{n}} + a_n\Big).$$

The proof of the theorem is outlined in Appendix C.5. This theorem shows that, similarly to $\hat{\theta}_c$, the convergence rate of $\hat{\theta}_v$ depends on the tuning parameter λ_n as well as the choice of a penalty function. Taking a small tuning parameter to ensure $a_n = O(\frac{1}{\sqrt{n}})$ can yield a $\sqrt{n}-$consistent estimator of θ_0.

Let $U^c_{k\alpha,\mathrm{I}}(\theta,\phi;Y^c_k)$, $U^c_{k\beta,\mathrm{I}}(\theta,\phi;Y^c_k)$ and $U^c_{k\gamma,\mathrm{I}}(\theta,\phi;Y^c_k)$ denote the first $d_{1\alpha}$, the first $d_{1\beta}$ and the first $d_{1\gamma}$ components of $U^c_{1k}(\theta,\phi;Y^c_k)$, $U^c_{2k}(\theta,\phi;Y^c_k)$ and $U^c_{3k}(\theta,\phi;Y^c_k)$, respectively. Let $U^c_{k\alpha,\mathrm{II}}(\theta,\phi;Y^c_k)$, $U^c_{k\beta,\mathrm{II}}(\theta,\phi;Y^c_k)$ and $U^c_{k\gamma,\mathrm{II}}(\theta,\phi;Y^c_k)$ denote the last $d_{2\alpha}$, the last $d_{2\beta}$ and the first $d_{2\gamma}$ components of $U^c_{1k}(\theta,\phi;Y^c_k)$, $U^c_{2k}(\theta,\phi;Y^c_k)$ and $U^c_{3k}(\theta,\phi;Y^c_k)$, respectively. Write $U^c_{k,\mathrm{I}}(\theta,\phi;Y^c_k) = \{U^{c\intercal}_{k\alpha,\mathrm{I}}(\theta,\phi;Y^c_k), U^{c\intercal}_{k\beta,\mathrm{I}}(\theta,\phi;Y^c_k), U^{c\intercal}_{k\gamma,\mathrm{I}}(\theta,\phi;Y^c_k)\}^\intercal$ and $U^c_{k,\mathrm{II}}(\theta,\phi;Y^c_k) = \{U^{c\intercal}_{k\alpha,\mathrm{II}}(\theta,\phi;Y^c_k), U^{c\intercal}_{k\beta,\mathrm{II}}(\theta,\phi;Y^c_k), U^{c\intercal}_{k\gamma,\mathrm{I}}(\theta,\phi;Y^c_k)\}^\intercal$. Let $\hat\theta_v = (\hat\theta^\intercal_{v,\mathrm{I}}, \hat\theta^\intercal_{v,\mathrm{II}})^\intercal$, where $\hat\theta_{v,\mathrm{I}} = (\hat{\tilde\alpha}^\intercal_{v,\mathrm{I}}, \hat\beta^\intercal_{v,\mathrm{I}}, \hat\gamma^\intercal_{v,\mathrm{I}},)^\intercal$ and $\hat\theta_{v,\mathrm{II}} = (\hat{\tilde\alpha}^\intercal_{v,\mathrm{II}}, \hat\beta^\intercal_{v,\mathrm{II}}, \hat\gamma^\intercal_{v,\mathrm{II}})^\intercal$ which correspond to the subvectors of $\theta_{\mathrm{I}0}$ and $\theta_{\mathrm{II}0}$, respectively.

Theorem 4.4 *Assume Conditions (C.1)-(C.6) in Appendix C.1 hold and that p_v approaches a positive constant as $n \to \infty$, if*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0^+} \sqrt{n} p'_{\lambda_n}(\theta) = \infty, \tag{4.7}$$

then with the probability tending to one, the following results hold:

(a) $\hat\theta_{v,\mathrm{II}} = 0$;

(b) as $n \to \infty$,

$$\sqrt{n}\left(\hat\theta_{v,\mathrm{I}} - \theta_{\mathrm{I}0} - \Gamma_{\mathrm{U}^v}(\theta_0,\phi_0)^{-1} g_\theta\right) \xrightarrow{d} N\left(0, \Gamma_{\mathrm{U}^v}(\theta_0,\phi_0)^{-1}\Sigma_{\mathrm{U}^v}(\theta_{\mathrm{I}0},\phi_0)\Gamma_{\mathrm{U}^v}(\theta_0,\phi_0)^{-1\intercal}\right),$$

where $\Gamma_{\mathrm{U}^v}(\theta_0,\phi_0) = E\left\{\frac{\partial U^c_{k,\mathrm{I}}(\theta_{\mathrm{I}0},\phi_0;Y^c_k)}{\partial\theta^\intercal}\right\} - \Sigma_\theta$, $\Sigma_{\mathrm{U}^v}(\theta_{\mathrm{I}0},\phi_0) = E\{U^{*c}_{k,\mathrm{I}}(\theta_{\mathrm{I}0},\phi_0;Y^c_k)U^{*c\intercal}_{k,\mathrm{I}}(\theta_{\mathrm{I}0},\phi_0;Y^c_k)\}$ *and*

$$U^{*c}_{k,\mathrm{I}}(\theta_{\mathrm{I}0},\phi_0;Y^c_k) = U^c_{k,\mathrm{I}}(\theta_{\mathrm{I}0},\phi_0;Y^c_k) - \left\{\frac{1}{n}\sum_{k=1}^n \partial U^c_{k,\mathrm{I}}(\theta_{\mathrm{I}0},\phi_0;Y^c_k)/\partial\phi\right\} \times \left\{\frac{1}{n}\sum_{k=1}^n \delta_k \times \partial S_k(\phi_0)/\partial\phi\right\}^{-1} \times \{\delta_k S_k(\phi_0)\}.$$

The proof of Theorem 4.4 is outlined in Appendix C.6. Theorem 4.4(a) shows that the oracle property is retained for $\hat\theta_v$, just like that of $\hat\theta_c$ established in Theorem 4.2(a). Theorem 4.4(b) establishes that the estimator $\hat\theta_{v,\mathrm{I}}$ has the asymptotic normal distribution, similar to that of the estimator $\hat\theta_{c,\mathrm{I}}$ reported in Theorem 4.2(a). However, the asymptotic covariance of $\hat\theta_{v,\mathrm{I}}$ differs from that of $\hat\theta_{c,\mathrm{I}}$; the inclusion of the second term in $U^{*c}_{k,\mathrm{I}}(\theta_{\mathrm{I}0},\phi_0;Y^c_k)$ reflects the variability induced from estimation of ϕ.

4.2 Regularized Likelihood Method

4.2.1 Inference Method with Known Misclassification Probabilities

We now consider an alternative method to estimate the parameter using the observed likelihood function. Here we assume that the parameters for the misclassification probabilities are known and develop an estimation method for θ. An estimator, say $\hat{\theta}$, of θ can be obtained by maximizing the penalized log-likelihood $\sum\limits_{k=1}^{n} \ell_k^o(\theta; Y_k^*) - np_{\lambda_n}(\theta)$, where $p_{\lambda_n}(\theta)$ is a penalized function with a tuning parameter λ_n. For $k = 1, ..., n$, the log-likelihood for the observed data contributed from subject k is

$$\ell_k^o(\theta; Y_k^*) = Y_k^* \mathrm{log}\mu_k^* + (1 - Y_k^*)\mathrm{log}(1 - \mu_k^*), \tag{4.8}$$

where μ_k^* is determined by (3.9) in combination with (3.1) and (3.4).

Under regularity conditions, $\hat{\theta}$ can be equivalently obtained by solving

$$\sum_{k=1}^{n} \begin{pmatrix} U_{1k}^o(\theta; Y_k^*) - p'_{\lambda_n}(\tilde{\alpha}) \\ U_{2k}^o(\theta; Y_k^*) - p'_{\lambda_n}(\beta) \\ U_{3k}^o(\theta; Y_k^*) - p'_{\lambda_n}(\gamma) \end{pmatrix} = 0. \tag{4.9}$$

where $U_{1k}^o(\theta; Y_k^*) = \partial \ell_k^o(\theta; Y_k^*)/\partial\tilde{\alpha}$, $U_{2k}^o(\theta; Y_k^*) = \partial \ell_k^o(\theta; Y_k^*)/\partial\beta$, $U_{3k}^o(\theta; Y_k^*) = \partial \ell_k^o(\theta; Y_k^*)/\partial\gamma$.

Let $U_k^o(\theta; Y_k^*) = \{U_{1k}^{o\intercal}(\theta; Y_k^*), U_{2k}^{o\intercal}(\theta; Y_k^*), U_{3k}^{o\intercal}(\theta; Y_k^*)\}^\intercal$. Let $U_{k\alpha,\mathrm{I}}^o(\theta; Y_k^*)$, $U_{k\beta,\mathrm{I}}^o(\theta; Y_k^*)$ and $U_{\mathrm{k}\gamma,\mathrm{I}}^o(\theta; Y_k^*)$ denote the first $d_{1\alpha}$, the first $d_{1\beta}$ and the first $d_{1\gamma}$ components of $U_{1k}^o(\theta; Y_k^*)$, $U_{2k}^o(\theta; Y_k^*)$ and $U_{3k}^o(\theta; Y_k^*)$, respectively. Write $U_{k,\mathrm{I}}^o(\theta; Y_k^*) = \{U_{k\alpha,\mathrm{I}}^{o\intercal}(\theta; Y_k^*), U_{\mathrm{k}\beta,\mathrm{I}}^{o\intercal}(\theta; Y_k^*), U_{k\gamma,\mathrm{I}}^{o\intercal}(\theta; Y_k^*)\}^\intercal$. Let $\hat{\theta} = (\hat{\theta}_{\mathrm{I}}^\intercal, \hat{\theta}_{\mathrm{II}}^\intercal)^\intercal$ where $\hat{\theta}_{\mathrm{I}} = (\hat{\tilde{\alpha}}_{\mathrm{I}}^\intercal, \hat{\beta}_{\mathrm{I}}^\intercal, \hat{\gamma}_{\mathrm{I}}^\intercal)^\intercal$ and $\hat{\theta}_{\mathrm{II}} = (\hat{\tilde{\alpha}}_{\mathrm{II}}^\intercal, \hat{\beta}_{\mathrm{II}}^\intercal, \hat{\gamma}_{\mathrm{II}}^\intercal)^\intercal$ corresponding to the subvectors of $\theta_{\mathrm{I}0}$ and $\theta_{\mathrm{II}0}$, respectively. Adapting the proofs of Fan and Li (2001), We establish the following asympotitic results.

Theorem 4.5 *If the Conditions (C.1)-(C.5) in Appendix C.1 hold, and a_n and b_n tend to 0 as $n \to \infty$, then, there exists a solution to (4.9), $\hat{\theta}$, such that*

$$\|\hat{\theta} - \theta_0\| = O_p\Big(\frac{1}{\sqrt{n}} + a_n\Big).$$

Theorem 4.5 suggests that the estimator $\hat{\theta}$ has similarity to $\hat{\theta}_c$ in that choosing a small tuning parameter to ensure $a_n = O(\frac{1}{\sqrt{n}})$ can make $\hat{\theta}$ be a $\sqrt{n}-$consistent estimator of θ.

Theorem 4.6 *Under Conditions (C.1)-(C.6) in Appendix C.1, if*

$$\liminf_{n\to\infty}\liminf_{\theta\to 0^+}\sqrt{n}p'_{\lambda_n}(\theta)=\infty,$$

then with the probability tending to one, the following results hold:

(a) $\hat{\theta}_{\mathrm{II}}=0$;

(b) *as* $n\to\infty$,

$$\sqrt{n}\left(\hat{\theta}_{\mathrm{I}}-\theta_{\mathrm{I}0}-\Gamma_{\mathrm{U}^o}^{-1}g_\theta\right)\xrightarrow{d}N\left(0,\Gamma_{\mathrm{U}^o}^{-1}\Sigma\Gamma_{\mathrm{U}^o}^{-1\intercal}\right),$$

where $\Sigma=-E\{\partial U^o_{k,\mathrm{I}}(\theta_{\mathrm{I}0};Y_k^*)/\partial\theta^\intercal\}$ *and* $\Gamma_{\mathrm{U}^o}=-\Sigma-\Sigma_\theta$.

Theorem 4.6 shows that like for the estimating equation method described in Section 4.1.1, the oracle property is retained by the likelihood based method. Although both $\hat{\theta}_{c,\mathrm{I}}$ and $\hat{\theta}_{\mathrm{I}}$ have asymptotic normal distributions after certain transformations, shown in Theorems 4.2(b) and 4.6(b), their asymptotic covariance matrices are different, suggesting that they differ in efficiency, which is confirmed from the simulation studies in Section 4.3.

4.2.2 Inference Method with Unknown Misclassification Probabilities

In this subsection we extend the development in Section 4.2.1 to accommodating settings where misclassification probabilities are unknown. We consider the same setting as Section 4.1.2 where a random internal validation subsample is available.

The inference about η, defined in Section 4.1.2, can be carried out based on the likelihood function for the observed data, given by (3.12). Correspondingly, the log-likelihood function with penalty terms is

$$\ell_v^o(\eta)=\Big[\sum_{\delta_k=1}Y_k\log(\mu_k)+(1-Y_k)\log(1-\mu_k)+Y_k\log\left\{(a_{k1}(Y_k^*)\right\}+(1-Y_k)\log\left\{a_{k0}(Y_k^*)\right\}\Big]$$

$$+\Big\{\sum_{\delta_k=0}Y_k^*\log(\mu_k^*)+(1-Y_k^*)\log(1-\mu_k^*)\Big\}-np_{\lambda_n}(\theta), \tag{4.10}$$

The log-likelihood score equation for θ can be derived using (4.10),

$$\sum_{k=1}^{n} U_k^{ov}(\eta) = \sum_{\delta_k=1} U_{1k}^{ov}(\eta; Y_k) + \sum_{\delta_k=0} U_{2k}^{ov}(\eta; Y_k^*) = 0, \tag{4.11}$$

where

$$U_{1k}^{ov}(\eta; Y_k) = \left\{ \frac{Y_k - \mu_k}{\mu_i(1-\mu_i)} \right\} \left(\frac{\partial \mu_k}{\partial \theta^{\intercal}} \right)$$

and

$$U_{2k}^{ov}(\eta; Y_k^*) = \left\{ \frac{Y_k^* - \mu_k^*}{\mu_i^*(1-\mu_i^*)} \right\} \left(\frac{\partial \mu_k^*}{\partial \theta^{\intercal}} \right).$$

Then estimation of the parameter η can be carried out by a two-stage procedure. At the first stage, we employ (3.7) to obtain the estimate of ϕ using the validation subsample. At the second stage, estimation of θ is carried out by solving

$$\sum_{k=1}^{n} U_k^{ov}(\eta) - np_{\lambda_n}^{'}(\theta) = 0 \tag{4.12}$$

for θ, where $U_k^{ov}(\eta)$ is defined in (4.11), with ϕ replaced by the estimate for ϕ obtained from the first stage. Let $\hat{\theta}_{vo}$ denote the resultant estimator of θ. Asymptotic properties of $\hat{\theta}_{vo}$ can be established following the same arguments as for Theorems 4.3 and 4.4 but with different technical details.

Theorem 4.7 *Assume Conditions (C.1)-(C.5) in Appendix C.1 hold and that p_v approaches a positive constant as $n \to \infty$. If a_n and b_n tends to 0 as $n \to \infty$, then, there exists a solution to (4.12), $\hat{\theta}_{vo}$, such that*

$$\|\hat{\theta}_{vo} - \theta_0\| = O_p\Big(\frac{1}{\sqrt{n}} + a_n\Big).$$

To show the *oracle* property of $\hat{\theta}_{vo}$, let $U_{k\alpha,\mathrm{I}}^{ov}(\theta,\phi)$ denote the first $d_{1\alpha}$ components of $U_k^{ov}(\eta)$, let $U_{k\beta,\mathrm{I}}^{ov}(\theta,\phi)$ denote the components from $(d_\alpha+1)$ to $(d_\alpha+d_{1\beta})$ of $U_k^{ov}(\eta)$, and let $U_{k\gamma,\mathrm{I}}^{ov}(\theta,\phi)$ denote the components from $(d_\beta+1)$ to $(d_\beta+d_{1\gamma})$ of $U_k^{ov}(\eta)$. Write $U_{k,\mathrm{I}}^{ov}(\theta,\phi) = \{U_{k\alpha,\mathrm{I}}^{ov\intercal}(\theta,\phi), U_{k\beta,\mathrm{I}}^{ov\intercal}(\theta,\phi), U_{k\gamma,\mathrm{I}}^{ov\intercal}(\theta,\phi)\}^{\intercal}$. Let $\hat{\theta}_{vo} = (\hat{\theta}_{vo,\mathrm{I}}^{\intercal}, \hat{\theta}_{vo,\mathrm{II}}^{\intercal})^{\intercal}$, where $\hat{\theta}_{vo,\mathrm{I}} = (\hat{\alpha}_{vo,\mathrm{I}}^{\intercal}, \hat{\beta}_{vo,\mathrm{I}}^{\intercal}, \hat{\gamma}_{vo,\mathrm{I}}^{\intercal},)^{\intercal}$ and $\hat{\theta}_{vo,\mathrm{II}} = (\hat{\alpha}_{vo,\mathrm{II}}^{\intercal}, \hat{\beta}_{vo,\mathrm{II}}^{\intercal}, \hat{\gamma}_{vo,\mathrm{II}}^{\intercal})^{\intercal}$ which correspond to the subvectors of $\theta_{\mathrm{I}0}$ and $\theta_{\mathrm{II}0}$, respectively.

Theorem 4.8 *Assume Conditions (C.1)-(C.6) in Appendix C.1 hold and that p_v approaches a positive constant as $n \to \infty$, if*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0^+} \sqrt{n} p'_{\lambda_n}(\theta) = \infty,$$

then with the probability tending to one, the following results hold:

(a) $\hat{\theta}_{vo,\mathrm{II}} = 0$;

(b) as $n \to \infty$,

$$\sqrt{n}\left(\hat{\theta}_{vo,\mathrm{I}} - \theta_{\mathrm{I}0} - \Gamma_{\mathrm{U}^{ov}}(\theta_{\mathrm{I}0}, \phi_0)^{-1} g_\theta\right) \xrightarrow{d} N\left(0, \Gamma_{\mathrm{U}^{ov}}(\theta_{\mathrm{I}0}, \phi_0)^{-1} \Sigma_{\mathrm{U}^{ov}}(\theta_{\mathrm{I}0}, \phi_0) \Gamma_{\mathrm{U}^{ov}}(\theta_{\mathrm{I}0}, \phi_0)^{-1\intercal}\right).$$

where $\Gamma_{\mathrm{U}^{ov}}(\theta_{\mathrm{I}0}, \phi_0) = E\left\{\frac{\partial U^{ov}_{k,I}(\theta_{\mathrm{I}0},\phi_0)}{\partial \theta^\intercal}\right\} - \Sigma_\theta$, $\Sigma_{\mathrm{U}^{ov}}(\theta_{\mathrm{I}0}, \phi_0) = E\{U^{*ov}_{k,\mathrm{I}}(\theta_{\mathrm{I}0}, \phi_0) U^{*ov\intercal}_{k,\mathrm{I}}(\theta_{\mathrm{I}0}, \phi_0)\}$, *and*

$$U^{*ov}_{k,\mathrm{I}}(\theta_{\mathrm{I}0}, \phi_0) = U^{ov}_{k,\mathrm{I}}(\theta_{\mathrm{I}0}, \phi_0) - E\Big\{\frac{\partial U^{ov}_{k,\mathrm{I}}(\theta_{\mathrm{I}0}, \phi_0)}{\partial \phi}\Big\} \times E\Big\{\frac{\partial \delta_k S_k(\phi_0)}{\partial \phi}\Big\}^{-1} \times \{\delta_k S_k(\phi_0)\}.$$

4.3 Numerical Studies

In this section, we design different simulations to evaluate the performance of the proposed methods, in addition to assessing the impact of various degrees of response misclassification on parameter estimation. We consider settings with $p + 1 = q$, denoted as p_x for ease of exposition. The sample size is set as $n = 1000$. Five hundred simulations are run for each setting.

For $k = 1, ..., n$, we simulate x_k, z_k and Y_k using the same design as those in Section 3.4.1, but set p_x to be 5 or 10. When $p_x = 5$, we consider the same values of α and β as those in Section 3.4.1; when $p_x = 10$, we take $\alpha = (0, 1, 0, 0.5, 0, -0.5, 0, 0.5, 0, 0.5)^\intercal$ and $\beta = (0.5, -0.5, 0, 0.5, 0, 0, -0.5, 0, 0.5, 0)^\intercal$. The surrogate responses Y^*_k are generated from model (3.4) with L_k set to be constant 1. We set $\tau_{01} = \tau_{10} = 2.5\%$, 5.0%, or 10.0% to feature increasing misclassification rates. We estimate the model parameters α, β and γ using six methods. The two naive methods (called "Naive 1" and "Naive 2") discard the difference between the Y^*_k and the Y_k and fit data with model (3.1) using the block relaxation algorithm. Naive 1 employs (4.1) by replacing $U^c_{jk}(\theta; Y^c_k)$ with $U_{jk}(\theta; Y^*_k)$ for $j = 1, 2, 3$ and Naive 2 implements (3.3) with the Y_k replaced by Y^*_k. To correct for the misclassification effects, we conduct the two regularized estimation equation methods described in Sections 4.1.1 and 4.1.2 (respectively called Methods 1 and 2) and the regularized likelihood methods described in Sections 4.2.1 and 4.2.2 (respectively called Methods 3 and 4). For the methods in Sections 4.1.2 and 4.2.2, we take the internal validation sample to include 30% or 60% randomly selected subjects from the initial sample. Finally, we use the simulated true values of Y_k, x_k and z_k to fit model (3.1), and we call this the "Reference" method.

4.3.1 Simulation Results

Table 4.1 presents the results obtained from Naive 1 and Methods 1-4 for the settings with $p_x = 5$ (case 1) and $p_x = 10$ (case 2), where we report the differences for the *specificity* and *sensitivity* obtained from each of Naive 1 and Methods 1-4 minus those obtained from the reference method for the *row* and *column* parameters. The *specificity* is defined as the average of those proportions of zero coefficients that are correctly estimated to be zeros in those 500 simulations; the *sensitivity* is the average of those proportion of non-zero coefficients that are estimated to be non-zeros in those 500 simulations. It is interesting that Naive method 1 works reasonably well and produces comparable results to those obtained from Methods 1-4, suggesting that misclassification effects do not seem profound in shrinking unimportant coefficients or retaining parameters. All the methods yield fairly close values for the specificity and almost identical values for the sensitivity. As the degree of misclassification increases, the performance of all the methods tends to deteriorate. Method 3 seems to slightly outperform Method 1, and Method 4 tends to perform better than Method 1.

Furthermore, we report the simulation results for the estimators obtained from the two naive methods and Methods 1-4 in the terms of the finite sample biases in percent (bias%), empirical standard errors (ESE), model-based asymptotic standard errors (ASE), and coverage rates in percent (CR%) for 95% confidence intervals. The results for $p_x = 5$ are displayed in Tables 4.2-4.4, and the results for $p_x = 10$ are displayed in Tables 4.5-4.7.

Regarding estimation of the *row* coefficients α, all the methods yield similar results, regardless of the degrees of the misclassification or the size of internal validation data. However, for the *column* parameter β and the vector-covariate parameter γ, these methods perform differently. When misclassification is minor, the two naive methods do not seem to produce noticeably biased results. However, as the degree of misclassification increases, the two naive methods yield considerable biases. Methods 1-4 all improve the results obtained from the two naive methods. Method 3 tends to be more efficient than Method 1, and Method 4 is more efficient than Method 2, which agrees with the expectation because

Methods 3 and 4 are likelihood-based. Unsurprisingly, the performance of Methods 1-4 would deteriorate as misclassification becomes more substantial. Furthermore, we observe that Methods 1-4 perform better for $p_x = 5$ than for $p_x = 10$.

In summary, although response misclassification may not show serious effects on variable selection in our simulation studies, estimation results can be seriously biased if response misclassification is ignored in inferential procedures. The proposed methods significantly improve the performance of the naive methods and effectively account for the effects of response misclassification.

4.3.2 Analysis of the Breast Cancer Wisconsin Prognostic Data

We apply the proposed methods, in contrast to the naive approach, to analyze the breast cancer Wisconsin imaging data which we introduced in Chapter 3.

Tables 4.8 and 4.9 report the estimation results for the breast cancer Wisconsin prognostic data by fitting model (3.1) using the two naive methods as well as Methods 1 and 3. While there is no obvious pattern between the point estimates for the two naive methods, unsurprisingly, Naive method 1 yields smaller standard errors than Naive method 2. *Radius*, *Perimeter*, *Concavity* and *Fractal Dimension* are all found to be significant by Naive method 1 but not by Naive method 2. The two methods with misclassification effects accounted for yield very close point estimates, whereas the associated standard errors differ noticeably. Method 3 seems to involve more variability and tends to be less stable than Method 1. The evidence shown from Method 3 may vary with different degrees of misclassification. But Method 1 reveals the same evidence of the significance or insignificance for all the covariates, regardless of the misclassification rate.

Table 4.1: The specificity and sensitivity for row and column effects

Parameters	Model	Specif.	Sensit.	Specif.	Sensit.	Specif.	Sensit.
		$\tau_{01} = \tau_{10} = 2.5\%$		$\tau_{01} = \tau_{10} = 5\%$		$\tau_{01} = \tau_{10} = 10\%$	
		Case 1: $p_x = 5$					
Reference Method: α		0.785	1.000				
Reference Method: β		0.898	1.000				
α	Naïve Method	-0.034	0.000	-0.077	0.000	-0.142	0.000
	Method 1	-0.033	0.000	-0.070	0.000	-0.136	0.000
	Method 2 with 60% internal validation	-0.030	0.000	-0.069	0.000	-0.133	0.000
	Method 2 with 30% internal validation	-0.032	0.000	-0.070	0.000	-0.141	0.000
	Method 3	-0.023	0.000	-0.038	0.000	-0.052	0.000
	Method 4 with 60% internal validation	-0.011	0.000	-0.019	0.000	-0.020	0.000
	Method 4 with 30% internal validation	-0.016	0.000	-0.025	0.000	-0.025	0.000
β	Naïve Method	0.002	0.000	0.007	0.000	0.015	0.000
	Method 1	-0.028	0.000	-0.071	0.000	-0.129	0.000
	Method 2 with 60% internal validation	-0.024	0.000	-0.073	0.000	-0.136	0.000
	Method 2 with 30% internal validation	-0.026	0.000	-0.074	0.000	-0.132	0.000
	Method 3	-0.025	0.000	-0.032	0.000	-0.047	0.000
	Method 4 with 60% internal validation	-0.006	0.000	-0.011	0.000	-0.009	0.000
	Method 4 with 30% internal validation	-0.018	0.000	-0.019	0.000	-0.023	0.000
		Case 2: $p_x = 10$					
Reference Method: α		0.838	1.000				
Reference Method: β		0.913	1.000				
α	Naïve Method	-0.021	0.000	-0.066	0.000	-0.151	0.000
	Method 1	-0.018	0.000	-0.069	0.000	-0.152	0.000
	Method 2 with 60% internal validation	-0.018	0.000	-0.069	0.000	-0.152	0.000
	Method 2 with 30% internal validation	-0.016	0.000	-0.068	0.000	-0.152	0.000
	Method 3	-0.026	0.000	-0.070	0.000	-0.143	0.000
	Method 4 with 60% internal validation	-0.002	0.000	-0.014	0.000	-0.037	0.000
	Method 4 with 30% internal validation	-0.013	0.000	-0.046	0.000	-0.079	0.000
β	Naïve Method	0.014	0.000	0.017	0.000	0.024	0.000
	Method 1	-0.022	0.000	-0.052	0.000	-0.137	0.000
	Method 2 with 60% internal validation	-0.023	0.000	-0.053	0.000	-0.135	0.000
	Method 2 with 30% internal validation	-0.024	0.000	-0.058	0.000	-0.135	0.000
	Method 3	-0.020	0.000	-0.050	0.000	-0.126	0.000
	Method 4 with 60% internal validation	-0.006	0.000	-0.022	0.000	-0.036	0.000
	Method 4 with 30% internal validation	-0.014	0.000	-0.030	0.000	-0.079	0.000

The entries for the naive method 1 and Methods 1-4 are the difference between those method with reference method. Negative value means the method preforms worse result than reference model.

Table 4.2: Simulation results for the naive, regularized estimation equation, and regularized likelihood methods: $p_x = 5$, $\tau_{01} = \tau_{10} = 2.5\%$

	Naïve Method 1				Method 1				Method 2 with 60% internal validation				Method 2 with 30% internal validation			
Parameters	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%
α_4	-0.927	0.105	0.102	94.4	-0.964	0.105	0.102	94.6	-0.939	0.105	0.102	94.4	-0.953	0.105	0.102	94.4
α_5	-0.344	0.097	0.102	96.7	-0.334	0.097	0.102	96.7	-0.322	0.097	0.102	96.7	-0.339	0.097	0.102	96.7
β_1	-5.836	0.068	0.068	92.6	3.121	0.076	0.076	95.5	3.070	0.076	0.076	94.8	3.332	0.078	0.082	94.6
β_2	-6.554	0.071	0.068	89.9	2.323	0.080	0.076	93.8	2.286	0.079	0.076	94.0	2.499	0.081	0.081	93.0
β_4	-6.414	0.072	0.068	88.7	2.486	0.080	0.076	94.0	2.437	0.080	0.076	94.4	2.677	0.082	0.081	93.6
γ_0	-6.452	0.076	0.076	91.3	2.007	0.085	0.084	94.8	2.033	0.084	0.084	95.7	2.380	0.095	0.123	91.8
γ_1	-6.661	0.079	0.077	90.3	2.217	0.088	0.085	94.8	2.167	0.088	0.085	95.3	2.391	0.089	0.091	95.3
	Naïve Method 2				Method 3				Method 4 with 60% internal validation				Method 4 with 30% internal validation			
α_4	-0.912	0.105	0.102	94.6	-0.988	0.106	0.114	96.1	-0.703	0.103	0.099	94.1	-0.813	0.103	0.105	95.3
α_5	-0.284	0.098	0.102	96.7	-0.032	0.098	0.113	97.6	-0.013	0.094	0.099	95.9	-0.194	0.096	0.104	96.1
β_1	-5.800	0.068	0.068	93.0	2.946	0.076	0.084	96.5	2.625	0.071	0.071	93.9	3.029	0.075	0.076	94.7
β_2	-6.564	0.071	0.068	89.7	2.454	0.079	0.083	96.1	2.460	0.075	0.071	93.3	2.715	0.079	0.075	92.9
β_4	-6.395	0.072	0.068	88.2	2.042	0.079	0.084	95.5	1.821	0.073	0.071	94.1	2.059	0.077	0.076	93.5
γ_0	-6.435	0.076	0.076	91.3	1.745	0.084	0.084	94.5	1.694	0.084	0.078	93.9	1.929	0.092	0.083	92.3
γ_1	-6.627	0.079	0.077	90.7	1.896	0.088	0.094	95.3	1.346	0.084	0.080	92.7	1.769	0.087	0.085	93.5

Table 4.3: Simulation results for the naive, regularized estimation equation, and regularized likelihood methods: $p_x = 5$, $\tau_{01} = \tau_{10} = 5\%$

	Naïve Method 1				Method 1				Method 2 with 60% internal validation				Method 2 with 30% internal validation			
Parameters	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%
α_4	-0.169	0.115	0.110	94.2	-0.155	0.115	0.110	94.0	-0.156	0.115	0.110	94.2	-0.146	0.115	0.110	94.2
α_5	-0.389	0.106	0.110	95.8	-0.399	0.106	0.110	95.6	-0.393	0.106	0.110	95.6	-0.384	0.106	0.110	95.6
β_1	-13.915	0.067	0.067	79.4	3.102	0.084	0.083	95.8	3.186	0.084	0.083	95.6	3.428	0.088	0.084	94.4
β_2	-14.806	0.069	0.067	76.8	2.063	0.086	0.083	95.6	2.149	0.087	0.083	95.2	2.421	0.092	0.084	93.6
β_4	-14.588	0.068	0.067	77.2	2.343	0.085	0.083	94.2	2.407	0.085	0.083	94.4	2.701	0.090	0.084	93.0
γ_0	-14.018	0.072	0.074	73.1	2.161	0.089	0.092	95.4	2.141	0.088	0.092	96.4	2.333	0.107	0.092	90.2
γ_1	-14.880	0.078	0.075	82.0	1.987	0.096	0.093	94.2	2.009	0.095	0.093	94.0	2.255	0.098	0.093	94.6
	Naïve Method 2				Method 3				Method 4 with 60% internal validation				Method 4 with 30% internal validation			
α_4	-0.173	0.115	0.110	94.2	-0.980	0.107	0.116	95.9	-0.858	0.103	0.099	93.1	-0.800	0.104	0.106	95.5
α_5	-0.332	0.107	0.110	95.8	-0.058	0.098	0.115	97.2	-0.009	0.094	0.099	96.1	-0.285	0.096	0.105	96.6
β_1	-13.922	0.068	0.067	79.0	3.103	0.076	0.086	96.6	2.675	0.072	0.071	93.3	3.113	0.075	0.076	94.3
β_2	-14.806	0.069	0.067	76.8	2.333	0.080	0.084	96.1	2.430	0.075	0.070	93.1	2.624	0.080	0.075	92.5
β_4	-14.547	0.068	0.067	77.0	2.035	0.080	0.086	95.5	1.836	0.073	0.070	94.5	2.062	0.077	0.076	93.3
γ_0	-13.991	0.072	0.074	73.3	1.850	0.085	0.086	94.5	1.716	0.085	0.078	92.9	1.819	0.096	0.083	91.5
γ_1	-14.858	0.078	0.075	82.0	1.908	0.090	0.096	95.7	1.267	0.085	0.079	92.3	1.696	0.087	0.086	93.3

Table 4.4: Simulation results for the naive, regularized estimation equation, and regularized likelihood methods: $p_x = 5$, $\tau_{01} = \tau_{10} = 10\%$

	Naïve Method 1				Method 1				Method 2 with 60% internal validation				Method 2 with 30% internal validation			
Parameters	Bias%	ESE	ASE	CR%	Bias%	ESE	ASE	CR%	Bias%	ESE	ASE	CR%	Bias%	ESE	ASE	CR%
α_4	0.113	0.136	0.128	93.4	0.083	0.136	0.128	93.4	0.109	0.136	0.128	93.4	0.095	0.136	0.128	93.2
α_5	-0.299	0.125	0.128	96.0	-0.231	0.125	0.128	95.8	-0.182	0.125	0.128	95.8	-0.188	0.125	0.128	95.8
β_1	-27.300	0.065	0.064	42.0	4.605	0.101	0.099	94.6	4.598	0.100	0.100	94.4	5.275	0.108	0.101	94.6
β_2	-28.582	0.068	0.064	39.6	2.796	0.106	0.099	92.6	2.722	0.104	0.099	93.8	3.464	0.112	0.101	93.0
β_4	-27.924	0.066	0.064	41.8	3.745	0.102	0.099	94.0	3.740	0.101	0.099	95.6	4.460	0.109	0.101	94.8
γ_0	-27.207	0.070	0.071	24.4	3.090	0.108	0.109	95.6	2.864	0.098	0.110	98.0	3.571	0.136	0.112	91.4
γ_1	-28.312	0.073	0.072	49.2	3.171	0.112	0.110	94.2	3.101	0.109	0.111	95.0	3.665	0.114	0.112	93.6
	Naïve Method 2				Method 3				Method 4 with 60% internal validation				Method 4 with 30% internal validation			
α_4	0.147	0.136	0.128	93.4	-0.853	0.115	0.122	95.9	-0.773	0.105	0.099	92.9	-0.571	0.109	0.107	94.7
α_5	-0.263	0.125	0.128	96.0	-0.596	0.100	0.121	97.0	-0.262	0.096	0.099	95.9	-0.678	0.099	0.106	96.1
β_1	-27.302	0.065	0.064	42.0	3.453	0.080	0.092	96.8	2.539	0.073	0.070	92.9	3.000	0.077	0.077	93.7
β_2	-28.582	0.068	0.064	40.2	2.352	0.085	0.090	96.1	2.372	0.075	0.070	93.1	2.411	0.081	0.076	92.5
β_4	-27.880	0.066	0.064	41.6	2.705	0.083	0.091	95.5	2.107	0.074	0.070	93.7	2.324	0.078	0.076	92.5
γ_0	-27.171	0.070	0.071	24.8	2.168	0.089	0.089	94.7	1.696	0.086	0.078	92.1	1.610	0.099	0.084	90.9
γ_1	-28.288	0.073	0.072	49.2	2.344	0.092	0.102	95.9	1.313	0.085	0.079	93.1	1.673	0.088	0.086	93.9

Table 4.5: Simulation results for the naive, regularized estimation equation, and regularized likelihood methods: $p_x = 10$, $\tau_{01} = \tau_{10} = 2.5\%$

	Naïve Method 1				Method 1				Method 2 with 60% internal validation				Method 2 with 30% internal validation			
Parameters	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%
α_4	-0.720	0.090	0.085	94.14	-0.709	0.090	0.085	94.17	-0.710	0.090	0.085	94.17	-0.713	0.090	0.085	94.17
α_6	0.684	0.087	0.085	93.51	0.650	0.087	0.085	93.54	0.653	0.087	0.085	93.54	0.654	0.087	0.085	93.33
α_8	1.884	0.085	0.085	94.98	1.892	0.085	0.085	95.21	1.878	0.085	0.085	95.21	1.898	0.085	0.085	95.00
α_{10}	0.791	0.087	0.085	95.61	0.812	0.087	0.085	95.63	0.819	0.087	0.085	95.63	0.812	0.087	0.085	95.63
β_1	-7.848	0.069	0.065	86.40	3.549	0.080	0.075	93.13	3.695	0.080	0.075	93.96	3.939	0.081	0.078	93.13
β_2	-7.920	0.069	0.065	87.87	3.528	0.080	0.075	92.71	3.650	0.079	0.075	92.92	3.898	0.081	0.078	93.75
β_4	-7.518	0.065	0.065	88.91	3.944	0.076	0.075	94.58	4.143	0.077	0.075	95.00	4.413	0.079	0.078	93.13
β_7	-8.015	0.065	0.064	89.12	3.356	0.075	0.075	95.21	3.521	0.076	0.075	95.21	3.741	0.077	0.077	95.00
β_9	-7.240	0.065	0.065	88.91	4.219	0.075	0.075	95.21	4.397	0.076	0.075	95.00	4.630	0.077	0.078	96.04
γ_0	-8.124	0.084	0.082	88.08	2.851	0.097	0.095	94.58	3.365	0.099	0.095	93.54	3.660	0.112	0.109	91.04
γ_1	-7.075	0.087	0.082	91.21	4.346	0.100	0.094	94.58	4.547	0.101	0.094	93.75	4.765	0.102	0.096	93.96
	Naïve Method 1				Method 3				Method 4 with 60% internal validation				Method 4 with 30% internal validation			
α_4	-0.696	0.090	0.085	94.17	-0.404	0.089	0.085	94.09	-0.589	0.084	0.076	92.46	-0.703	0.087	0.078	93.48
α_6	0.620	0.087	0.085	93.33	0.621	0.087	0.085	94.30	0.538	0.083	0.076	92.67	0.456	0.085	0.078	92.67
α_8	1.895	0.085	0.085	95.42	1.816	0.086	0.085	94.91	1.671	0.082	0.077	93.48	1.662	0.084	0.078	93.48
α_{10}	0.818	0.087	0.085	95.63	1.198	0.088	0.085	95.52	1.255	0.085	0.077	93.48	1.073	0.086	0.078	93.08
β_1	-7.802	0.069	0.065	86.46	3.496	0.080	0.074	94.09	3.242	0.073	0.063	90.02	3.636	0.078	0.065	89.21
β_2	-7.835	0.069	0.065	88.13	3.107	0.079	0.074	92.67	3.072	0.074	0.063	89.61	3.483	0.077	0.065	89.21
β_4	-7.469	0.065	0.065	89.58	4.003	0.076	0.074	94.30	3.520	0.074	0.063	90.43	4.028	0.076	0.065	89.41
β_7	-7.953	0.065	0.065	89.38	2.876	0.074	0.074	95.72	2.594	0.070	0.063	91.45	3.068	0.075	0.065	91.04
β_9	-7.201	0.065	0.065	88.96	3.979	0.077	0.074	95.11	3.407	0.073	0.063	90.02	3.811	0.075	0.065	89.82
γ_0	-8.142	0.084	0.082	88.54	3.003	0.097	0.094	94.50	2.996	0.092	0.080	89.61	3.190	0.105	0.083	86.97
γ_1	-7.120	0.087	0.082	90.63	4.480	0.099	0.093	93.48	3.849	0.093	0.081	89.61	4.425	0.097	0.083	89.61

Table 4.6: Simulation results for the naive, regularized estimation equation, and regularized likelihood methods:$p_x = 10$, $\tau_{01} = \tau_{10} = 5\%$

	Naïve Method 1				Method 1				Method 2 with 60% internal validation				Method 2 with 30% internal validation			
Parameters	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%
α_4	-0.051	0.097	0.092	94.78	-0.098	0.097	0.092	94.78	-0.091	0.097	0.092	94.98	-0.088	0.097	0.092	94.98
α_6	0.808	0.093	0.092	94.78	0.817	0.093	0.092	95.18	0.827	0.093	0.092	95.38	0.820	0.093	0.092	95.38
α_8	2.189	0.091	0.092	95.78	2.180	0.092	0.093	95.78	2.168	0.092	0.093	95.98	2.176	0.092	0.093	95.78
α_{10}	1.085	0.096	0.093	94.18	1.126	0.096	0.093	94.58	1.143	0.097	0.093	94.38	1.118	0.097	0.093	94.38
β_1	-17.002	0.067	0.062	68.47	4.358	0.090	0.084	94.38	4.407	0.092	0.084	93.57	5.023	0.094	0.085	92.17
β_2	-17.605	0.067	0.062	64.46	3.619	0.090	0.083	93.37	3.604	0.089	0.083	93.37	4.240	0.093	0.084	92.97
β_4	-16.584	0.063	0.063	70.28	4.867	0.085	0.084	93.37	4.922	0.087	0.084	93.37	5.544	0.090	0.085	93.37
β_7	-17.317	0.062	0.062	68.27	3.892	0.083	0.083	94.98	3.918	0.084	0.083	94.98	4.543	0.087	0.084	94.38
β_9	-16.703	0.064	0.062	69.68	4.658	0.086	0.083	94.18	4.655	0.086	0.084	94.58	5.294	0.089	0.085	93.37
γ_0	-17.122	0.082	0.079	65.66	3.569	0.109	0.105	93.57	4.061	0.107	0.106	93.98	5.147	0.126	0.107	91.16
γ_1	-16.724	0.083	0.079	80.32	4.740	0.108	0.104	93.98	4.800	0.110	0.104	94.58	5.349	0.112	0.105	95.18
	Naïve Method 2				Method 3				Method 4 with 60% internal validation				Method 4 with 30% internal validation			
α_4	-0.043	0.097	0.093	94.98	-0.154	0.096	0.091	94.20	-0.619	0.086	0.075	91.80	-0.677	0.091	0.078	91.00
α_6	0.822	0.093	0.092	95.18	0.675	0.092	0.091	94.80	0.639	0.085	0.075	91.80	0.486	0.088	0.078	92.20
α_8	2.210	0.091	0.093	95.78	1.950	0.091	0.091	96.20	1.771	0.085	0.075	91.20	1.853	0.089	0.078	91.20
α_{10}	1.131	0.096	0.093	94.38	1.231	0.094	0.091	94.60	1.296	0.084	0.075	92.40	0.951	0.089	0.078	91.80
β_1	-16.984	0.067	0.062	68.67	4.470	0.088	0.081	94.00	3.355	0.075	0.060	89.80	3.646	0.083	0.063	85.20
β_2	-17.601	0.067	0.062	64.26	3.433	0.089	0.081	92.60	2.708	0.077	0.060	87.20	3.093	0.084	0.063	85.40
β_4	-16.581	0.063	0.063	70.88	5.320	0.084	0.081	94.20	3.707	0.076	0.060	88.40	4.316	0.080	0.063	87.00
β_7	-17.290	0.062	0.062	68.67	3.844	0.081	0.081	95.60	2.693	0.073	0.060	88.40	3.291	0.081	0.063	87.60
β_9	-16.705	0.064	0.062	69.88	4.713	0.085	0.081	94.00	3.433	0.074	0.060	88.00	3.732	0.079	0.063	88.40
γ_0	-17.100	0.082	0.079	66.06	3.995	0.109	0.103	93.60	3.390	0.096	0.077	86.60	3.266	0.116	0.081	83.00
γ_1	-16.695	0.083	0.080	80.32	5.023	0.108	0.102	93.80	3.680	0.095	0.078	89.80	4.207	0.102	0.081	88.00

Table 4.7: Simulation results for the naive, regularized estimation equation, and regularized likelihood methods:$p_x = 10$, $\tau_{01} = \tau_{10} = 10\%$

	Naïve Method 1				Method 1				Method 2 with 60% internal validation				Method 2 with 30% internal validation			
Parameters	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%	Bias%	SEE	ASE	CR%
α_4	-0.053	0.113	0.108	93.20	-0.138	0.113	0.108	93.60	-0.148	0.113	0.108	93.60	-0.154	0.113	0.108	93.40
α_6	1.020	0.111	0.108	94.80	0.988	0.111	0.109	94.00	0.974	0.111	0.109	94.20	0.982	0.111	0.109	94.40
α_8	2.219	0.113	0.109	95.80	2.189	0.113	0.109	96.00	2.162	0.113	0.109	95.80	2.171	0.113	0.109	96.00
α_{10}	0.875	0.114	0.108	93.60	0.929	0.114	0.109	93.20	0.895	0.114	0.109	93.20	0.859	0.114	0.109	93.40
β_1	-31.947	0.061	0.059	22.80	6.767	0.108	0.104	94.60	7.041	0.110	0.105	93.60	8.189	0.118	0.109	93.20
β_2	-32.578	0.063	0.058	23.00	5.751	0.111	0.104	93.80	5.833	0.108	0.104	94.80	7.156	0.121	0.108	93.80
β_4	-31.773	0.061	0.059	24.60	6.960	0.107	0.104	94.20	7.279	0.110	0.105	95.00	8.576	0.123	0.109	94.00
β_7	-32.014	0.061	0.058	23.80	6.563	0.106	0.104	94.80	6.863	0.110	0.105	94.80	8.146	0.121	0.109	93.60
β_9	-32.190	0.063	0.059	26.00	6.279	0.110	0.104	94.40	6.610	0.115	0.105	94.80	7.773	0.125	0.110	93.20
γ_0	-32.270	0.078	0.075	16.40	5.017	0.135	0.131	95.80	5.862	0.126	0.132	95.80	7.997	0.176	0.139	89.00
γ_1	-32.107	0.082	0.075	43.00	6.591	0.138	0.128	93.60	6.705	0.136	0.129	93.80	7.787	0.143	0.133	93.80
	Naïve Method 2				Method 3				Method 4 with 60% internal validation				Method 4 with 30% internal validation			
α_4	-0.035	0.113	0.108	93.20	0.222	0.112	0.105	93.80	-0.590	0.090	0.072	88.60	-0.384	0.099	0.077	87.20
α_6	0.981	0.110	0.109	95.00	1.197	0.110	0.105	93.60	0.738	0.088	0.072	89.20	0.802	0.097	0.077	87.60
α_8	2.202	0.112	0.109	95.60	2.237	0.111	0.105	94.40	1.873	0.090	0.072	87.80	1.879	0.098	0.078	88.00
α_{10}	0.857	0.114	0.109	93.40	1.213	0.110	0.105	93.80	1.072	0.091	0.072	88.40	1.221	0.100	0.078	87.20
β_1	-31.931	0.061	0.059	23.40	7.216	0.106	0.098	92.40	4.051	0.077	0.056	85.40	4.519	0.093	0.061	78.80
β_2	-32.561	0.063	0.059	23.20	5.840	0.108	0.097	92.00	3.249	0.080	0.056	82.00	4.094	0.091	0.061	79.00
β_4	-31.757	0.061	0.059	24.60	7.325	0.105	0.098	93.40	3.872	0.080	0.057	84.00	4.596	0.089	0.062	80.60
β_7	-31.993	0.061	0.059	24.20	6.292	0.103	0.097	93.60	3.408	0.078	0.056	84.80	3.783	0.092	0.061	82.40
β_9	-32.163	0.063	0.059	26.20	6.253	0.106	0.097	91.80	3.872	0.078	0.056	83.40	3.993	0.090	0.061	82.00
γ_0	-32.247	0.078	0.075	16.80	5.509	0.131	0.123	94.80	3.137	0.104	0.073	84.40	3.296	0.135	0.080	75.20
γ_1	-32.089	0.082	0.075	43.40	7.160	0.134	0.121	90.80	4.204	0.099	0.073	83.80	5.206	0.115	0.080	81.40

Table 4.8: Analysis of the breast cancer Wisconsin data without accounting for misclassification effects

	Naive Method 1			Naive Method 2		
Parameter	Est.	SE	95% CI	Est.	SE	95% CI
Radius	-4.551	0.344	(-5.225, -3.877)	-4.790	3.778	(-12.195, 2.614)
Texture	-0.048	0.032	(-0.110, 0.014)	-0.122	0.115	(-0.348, 0.104)
Perimeter	3.732	0.319	(3.107, 4.356)	3.909	3.951	(-3.835, 11.654)
Smoothness	0.268	0.144	(-0.015, 0.551)	0.242	0.233	(-0.214, 0.698)
Compactness	0.000	0.000	(-0.000, 0.000)	0.073	0.273	(-0.463, 0.609)
Concavity	-0.315	0.149	(-0.607, -0.022)	-0.372	0.293	(-0.945, 0.202)
Concave Points	0.000	0.000	(-0.000, 0.000)	0.117	0.251	(-0.376, 0.610)
Symmetry	0.000	0.000	(-0.000, 0.000)	-0.041	0.119	(-0.276, 0.193)
Fractal Dimension	-0.342	0.125	(-0.587, -0.096)	-0.411	0.266	(-0.933, 0.111)
Mean	3.481	0.998	(1.524, 5.438)	3.089	1.791	(-0.422, 6.599)
SE	0.313	0.300	(-0.274, 0.900)	0.276	0.388	(-0.484, 1.037)
Worst	-0.818	0.418	(-1.638, 0.001)	-0.773	0.660	(-2.066, 0.520)
Tumor Size	0.271	0.165	(-0.054, 0.595)	0.267	0.167	(-0.060, 0.594)
Intercept	-1.357	0.196	(-1.741, -0.974)	-1.371	0.205	(-1.772, -0.969)

Entries of 95% CI with the form 0.000 are positive and very close to zero; entries with the from -0.000 are negative and very close to zero.

Table 4.9: Sensitivity analyses of the breast cancer Wisconsin data

Parameter	Method 1			Method 3		
	Est.	SE	95% CI	Est.	SE	95% CI
Scenario (i): assuming $\tau_{10} = 1\%$						
Radius	-4.534	0.351	(-5.222, -3.846)	-4.555	5.008	(-14.370, 5.260)
Texture	-0.049	0.032	(-0.112, 0.014)	-0.052	0.320	(-0.680, 0.576)
Perimeter	3.714	0.327	(3.073, 4.354)	3.735	5.237	(-6.529, 13.999)
Smoothness	0.266	0.143	(-0.015, 0.547)	0.266	0.253	(-0.229, 0.761)
Compactness	0.000	0.000	(-0.000, 0.000)	0.000	0.000	(-0.000, 0.000)
Concavity	-0.312	0.149	(-0.603, -0.021)	-0.312	0.363	(-1.024, 0.400)
Concave Points	0.000	0.000	(-0.000, 0.000)	0.000	0.000	(-0.000, 0.000)
Symmetry	0.000	0.000	(-0.000, 0.000)	0.000	0.000	(-0.000, 0.000)
Fractal Dimension	-0.340	0.124	(-0.583, -0.096)	-0.341	0.410	(-1.146, 0.463)
Mean	3.518	1.006	(1.546, 5.489)	3.503	3.928	(-4.196, 11.202)
SE	0.317	0.302	(-0.276, 0.910)	0.317	0.563	(-0.785, 1.420)
Worst	-0.828	0.421	(-1.653, -0.002)	-0.825	1.247	(-3.269, 1.619)
Tumor Size	0.272	0.167	(-0.054, 0.599)	0.274	0.170	(-0.059, 0.608)
Intercept	-1.344	0.196	(-1.729, -0.959)	-1.345	0.202	(-1.740, -0.950)
Scenario (2): assuming $\tau_{10} = 3\%$						
Radius	-4.499	0.366	(-5.218, -3.781)	-4.566	0.578	(-5.699, -3.432)
Texture	-0.050	0.033	(-0.116, 0.015)	-0.056	0.376	(-0.794, 0.681)
Perimeter	3.677	0.343	(3.004, 4.350)	3.745	0.551	(2.665, 4.824)
Smoothness	0.262	0.142	(-0.016, 0.540)	0.264	0.163	(-0.055, 0.584)
Compactness	0.000	0.000	(-0.000, 0.000)	0.000	0.000	(-0.000, 0.000)
Concavity	-0.307	0.147	(-0.596, -0.019)	-0.308	0.193	(-0.686, 0.070)
Concave Points	0.000	0.000	(-0.000, 0.000)	0.000	0.000	(-0.000, 0.000)
Symmetry	0.000	0.000	(-0.000, 0.000)	0.000	0.000	(-0.000, 0.000)
Fractal Dimension	-0.336	0.122	(-0.576, -0.096)	-0.342	0.124	(-0.585, 0.099)
Mean	3.595	1.021	(1.594, 5.597)	3.540	1.275	(1.040, 6.039)
SE	0.324	0.308	(-0.280, 0.929)	0.324	0.351	(-0.363, 1.011)
Worst	-0.847	0.428	(-1.686, -0.008)	-0.835	0.467	(-1.751, 0.081)
Tumor Size	0.276	0.169	(-0.056, 0.608)	0.283	0.176	(-0.062, 0.627)
Intercept	-1.317	0.198	(-1.706, -0.929)	-1.318	0.201	(-1.712, -0.925)
Scenario (iii): assuming $\tau_{10} = 5\%$						
Radius	-4.464	0.366	(-5.183, -3.746)	-4.531	0.334	(-5.187, -3.876)
Texture	-0.052	0.033	(-0.118, 0.013)	-0.053	0.141	(-0.330, 0.225)
Perimeter	3.640	0.343	(2.968, 4.313)	3.707	0.303	(3.113, 4.301)
Smoothness	0.258	0.142	(-0.020, 0.535)	0.262	0.027	(0.210, 0.314)
Compactness	0.000	0.000	(-0.000, 0.000)	0.000	0.000	(-0.000, 0.000)
Concavity	-0.302	0.147	(-0.591, -0.014)	-0.303	0.037	(-0.376, -0.230)
Concave Points	0.000	0.000	(-0.000, 0.000)	0.000	0.000	(-0.000, 0.000)
Symmetry	0.000	0.000	(-0.000, 0.000)	0.000	0.000	(-0.000, 0.000)
Fractal Dimension	-0.332	0.122	(-0.572, -0.092)	-0.339	0.015	(-0.369, -0.309)
Mean	3.677	1.021	(1.676, 5.679)	3.609	1.626	(0.422, 6.797)
SE	0.332	0.308	(-0.273, 0.937)	0.326	0.123	(0.085, 0.567)
Worst	-0.867	0.428	(-1.706, -0.028)	-0.850	0.218	(-1.277, -0.422)
Tumor Size	0.280	0.169	(-0.052, 0.612)	0.291	0.031	(0.230, 0.351)
Intercept	-1.289	0.198	(-1.678, -0.901)	-1.289	0.040	(-1.368, -1.210)

Entries of 95% CI with the form 0.000 are positive and very close to zero; entries with the from -0.000 are negative and very close to zero.

Chapter 5

Bayesian Analysis for Matrix-Variate Logistic Regression with/without Response Misclassification

In this chapter, we propose a Bayesian inference procedure to the matrix-variate logistic regression using horseshoe prior under matrix-variate logistic regression with the help of augmented data from Pólya-Gamma distribution. Meanwhile, we develop a Bayesian estimation procedure with missclassification on response. The remainder is organized as follows. In Section 5.1, we introduce the model setup related to model (2.1). In Section 5.2, we propose the Bayesian inference method for the error-free context. In Section 5.3, we develop the Bayesian estimation procedure with missclassification on response. In Section 5.4, we conduct simulation studies to assess the performance of the methods developed in Sections 5.2 and 5.3, as well as to demonstrate the biased effects of ignoring response missclassification. We also present an application to a LSVT data set in Section 5.5.

5.1 Matrix-variate Logistic Regression Model

For subject k with $k = 1, ..., n$, Y_k is defined in the same way as in Section 2.1.1. Write $\mathbb{Y} = (Y_1, ..., Y_n)^\intercal$. Let $x_k = [x_{k,ij}]_{p\times q}$ be the associated $p \times q$ covariate matrix where $x_{k,ij}$ is the observation at row i and column j for subject k.

Employing an assumed rank-R parafac decomposition (Guhaniyogi et al. 2017) to $\mathcal{B}$ in model (2.1) gives

$$\mathcal{B} = \sum_{r=1}^{R} \alpha^{(r)} \circ \beta^{(r)}, \tag{5.1}$$

where $\circ$ denotes the outer product, $\alpha^{(r)}$ is a $p \times 1$ row parameter vector, $\beta^{(r)}$ is a $q \times 1$ column parameter vector, and R is the positive integer so that $\mathcal{B}$ cannot be written as a sum of less than R outer products (Zhou et al. 2013).

With (5.1), model (2.1) becomes

$$\text{logit} P(Y_k = 1 | x_k) = \left\langle x_k, \sum_{r=1}^{R} \alpha^{(r)} \circ \beta^{(r)} \right\rangle. \tag{5.2}$$

When a rank-1 (i.e., $R = 1$) parafac decomposition is applied to $\mathcal{B}$, model (5.2) reduces to the matrix-variate logistic regression model (2.2) with $\gamma = 0$, the model considered by Hung and Wang (2013).

Finally, we comment that in model (5.2), the coefficients $\alpha^{(r)}$ and $\beta^{(r)}$ are not identifiable for $r = 1, ..., R$. For instance, respectively scaling $\alpha^{(r)}$ and $\beta^{(r)}$ by any nonzero constant c and its reciprocal $1/c$ makes (5.2) hold. However, if our interest focuses on $\mathcal{B}$ itself, nonidentifiability of the coefficients $\alpha^{(r)}$ and $\beta^{(r)}$ does not pose a concern, especially in the context of Bayesian analysis, as discussed by Guhaniyogi et al. (2017).

5.2 Bayesian Inference Procedure

We are interested in inference about $\mathcal{B}$ in model (2.1) via the formulation model (5.2) through a Bayesian approach. We denote the conditional probability density function (p.d.f.) of the response variable Y_k, given x_k, as $p_{Y_k|x_k}(y_k, x_k; \mathcal{B}, \gamma)$ or p_k for simplicity, i.e.,

$$p_k = \frac{\exp(< x_k, \mathcal{B} >)}{1 + \exp(< x_k, \mathcal{B} >)}. \tag{5.3}$$

In this subsection, we describe a Bayesian approach based on a family of Pólya-Gamma distributions (Polson et al. 2013), where the Gibbs sampler procedure is used, together with the specification of the prior distribution for parameters $\alpha^{(r)}$ and $\beta^{(r)}$ in model (5.2) for $r = 1, ..., R$.

Let $\pi(\alpha^{(r)})$ and $\pi(\beta^{(r)})$ denote the prior densities for $\alpha^{(r)}$ and $\beta^{(r)}$, respectively, and write the posterior densities of $\alpha^{(r)}$ and $\beta^{(r)}$, given the observed data $\{\mathbb{Y}, x\}$, as $\pi(\alpha^{(r)}|\mathbb{Y}, x; \beta^{(r)}, \mathcal{B}_{-r})$ and $\pi(\beta^{(r)}|\mathbb{Y}, x; \alpha^{(r)}, \mathcal{B}_{-r})$, respectively, where $\mathcal{B}_{-r} = \{\mathcal{B}^l : l \neq r\}$ for $r = 1, ..., R$.

With the prior densities for $\alpha^{(r)}$, $\beta^{(r)}$ and model (2.1) together with (5.1), the posterior density for $\mathcal{B}$, given the data, is possible to construct, at least in principle. However, the actual calculation of the posterior density is not trivial due to the lack of its closed form; even with the application of approximations, such as the Markov chain Monte Carlo (MCMC) method, this can be computationally difficult. To get around these issues, Polson et al. (2013) developed a data augmentation algorithm for logistic regression which is simple and fast to implement. The idea is to introduce an independent random variable, say W, as an intermediate tool to form a Pólya-Gamma distribution, then the posterior densities of α and β, given the data and W, have a normal distribution which is easy to handle.

5.2.1 Pólya-Gamma Distribution with Logistic Regression

Here we describe the connection between the Pólya-Gamma distribution and the logistic regression model. A random variable U follows a Pólya-Gamma distribution, $PG(1, c)$ for $c \geq 0$, if it has the density function

$$f(u|c) = \cosh\Big(\frac{1}{2}\Big) \exp\Big(-\frac{c^2u}{2}\Big) g(u),$$

where $g(u)$ is given by

$$g(u) = \sum_{k=0}^{\infty} (-1)^k \frac{(2k+1)}{\sqrt{2\pi u^3}} \exp\left\{-\frac{(2k+1)^2}{8u}\right\} \mathbf{I}_{(0,\infty)}(u),$$

with $\mathbf{I}_{(0,\infty)}(u)$ defined as 1 for $0 < u < \infty$ and 0 otherwise (Biane et al. 2001).

Next, we make a connection of the Pólya-Gamma Distribution with model (5.2) via (2.1) and (5.1). Let $W_1, ..., W_n$ be independent of each other and of the Y_k, each having a Pólya-Gamma distribution with $W_k \sim PG(1, c_k)$ where $c_k = | < x_k, \mathcal{B} > |$ with covariates x_k fixed. Then the joint probability density function for $W = (W_1, ...W_n)^\intercal$, $f(w|\mathcal{B})$, indexed by $\mathcal{B}$, is given by $\prod_{k=1}^n f(w_k|c_k)$.

Using the intermediate variables W_k, we augment the observed data $\{\mathbb{Y}, x\}$ with W and construct the augmented posterior densities for the parameters by combining model (5.2) with the priors of $\alpha^{(r)}$ and $\beta^{(r)}$, which are straightforward to analyze (Tanner and Wong

1987). For example, given the covariates, the posterior density for $\alpha^{(r)}$ is determined by

$$\pi(\alpha^{(r)}|\mathbb{Y},x;\beta^{(r)},\mathcal{B}_{-r}) = \int_{R_+^n} \pi(\alpha^{(r)},w|\mathbb{Y},x;\beta^{(r)},\mathcal{B}_{-r})dw,$$

where

$$\pi(\alpha^{(r)},w|\mathbb{Y},x;\beta^{(r)},\mathcal{B}_{-r}) = \frac{\{\prod_{k=1}^{n} P(Y_k = y_k|\mathcal{B})\}f(w|\mathcal{B})\pi(\alpha^{(r)}|\mathcal{B}_{-r},\beta^{(r)})}{c(\{\mathbb{Y},x\})} \tag{5.4}$$

with $c(\{\mathbb{Y},x\})$ being the normalizing constant.

In the appendix, we show that the augmented posterior distribution for the coefficients for $\alpha^{(r)}$ is

$$\pi(\alpha^{(r)}|\mathbb{Y},x,w;\beta^{(r)},\mathcal{B}_{-r}) \quad \propto \Big\{\prod_{k=1}^{n} P(Y_k = y_k|\mathcal{B})\Big\}f(w|\mathcal{B})\pi(\alpha^{(r)}|\beta^{(r)},\mathcal{B}_{-r}), \tag{5.5}$$

which is a multivariate normal distribution if the prior distribution for $\alpha^{(r)}$ is specified as a normal distribution (Choi and Hobert 2013), let $m_\alpha(w)$ and $\Sigma_\alpha(w)$ denote the mean and covariance matrix of the posterior normal distribution of $\alpha^{(r)}$. Thus, the Bayesian inference can proceed with sampling from $f(w|\mathcal{B})$, $\pi(\alpha^{(r)}|\mathbb{Y},x,w;\beta^{(r)},\mathcal{B}_{-r})$ and $\pi(\beta^{(r)}|\mathbb{Y},x,w;\alpha^{(r)},\mathcal{B}_{-r})$ iteratively.

5.2.2 Prior Specification

Guhaniyogi et al. (2017) discussed an adequate global-local shrinkage prior distribution for $\alpha^{(r)}$ and $\beta^{(r)}$, which typically suits high dimensional linear regression models. However, under the logistic regression, the horseshoe shrinkage prior performs better, suggested by Wei and Ghosal (2020). Thus, in our framework, we employ the horseshoe shrinkage priors for $\alpha^{(r)}$ and $\beta^{(r)}$ marginally to deal with the sparsity problem.

Assuming that the $\alpha_i^{(r)}$ and $\beta_j^{(r)}$ are conditionally independent, for $r = 1,...,R$, $i = 1,...,p$, $j = 1,...,q$, and $l = 1,...,p_z$, we specify the priors as:

$$
\begin{aligned}
&\pi(\alpha_i^{(r)}|\lambda_{\alpha_i^{(r)}}, a) \sim N(0, \lambda^2_{\alpha_i^{(r)}} a^2);\\
&\pi(\beta_j^{(r)}|\lambda_{\beta_j^{(r)}}, a) \sim N(0, \lambda^2_{\beta_j^{(r)}} a^2);\\
&\pi(\lambda_{\alpha_i^{(r)}}) \sim C^+(0,1);\\
&\pi(\lambda_{\beta_j^{(r)}}) \sim C^+(0,1);\\
&\pi(a) \sim C^+(0,1);
\end{aligned}
\tag{5.6}
$$

where $C^+(0,1)$ is the half-Cauchy distribution, and hyperparameter $a > 0$ controls the global shrinkage.

5.2.3 Computation of Posterior

The details of the posterior distribution of each parameter are included in Appendix D. Here we describe estimation procedures using the MCMC algorithm with Gibbs sampling, which consist of three blocks. At iteration $(t+1)$:

Block 1. Sample the hyperparameters λ_α, λ_β, λ_γ and a using slice sampling based on the algorithm from of Polson et al. (2014). Here we provide the steps for obtaining $\lambda_{\alpha^{(r)}}^{(t+1)} = (\lambda^{(t+1)}_{\alpha_1^{(r)}}, ..., \lambda^{(t+1)}_{\alpha_p^{(r)}})$, given a fixed rank r and $\lambda^{(t)}_{\alpha^{(r)}}$:

Step 1: *sample* $u_{\alpha_i^{(r)}}|\psi_{\alpha_i^{(r)}}$ *uniformly from interval* $\left(0, \frac{1}{1+\psi_{\alpha_i^{(r)}}}\right)$, *where* $\psi_{\alpha_i^{(r)}} = \frac{1}{(\lambda^{(t)}_{\alpha_i^{(r)}})^2}$,

Step 2: *sample* $(\psi_{\alpha_i^{(r)}}|u_{\alpha_i^{(r)}}, \alpha_i^{(r),(t)})$ *from the exponential density* $\mathrm{Exp}(\frac{2}{\alpha_{\mathrm{i}}^{\mathrm{(r),(t)}}})$, *truncated to have a zero probability outside the interval* $\left(0, \frac{1-u_{\alpha_i^{(r)}}}{u_{\alpha_i^{(r)}}}\right)$,

Step 3: *transform back to* $\lambda^{(t+1)}_{\alpha_i^{(r)}}$ *using* $\psi_{\alpha_i^{(r)}}$; $\lambda_\beta^{(t+1)}$ *and* $\lambda_\gamma^{(t+1)}$ *are generated using the same process as* $\lambda_{\alpha_i}^{(t+1)}$.

Block 2. Generate the random variables $W_1, ..., W_n$ independently using

$$W_k \sim PG(1, c_k)$$

and write the sampled value as $w^{(t)} = (w_1^{(t)}, ..., w_n^{(t)})$.

Block 3. Given $\mathcal{B}_{-r}^{(t)}$, sample the coefficients $\alpha^{(r)}$ and $\beta^{(r)}$ by three steps:

Step 1: *sample* $\alpha^{(r),(t+1)}$ *from* $N_p\{m_{\alpha^{(r)}}(w^{(t)}), \Sigma_{\alpha^{(r)}}(w^{(t)})\}$, *where*

$$\begin{aligned} m_{\alpha^{(r)}}(w^{(t)}) &= \Sigma_{\alpha^{(r)}}(w^{(t)}) x_{\beta^{(r)}}^{(t)} y(w^{(t)}), \\ \Sigma_{\alpha^{(r)}}(w^{(t)}) &= \left\{ x_{\beta^{(r)}}^{(t)\intercal} \Omega(w^{(t)}) x_{\beta^{(r)}}^{(t)} + \Sigma_{\alpha^{(r)}}^{(t+1),-1} \right\}^{-1}, \end{aligned}$$

$x_{\beta^{(r)}}^{(t)} = (x_1\beta^{(r),(t)}, ..., x_n\beta^{(r),(t)})^\intercal$, $y = (y_1, ..., y_n)^\intercal$, $y(w^{(t)}) = y - \frac{1}{2}\mathbf{1}_n - x_{\mathcal{B}_{-r}^{(t)}}(w^{(t)})$, $x_{\mathcal{B}_{-r}^{(t)}}(w^{(t)}) = \{(< x_1, \mathfrak{B}_{-r}^{(t)} >)w_1^{(t)}, ..., (< x_n, \mathfrak{B}_{-r}^{(t)} >)w_n^{(t)}\}^\intercal$, 1_n is an $n \times 1$ unit vector, $\Omega(w^{(t)}) = diag(w^{(t)})$ and $\Sigma_{\alpha^{(r)}}^{(t+1)} = diag\{(\lambda_{\alpha^{(r)}}^{(t+1)} a^{(t+1)})^2\}$;

Step 2: *sample* $\beta^{(r),(t+1)}$ *from* $N_q\{m_{\beta^{(r)}}(w^{(t)}), \Sigma_{\beta^{(r)}}(w^{(t)})\}$, *where*

$$\begin{aligned} m_{\beta^{(r)}}(w^{(t)}) &= \Sigma_{\beta^{(r)}}(w^{(t)}) x_{\alpha^{(r)}}^{(t+1)} y(w^{(t)}), \\ \Sigma_{\beta^{(r)}}(w^{(t)}) &= \left\{ x_{\alpha^{(r)}}^{(t+1),\intercal} \Omega(w^{(t)}) x_{\alpha^{(r)}}^{(t+1)} + \Sigma_{\beta^{(r)}}^{(t+1),-1} \right\}^{-1}, \end{aligned}$$

$x_{\alpha^{(r)}}^{(t+1)} = (x_1^\intercal\alpha^{(r),(t+1)}, ..., x_n^\intercal\alpha^{(r),(t+1)})^\intercal$, and $\Sigma_{\beta^{(r)}}^{(t+1))} = diag\{(\lambda_{\beta^{(r)}}^{(t+1)} a^{(t+1)})^2\}$;

The MCMC samples are generated by repeating the three blocks many times after discarding the early generated samples for a certain burn-in period.

5.3 Bayesian Estimation Procedure with Missclassification on Response

In applications, the true response Y_k for $k = 1, ..., n$ may be subject to misclassification, and a surrogate response, Y_k^*, is observed. Let O_k be an indicator variable for the kth subject such that $O_k = 1$ if $Y_k^* = Y_k$ and $O_k = 0$ otherwise. We denote $\mathbb{Y}^* = (Y_1^*, ..., Y_n^*)^\intercal$ and $\mathcal{O} = (O_1, ..., O_k)^\intercal$. Let $\rho = P(O_k = 1|Y_k = y_k)$ be the probability of observing Y_k correctly, which is assumed to known for now. The variables O_k, Y_k^*, and the true response, Y_k are connected via

$$Y_k = O_k \times Y_k^* + (1 - O_k) \times (1 - Y_k^*). \tag{5.7}$$

The conditional distribution of O_k is given by

$$P(O_k = 1|\mathcal{B}, Y_k^*, \rho) = \frac{1}{C_k}\rho \times p_k^{Y_k^*} \times (1 - p_k)^{1-Y_k^*} \tag{5.8}$$

where $C_k = \rho \times p_k^{Y_k^*} \times (1-p_k)^{1-Y_k^*} + (1-\rho) \times p_k^{1-Y_k^*} \times (1-p_k)^{Y_k^*}$ is the normalization constant. Then O_k can be sampled based on a binomial distribution with the probability (5.8).

To carry out inference about $\mathcal{B}$ using the surrogate measurements Y_k^*, we modify the procedure in Section 5.2.3, by bridging Y_k^* and Y_k. To be specific, this posterior density, denoted $P(\mathcal{B}|\rho, \mathbb{Y}^*, \mathcal{O})$, can be derived from (2.1), (5.7) and the Bayesian hierarchical model of Rekaya et al. (2001):

$$P(\mathcal{B}|\rho, \mathbb{Y}^*, \mathcal{O}) \propto \Pi_\tau(\mathcal{B}) \times \prod_{k=1}^{n} p_k^{(1-O_k)Y_k^*+O_k(1-Y_k^*)} \times (1-p_k)^{1-(1-O_k)Y_k^*-O_k(1-Y_k^*)}, \tag{5.9}$$

where p_k is given by (5.3), $\Pi_\tau(\mathcal{B}) = \prod_{r=1}^{R}\{\prod_{i=1}^{p} \pi_\tau(\alpha_i^r)\}\{\prod_{j=1}^{q} \pi_\tau(\beta_j^r)\}$ denotes the product of the prior distributions for the $\alpha^{(r)}$ and $\beta^{(r)}$, and τ represents the set of hyper-parameters that are suppressed in the notation $\pi_\tau(\alpha_i^r)$ and $\pi_\tau(\beta_j^r)$.

Then we modify the algorithm described in Section 5.2.3 by replacing its Block 2 with:

Block 2*. Given $\mathcal{B}^{(t)}$, x_k, Y_k^* and ρ,

Step 1: generate the random variables $W_1, ..., W_n$ independently using

$$W_k \sim PG(1, c_k)$$

and let $w = (w_1, ..., w_n)$ denote the sampled values.

Step 2: generate O_k from the Bernoulli distribution with the probability (5.8), and then recover Y_k based on (5.7). Let Y_k^s denote the resulting value which is used for the implementation of Block 3 in Section 5.2.3, where Y_k is replaced by Y_k^s.

5.4 Simulation Studies

In this subsection, various simulations are designed to evaluate the performance of the proposed methods, together with the impacts of different degrees of misclassification on parameter estimation. We consider settings with $p = q$, and denote this to be p_x for ease of exposition. We consider the case with the sample size $n = 1000$ or $n = 2000$. Matrix-variate data, x_k, from the matrix-normal distribution $MN(0, I_{p_x}, I_{p_x})$, for $k = 1, ..., n$, where p_x is taken as 5 or 20.

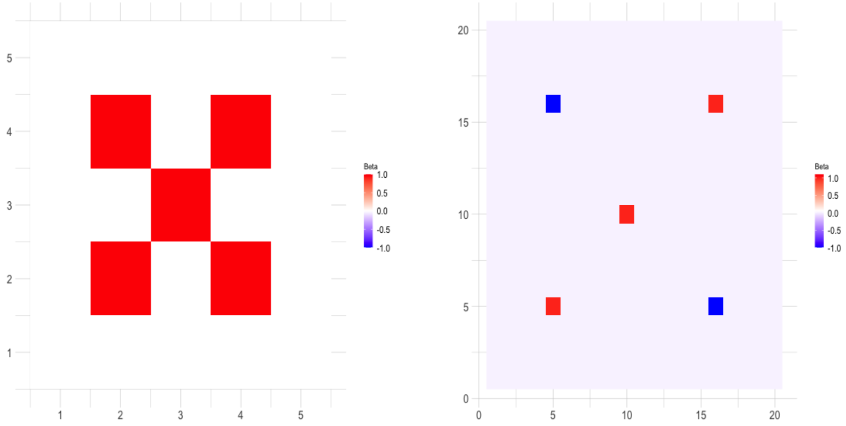


Figure 5.1: Designed $\mathcal{B}$ with $p_x = 5$ (left) and 10 (right). The vertical bar labeled with Beta represents the corresponding values of different color in the figure.

For the parameter $\mathcal{B}$, we design it as a rank-1 or rank-2 matrix. In details, let $\mathcal{B}_{i,j}$ denote the ith row and the jth column element of $\mathcal{B}$. For $p_x = 5$ cases, we set $\mathcal{B}$ to be a rank-2 matrix-variate where $\mathcal{B}_{2,2}, \mathcal{B}_{2,4}, \mathcal{B}_{3,3}, \mathcal{B}_{4,2}, \mathcal{B}_{4,4}$ to be 1 and other entries of $\mathcal{B}$ to be 0; for $p_x = 20$ cases, we consider two different ranks: (1) for the case with a rank-1 matrix-variate, we set $\mathcal{B}_{5,5}, \mathcal{B}_{3,3}, \mathcal{B}_{16,16}$ to be 1, $\mathcal{B}_{5,16}, \mathcal{B}_{15,5}$ to be -1, and other entries of $\mathcal{B}$ to be 0; (2) for the case with a rank-2 matrix-variate, we set $\mathcal{B}_{5,5}, \mathcal{B}_{10,10}, \mathcal{B}_{3,3}, \mathcal{B}_{16,16}$ to be 1, $\mathcal{B}_{5,16}, \mathcal{B}_{15,5}$ to be -1, and the rest entries of $\mathcal{B}$ to be 0. Figure 5.1 displays the designed $\mathcal{B}$ of the cases with $p_x = 5$ and 10, where blue and red squares show negative and positive values in the range $[-1, 1]$, respectively. For $k = 1, ..., n$, the binary response Y_k is independently generated from the Bernoulli distribution with the probability (5.3).

We evaluate the accuracy of the estimates in the terms of the L_2-error:

$$\|\mathcal{B} - \hat{\mathcal{B}}\| = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q}(\mathcal{B}_{i,j} - \hat{\mathcal{B}}_{i,j})^2} \quad ,$$

where $\hat{\mathcal{B}}$ represents the estimated posterior mean of $\mathcal{B}$. We also monitor the coverage of 95% and 90% credible intervals. We compare the variable selection performance using the average proportion of including zero effects (FP) and the average proportion of excluding non-zero effects (FN), where the covariate is excluded if zero is covered by its 95% credible interval.

5.4.1 Simulation Studies without Misclassification

In this subsection, we evaluate the performance of the procedure proposed in Section 5.2.3 where the precisely measured outcome Y_k is available, and call this method as "Method 1". In the implementation of Method 1, when fitting the data, the rank of $\mathcal{B}$ is set as that for generating the data. We apply the LASSO method as a reference method and call it "LASSO". After the burn-in period consisting of 1,000 Markov chain Monte Carlo (MCMC) iterations, we generate 5,000 posterior samples and save the samples of model parameters at every 5 iterations to reduce autocorrelation between the samples. Table 1 records the results for the cases with $p_x = 5$. Method 1 outperforms the LASSO method.

To also investigate the effect of a potentially misspecified rank, when fitting the simulated data with the case $p_x = 20$, we consider two methods by setting the rank of $\mathcal{B}$ to be 1 or 2, and call the resulting method "True R_T-Fit R_F", where R_T represents the true rank of coefficients of $\mathcal{B}$, and R_F represents the user specified rank when fitting the data. The results are presented in Tables 3-4, showing that Method 1 provides better performance than the LASSO method in both correctly specified or misspecified rank situations. Comparing Tables 3 and 4, as we expected, True 1-Fit 1 and True 2-Fit 2 yield better results than those obtained in the presence of rank misspecification. With rank misspecification involved, True 1-Fit 2 outperforms True 2-Fit 1, suggesting that engaging a lower rank to estimate a higher rank of $\mathcal{B}$ has worse performance than the opposite way.

To reduce the risk of rank misspecification, we recommend to use the LASSO method as a start to first decide a suitable rank of the matrix-variate and then apply our proposed methods. A.1 and A.2 in Figures 5.2-5.6 give the estimated posterior means of model parameters using Method 1 and the LASSO, showing the same patterns as we observed from Tables 5.1-5.3.

5.4.2 Simulation Studies with Misclassification

In this subsection, we evaluate the performance of the procedure proposed in Section 5.3 where only the surrogate response, Y_k^*, of Y_k is available, and we call this "Method 2". We consider three misclassification situations with $\rho = 0.95$, 0.90 or 0.85, to reflect an increasing degree of misclassification in Y_k, where ρ is defined in Section 5.3. We are interested in not only the misclassification effects, but also the effects of potential rank misspecification. Thus, when fitting the simulated data with the case $p_x = 20$, we consider "True R_T-Fit R_F", where $R_T = 1, 2$ and $R_F = 1, 2$ as well. The LASSO method is also applied as a reference method.

The simulation results for the case $p_x = 5$ with different ρ are summarized in Table 5.2. For comparison, we also present in this table the results obtained from naively applying the LASSO method and Method 1 by ignoring the response misclassification, denoted "LASSO-naive" and "Method-1-naive", respectively. We observe that the two naive methods provide biased results, and the LASSO-naive method performs worse than Method-1-naive does. Method 2 yields reasonable results. Tables 3-4 report the simulation results for the case with $p_x = 20$, showing similar patterns observed for the case with $p_x = 5$. For the influence of rank misspecification of $\mathcal{B}$, we find that Method 2 shows similar patterns as observed in Section 5.4.1. Misclassification effects do not seem dramatic in shrinking unimportant coefficients or retaining parameters.

In Figures 5.2-5.6, we report the estimated posterior means of $\mathcal{B}$ using the LASSO-naive, Method-1-naive, and Method 2 with different values for ρ. These figures show similar results to those of Tables 5.2-5.4. Figures 2-4 are obtained $\hat{\mathcal{B}}$ with the correctly specified rank of $\mathcal{B}$, and they show that Method 2 (in Column 3) provides the most precise $\hat{\mathcal{B}}$ under different ρ settings (in Rows B-D). Two naive methods (in Columns 1-2) display biased estimates with lighter red or blue squares. All the methods correctly select the non-zero $\mathcal{B}_{i,j}$ when the rank of $\mathcal{B}$ is correctly specified. Figures 5.5-5.6 summarize the results when the rank of $\mathcal{B}$ is misspecified for $p_x = 20$. Method 2 still provides the less biased $\hat{\mathcal{B}}$ compared to the two naive methods. Especially, in Figure 5.6, although Method 2 cannot find $\mathcal{B}_{10,10}$ like the LASSO-naive method, it provides the most precise estimates of the selected $\mathcal{B}_{i,j}$. That is the reason we suggest using the LASSO method to investigate the rank of $\mathcal{B}$ first when $\mathcal{B}$ has a large dimension.

5.5 Data Analysis of LSVT Data

In this subsection, we apply the proposed method, in contrast to the LASSO approach, to analyze a subset of the Lee Silverman voice treatment (LSVT) Companion data, available at the UCI Machine Learning Repository website: https://archive.ics.uci.edu/ml/datasets/LSVT+Voice+Rehabilitation. The study investigates the potential of using sustained vowel phonations for Parkinson's diseased (PD) patients whose voice is classified as "acceptable" or "unacceptable" during an in-person rehabilitation treatment. Each subject is originally instructed to produce phonations categorized into one of the nine combinations of pitch and amplitude, where the pitch is defined as "comfortable", "high" and "low"; and the amplitude is considered to be "acceptable", "too loud", or "too soft". The data set contains a total of 126 phonations for the 14 subjects, where each subject contributes nine phonations. The details of the study can be found in Tsanas et al. (2013).

The study processes the speech signals and extracts the features of 126 phonations into two groups of features which are suitably displayed in matrix form. Specifically, in the analysis of Tsanas et al. (2013), the first group of features is formed wavelet measures where 17 wavelet coefficients are calculated for a 10 level wavelet decomposition of the fundamental frequency time series (F0), resulting in a vector-covariate with 170 attributes. The second group of the features uses the dysphonia measures, jitter and shimmer, resulting in three features quantifying F0, the pitch deviations, and the amplitude deviations. Given that 13 characteristics of the three features are calculated, Tsanas et al. (2013) created a vector-covariate of 39 attributes for their analysis. However, such an approach of reporting data obscures the inherent relations among the features and characteristics. It is more reasonable to display the measurements in the first group as a *matrix*-variate with 10 levels and 17 features treated as rows and columns, respectively, and the measurements in the second group as a *matrix*-variate with 13 characteristics and 3 features treated as rows and columns, respectively.

A phonation is assessed by the LSVT clinicians to be acceptable (setting $Y = 1$) or unacceptable (setting $Y = 0$), where the assessment largely depends on the experience of a rater. Thus, there is a possibility that phonations may be misclassified due to no solid criteria can be applied for the assessment. For $k = 1, .., 126$, let Y_k^* denote the observed value for the true binary variable for phonation k, with value 1 for being in the acceptable group and 0 otherwise. The matrix-variate of phonation k, for the features in group 1, denoted as $x_k^{(1)}$, is a 10×17 matrix with entry (i, j) representing the value of the jth wavelet coefficient of the ith level, where $i = 1, ..., 10$ and $j = 1, .., 17$; for the features in group 2, denoted as $x_k^{(2)}$, is a 13×3 matrix with entry (i, j) representing the value of the jth feature of the ith characteristic, where $i = 1, ..., 13$ and $j = 1, .., 3$. Correspondingly, we let $\mathcal{B}^{(1)}$ and $\mathcal{B}^{(2)}$ denote the parameters for $x_k^{(1)}$ and $x_k^{(2)}$, respectively. Consistent with the notation in Section 5.3, we let ρ denote the probability of assessing a phonation correctly.

While it is interesting to understand the possible impacts of misclassification on the analysis, there is no information on the degree of misclassification in this data set. Consequently, we conduct sensitivity analyses by specifying different magnitudes of misclassification probabilities. In particular, we take $\rho = 0.95$, 0.90, or 0.85 to feature increasing degrees of misclassification.

Figure 5.7 shows the point estimators of $\mathcal{B}^{(1)}$ and $\mathcal{B}^{(2)}$ using the LASSO method. Figures 5.8 and 5.9 show the lower and upper bounds of 95% Credible Interval (CI) and estimated posterior means for $\mathcal{B}^{(1)}$ and $\mathcal{B}^{(2)}$ by applying Methods 1 and 2 described in Section 5. For $x^{(1)}$, there is no significant $\mathcal{B}_{i,j}^{(1)}$ selected for all the proposed methods, but the LASSO method selects two variables. For $x^{(2)}$, $\mathcal{B}_{3,1}^{(2)}$, $\mathcal{B}_{4,1}^{(2)}$, $\mathcal{B}_{4,2}^{(2)}$, $\mathcal{B}_{12,2}^{(2)}$, $\mathcal{B}_{5,3}^{(2)}$, $\mathcal{B}_{11,3}^{(2)}$ $\mathcal{B}_{12,3}^{(2)}$ are selected as significant parameters by Method 1. However, no characteristic of feature F0 is selected, three characteristics of the pitch deviation and three more characteristics of the amplitude deviation are selected in the model by the LASSO method. As the misclassification rate increases, the magnitude of the posterior means of the model parameters increases, and the 95% CIs become wider.

Table 5.1: Model performance for $p_x = 5$ without misclassifiction

Model	LASSO	Method 1
L_2 error	2.368(0.168)	0.361(0.078)
95% Coverage		93.6%
90% Coverage		88.6%
FP	0.476(0.065)	0.040(0.054)
FN	0(0)	0(0)

Numbers in $(\cdot)$ represents standard error.

Table 5.2: Model performance for $p_x = 5$ with misclassifiction

Model	$\rho = 0.95$			$\rho = 0.90$			$\rho = 0.85$		
	LASSO-naive	Method-1-naive	Method 2	LASSO-naive	Method-1-naive	Method 2	LASSO-naive	Method-1-naive	Method 2
L_2 error	2.190(0.169)	0.585(0.099)	0.368(0.068)	2.109(0.209)	0.912(0.099)	0.543(0.137)	2.08	1.178(0.091)	0.679(0.195)
95% Coverage		25.4%	96.2%		0.4%	93.8%		0.0%	92.6%
90% Coverage		16.7%	92.3%		0.1%	87.9%		0.0%	87%
FP	0.423(0.175)	0.040(0.053)	0.027(0.042)	0.399(0.170)	0.044(0.058)	0.050(0.060)	0.375(0.164)	0.043(0.052)	0.052(0.057)
FN	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)

Numbers in $(\cdot)$ represents standard error.

Table 5.3: Model performance for correctly specified rank when $p_x = 20$

	True 1-Fit 1			True 2-Fit 2		
	Without Misclassification					
	LASSO	Method 1	Method 2	LASSO	Method 1	Method 2
L_2 error	2.332(1.022)	0.378(0.050)		2.630(0.999)	0.565(0.053)	
95% Coverage		94.2			91.6	
90% Coverage		89.2			86.0	
FP	0.059(0.030)	0.011(0.009)		0.078(0.036)	0.017(0.009)	
FN	0(0)	0(0)		0(0)	0(0)	
	With Misclassification					
	LASSO-naive	Method-1-naive	Method 2	LASSO-naive	Method-1-naive	Method 2
	$\rho = 0.95$					
L_2 error	2.238(0.732)	0.538(0.0668)	0.461(0.066)	2.502(0.679)	0.711(0.067)	0.732(0.099)
95% Coverage		6.3	92.9		6.6	83.5
90% Coverage		3.6	86.5		3.5	74.1
FP	0.050(0.029)	0.011(0.009)	0.012(0.009)	0.064(0.032)	0.016(0.009)	0.018(0.009)
FN	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
	$\rho = 0.90$					
L_2 error	2.169(0.520)	0.801(0.070)	0.565(0.100)	2.415(0.462)	0.999(0.070)	1.008(0.197)
95% Coverage		0	90		0	67.4
90% Coverage		0	83		0	56.5
FP	0.047(0.029)	0.012(0.100)	0.012(0.011)	0.057(0.030)	0.017(0.010)	0.020(0.011)
FN	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
	$\rho = 0.85$					
L_2 error	2.116(0.356)	1.034(0.066)	0.717(0.148)	2.353(0.304)	1.259(0.066)	1.858(0.790)
95% Coverage		0	87		0	41.3
90% Coverage		0	80.6		0	32
FP	0.043(0.028)	0.012(0.010)	0.015(0.012)	0.050(0.029)	0.017(0.010)	0.027(0.018)
FN	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)

Numbers in $(\cdot)$ represents standard error.

Table 5.4: Model performance for mis-specified rank when $p_x = 20$

	True 1-Fit 2			True 2-Fit 1		
	Without Misclassification					
	LASSO	Method 1	Method 2	LASSO	Method 1	Method 2
L_2 error	2.333(1.024)	0.497(0.068)		2.630(0.999)	1.101(0.023)	
95% Coverage		87.2			24.2	
90% Coverage		79			16	
FP	0.060(0.031)	0.012(0.01)		0.078(0.036)	0.0125(0.010)	
FN	0(0)	0(0)		0(0)	0(0)	
	With Misclassification					
	LASSO-naive	Method-1-naive	Method 2	LASSO	Method-1-naive	Method 2
	$\rho = 0.95$					
L_2 error	2.238(0.732)	0.564(0.063)	0.666(0.106)	2.502(0.679)	1.222(0.0395)	1.118(0.030)
95% Coverage		19.2	70.8		0	43.1
90% Coverage		11.9	58.8		0	34.1
FP	0.050(0.029)	0.013(0.010)	0.0146(0.012)	0.064(0.032)	0.013(0.010)	0.013(0.010)
FN	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
	$\rho = 0.90$					
L_2 error	2.169(0.520)	0.803(0.071)	0.956(0.197)	2.415(0.462)	1.367(0.045)	1.145(0.035)
95% Coverage		0	47.4		0	57.8
90% Coverage		0	35.3		0	48.8
FP	0.047(0.029)	0.013(0.011)	0.018(0.015)	0.057(0.030)	0.012(0.010)	0.013(0.010)
FN	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
	$\rho = 0.85$					
L_2 error	2.116(0.356)	1.032(0.065)	1.925(0.875)	2.353(0.304)	1.512(0.048)	1.191(0.044)
95% Coverage		0	19.4		0	67.2
90% Coverage		0	12.1		0	61.7
FP	0.043(0.028)	0.012(0.010)	0.029(0.024)	0.050(0.029)	0.013(0.010)	0.014(0.011)
FN	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)

Numbers in $(\cdot)$ represents standard error.

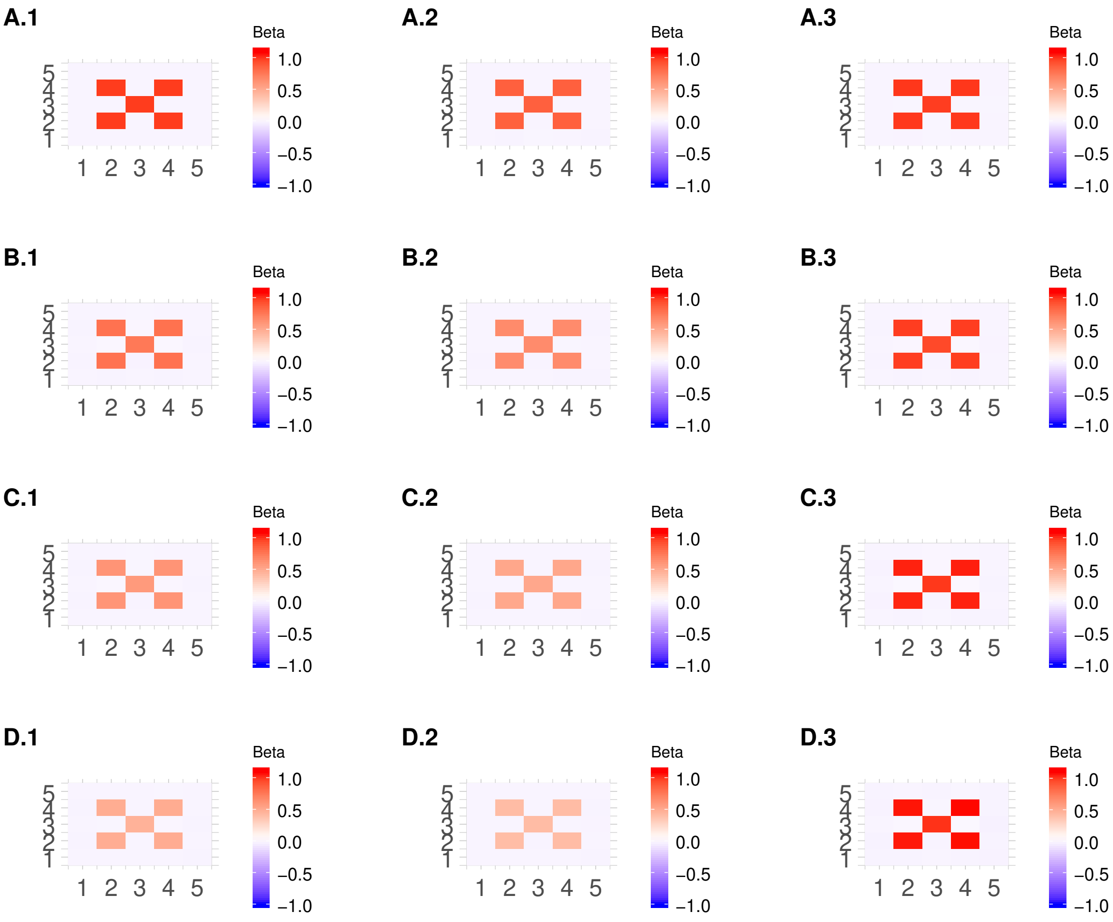


Figure 5.2: The estimated posterior means of model parameters for $p_x = 5$: Columns 1-3 record the results obtained from Method-1-naive, the LASSO-naive method and Method 2, respectively. Row A displays the results for the case with no misclassification, and Rows B, C and D summarize the results for $\rho = 0.95$, 0.90 and 0.85, respectively.

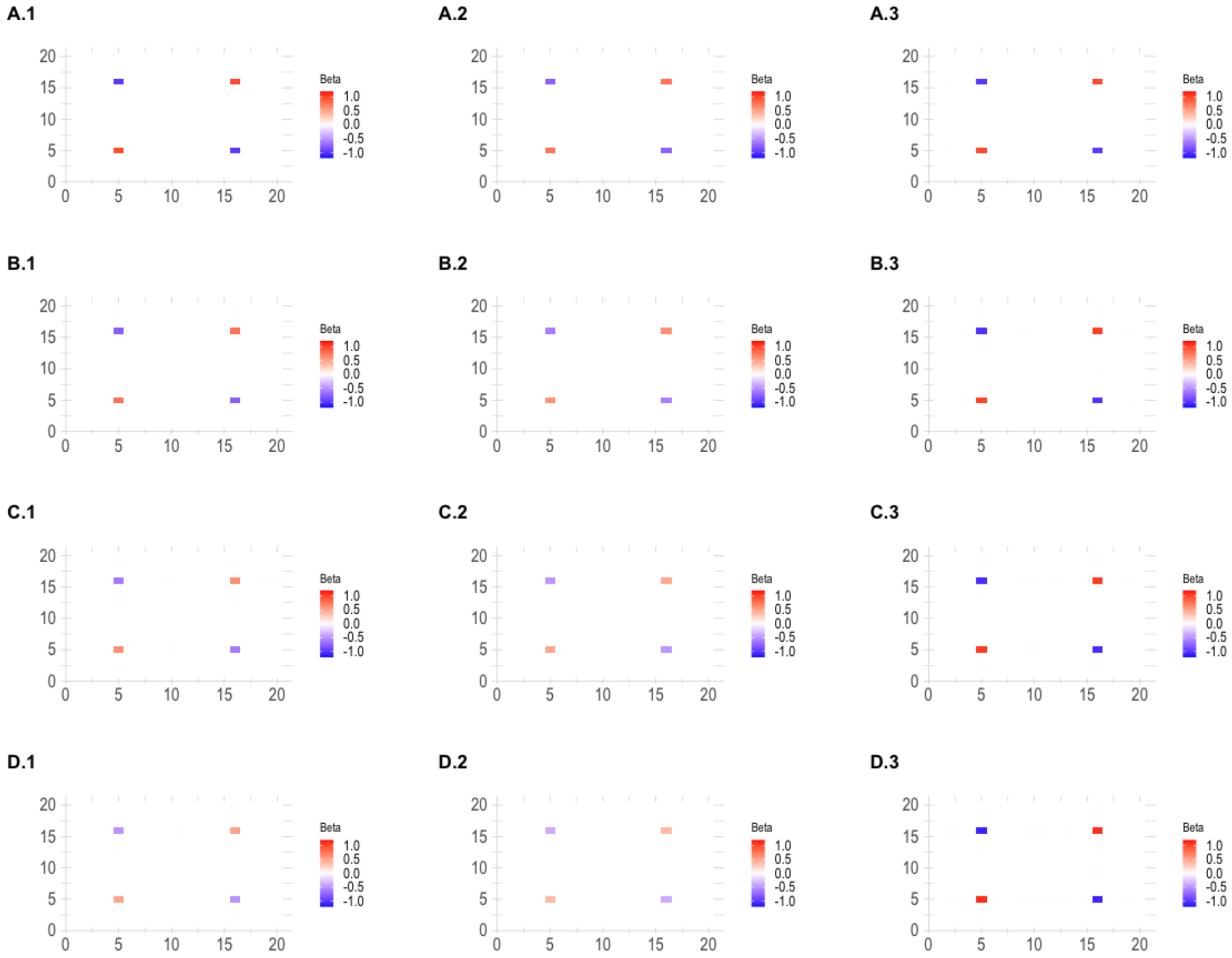


Figure 5.3: The estimated posterior means of model parameters for $p_x = 20$ using True 1-Fit 1: Columns 1-3 record the results obtained from Method-1-naive, the LASSO-naive method and Method 2, respectively. Row A displays the results for the case with no misclassification, and Rows B, C and D summarize the results for $\rho = 0.95$, 0.90 and 0.85, respectively.

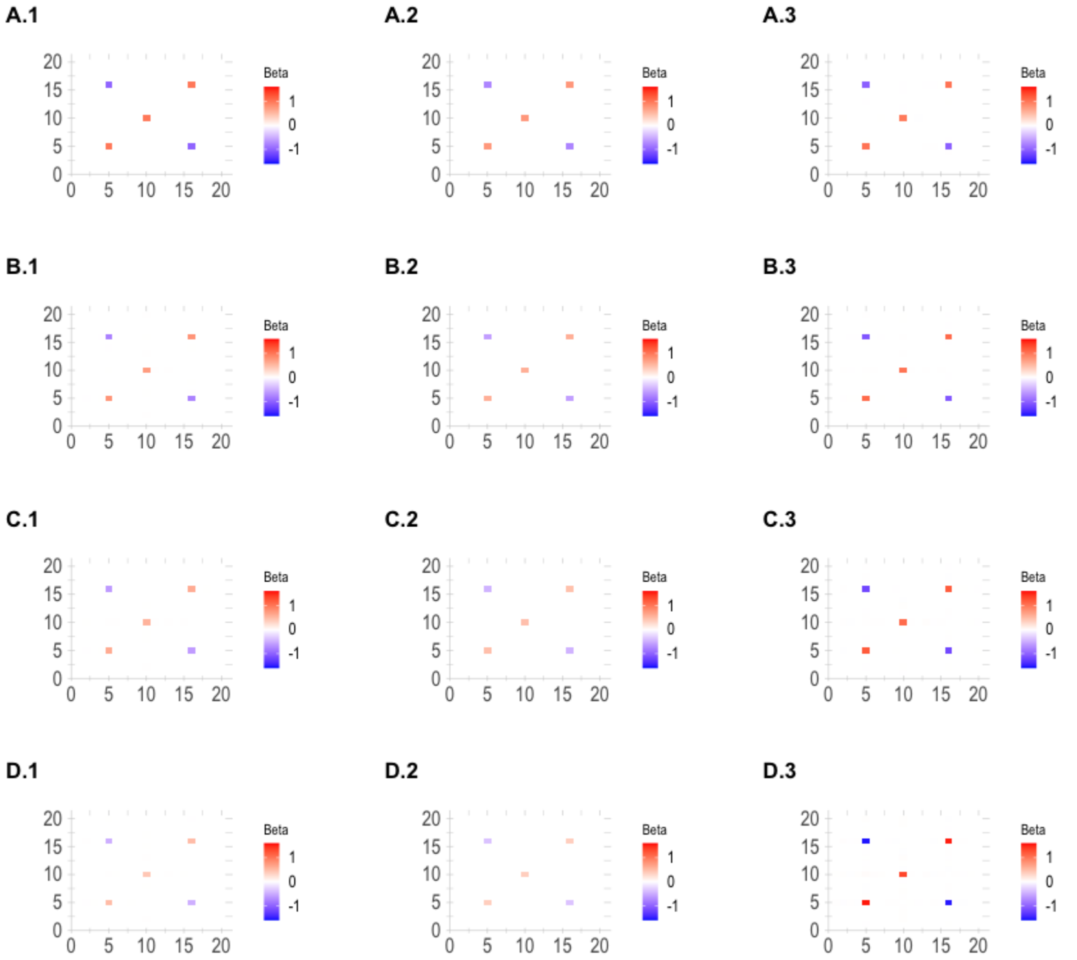


Figure 5.4: The estimated posterior means of model parameters for $p_x = 20$ using True 2-Fit 2:Columns 1-3 record the results obtained from Method-1-naive, the LASSO-naive method and Method 2, respectively. Row A displays the results for the case with no misclassification, and Rows B, C and D summarize the results for $\rho = 0.95$, 0.90 and 0.85, respectively.

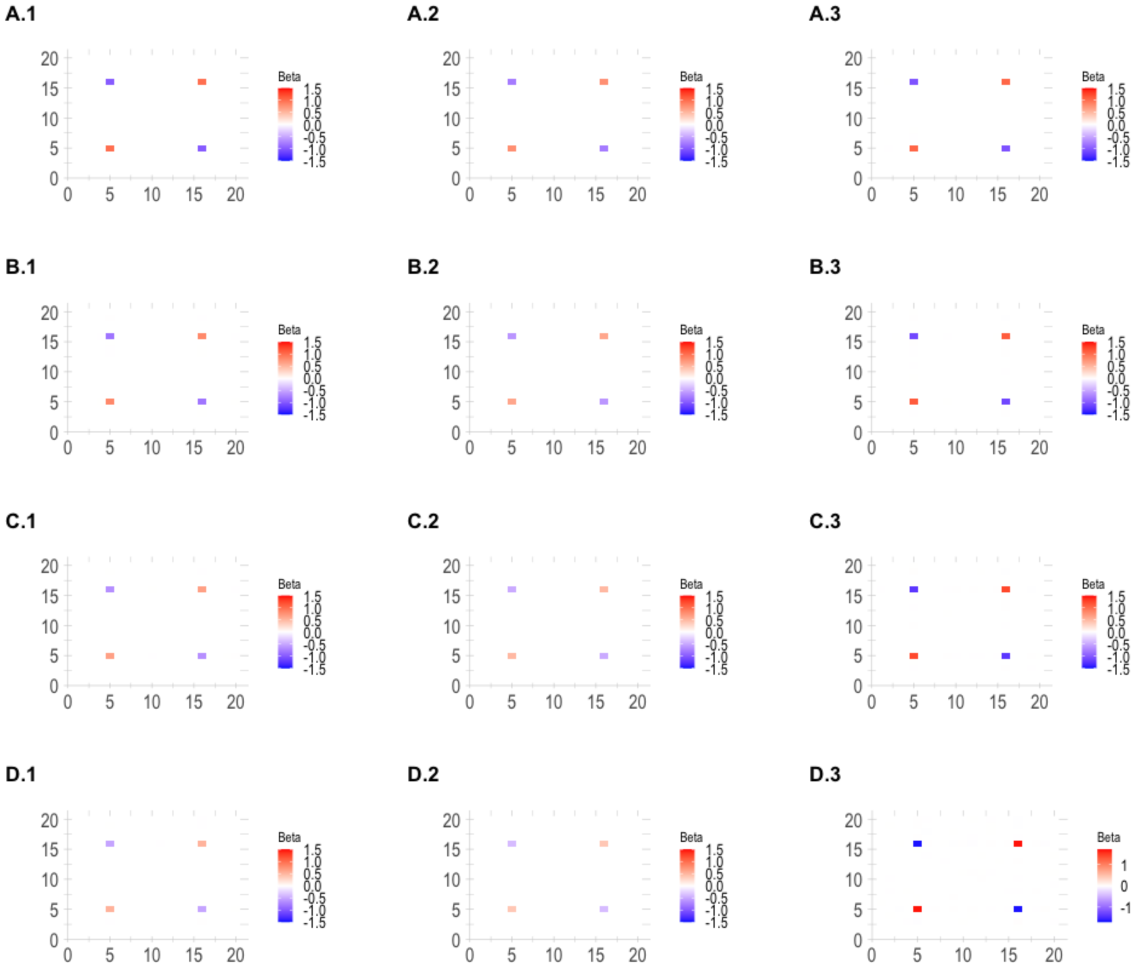


Figure 5.5: The estimated posterior means of model parameters for $p_x = 20$ using True 1-Fit 2:Columns 1-3 record the results obtained from Method-1-naive, the LASSO-naive method and Method 2, respectively. Row A displays the results for the case with no misclassification, and Rows B, C and D summarize the results for $\rho = 0.95$, 0.90 and 0.85, respectively.

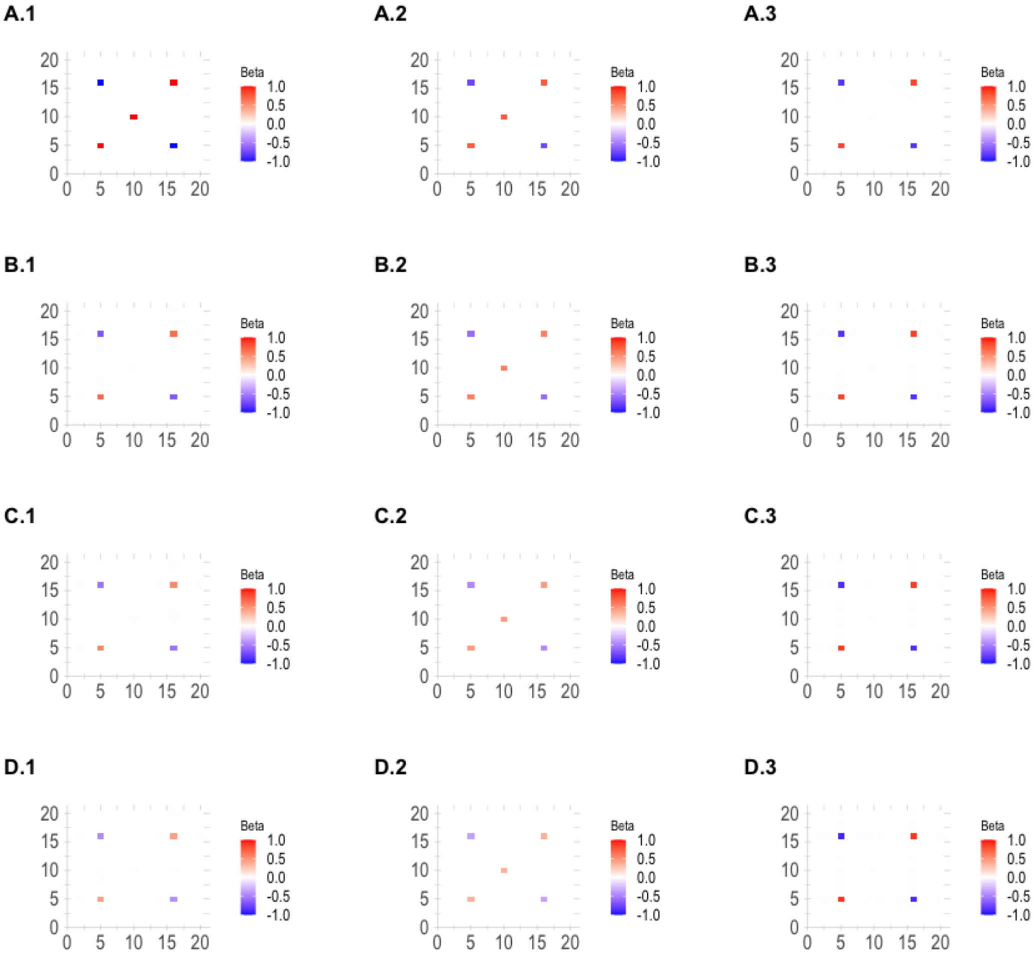


Figure 5.6: The estimated posterior means of model parameters for $p_x = 20$ using True 2-Fit 1: Columns 1-3 record the results obtained from Method-1-naive, the LASSO-naive method and Method 2, respectively. Row A displays the results for the case with no misclassification, and Rows B, C and D summarize the results for $\rho = 0.95$, 0.90 and 0.85, respectively.

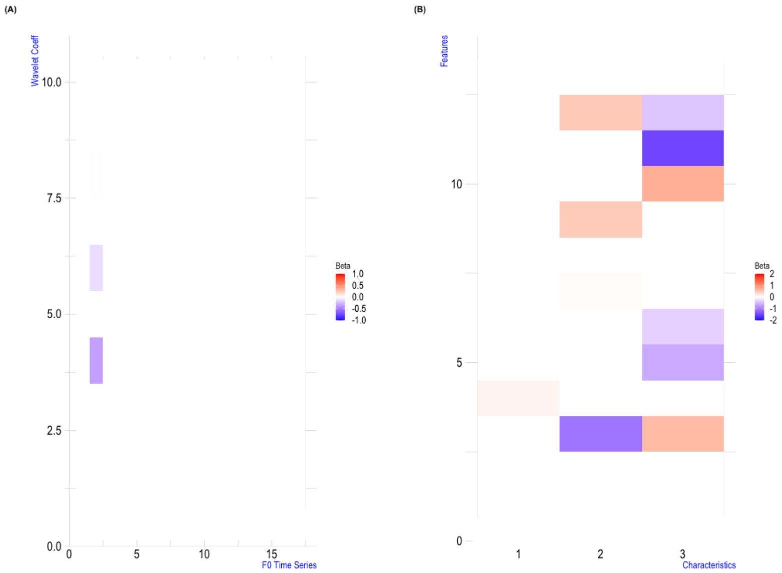
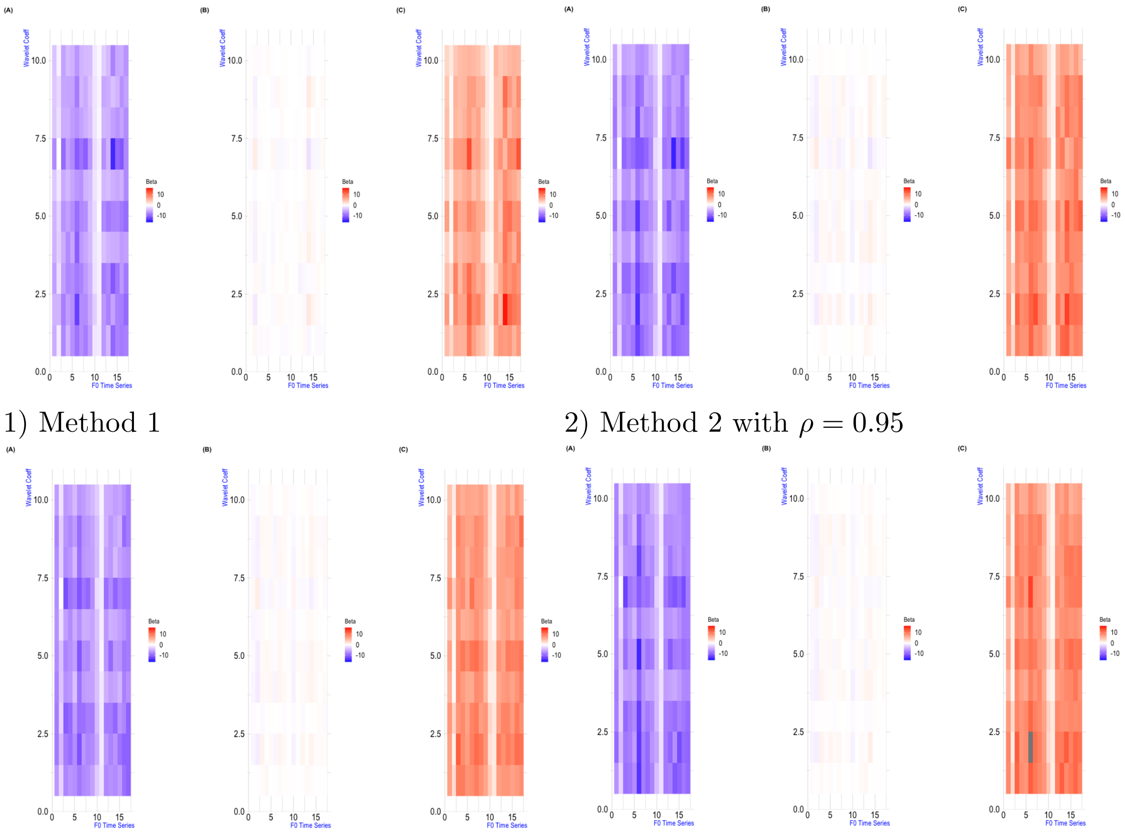


Figure 5.7: The point estimates of model parameters for $\mathcal{B}^{(1)}$ in (A) and $\mathcal{B}^{(2)}$ in (B) using the LASSO method.



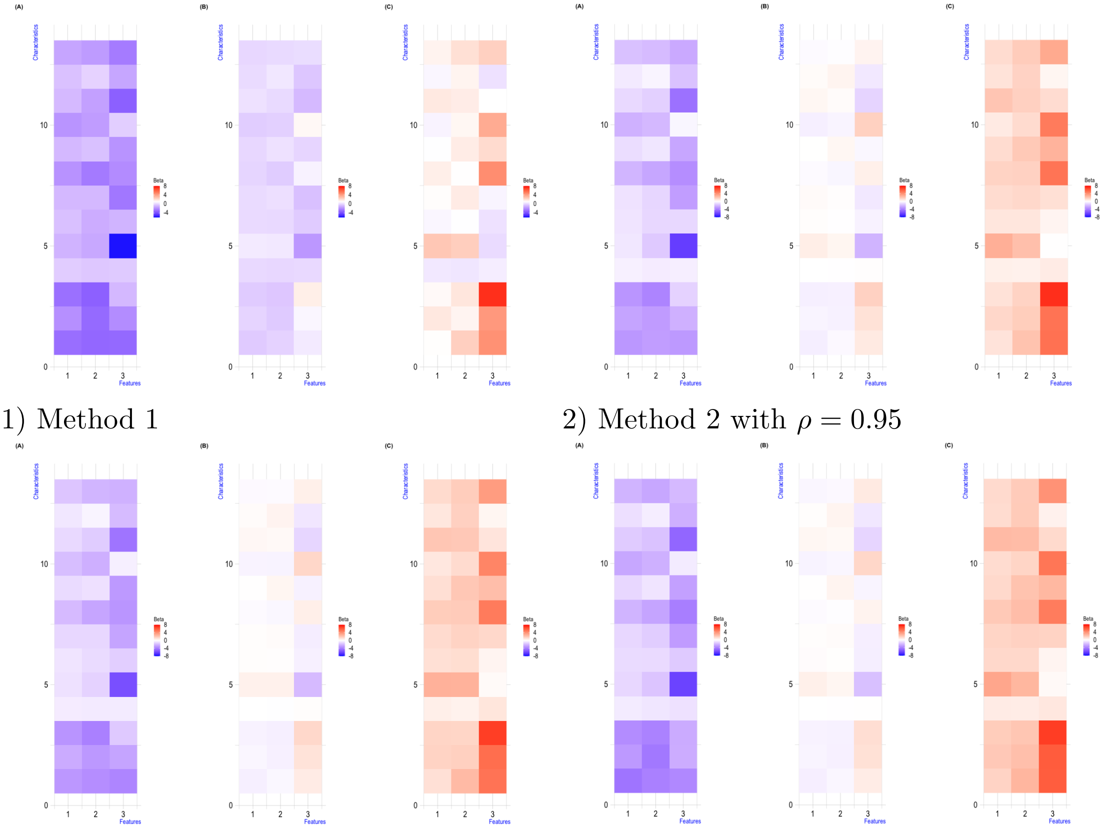
1) Method 1

2) Method 2 with $\rho = 0.95$

3) Method 2 with $\rho = 0.90$

4) Method 2 with $\rho = 0.85$

Figure 5.8: Estimation results for $\mathcal{B}^{(1)}$: in each sub-figure, (A) shows the 2.5% posterior quantiles of $\mathcal{B}^{(1)}$, (B) shows the estimated posterior means of $\mathcal{B}^{(1)}$, and (C) shows the 97.5% posterior quantiles of $\mathcal{B}^{(1)}$.



1) Method 1

2) Method 2 with $\rho = 0.95$

3) Method 2 with $\rho = 0.90$

4) Method 2 with $\rho = 0.85$

Figure 5.9: Estimation results for $\mathcal{B}^{(2)}$: in each sub-figure, (A) shows the 2.5% posterior quantiles of $\mathcal{B}^{(2)}$, (B) shows the estimated posterior means of $\mathcal{B}^{(2)}$, and (C) shows the 97.5% posterior quantiles of $\mathcal{B}^{(2)}$.

Chapter 6

Summary and Future Work

In this chapter, we present a summary for the previous chapters, together with discussions on possible future work or extensions.

Chapter 2:

Matrix-variate logistic regression models are useful in handling complex-structured covariates which commonly arise from imaging data. However, these models cannot be directly used when the number of model parameters is larger then the sample size. Furthermore, little discussion is available for using such models to analyze error-contaminated matrix-variate data. It is even unclear the impact would be if measurement error effects were ignored in such a setting. In Chapter 2, we study this important problem and develop two valid inference methods for accommodating measurement error effects in matrix-variate logistic regression. These two methods are developed under different distributional assumptions of the measurement error model; one makes a normality assumption for the measurement error while the other makes no assumptions. We establish theoretical results for the proposed methods and numerical studies demonstrate satisfactory finite sample performance.

In Chapter 2, we apply the $(2D)^2$PCA method (Zhang and Zhou 2005) to solve the inestimable problem where the sample size is smaller than the number of parameters; this method seems easier to implement than other matrix dimension reduction methods, such as generalized low rank approximations of matrices (GLRAM) (Ye 2005) and 2DPCA (Yang et al. 2004). Other dimension reduction methods can be considered as well prior to using the methods developed in Sections 2.3.1 and 2.3.2. For instance, one may consider to add a penalty function to the likelihood (2.3) in combination with the adjustment for the measurement error effects, and then employ the penalized likelihood function to reduce the dimension of the covariates. It would be interesting to explore this method in depth with some technical details here modified accordingly.

Chapter 3:

Vector-Matrix-variate logistic regression models are useful in characterizing the relationship between binary responses and matrix-expressed covariates as well as vector-expressed covariates. However, inference based on such models is challenged by the presence of response misclassification. In Chapter 3, we propose two valid methods, the imputation and likelihood methods to accommodate response misclassification effects in matrix-variate logistic regression. These two methods are developed for two settings where the misclassification rates are known or estimated from validation data. We establish theoretical results for the proposed methods and conduct numerical studies which demonstrate satisfactory finite sample performance of the methods.

In Chapter 3, we impose a constraint on the *row* effects to deal with the model identifiability problem, which allows us to focus on explaining the *row* and *column* effects separately. It is interesting to consider other types of constraint to express the effects of different combinations of *row* and *column* directly.

Chapter 4:

Regularized Matrix-variate logistic regression models are useful in characterizing the relationship between binary responses and matrix-expressed covariates, which commonly have sparsity property, as well as vector-expressed covariates. However, inference based on such models is challenged by the presence of response misclassification. In Chapter 4, we propose two valid methods, the imputation and likelihood methods to accommodate response misclassification effects in matrix-variate logistic regression combined with the SCAD penalty function. These two methods are developed for two settings where the misclassification rates are either known or estimated from validation data. We establish theoretical results for the proposed methods and conduct numerical studies which demonstrate satisfactory finite sample performance of the methods.

In Chapter 4, we still impose a constraint on the *row* effects to deal with the model identifiability problem. It is interesting to applying different inference methods to express the effects of different combinations of *row* and *column* directly. This is an ongoing project we are working with based on the Bayesian analysis. Moreover, we also add the SCAD penalty in the estimation procedure to deal with the sparsity property. Other penalty functions can be added as well to deal with this property.

Chapter 5:

Matrix-variate logistic regression models are newly emerging tools that are useful in featuring the relationship between binary responses and covariates in a matrix form as we shown in previous chapters. However, it is challenging due to the computational burden and intrinsic complex data structures under frequentist frame. In Chapter 5, we propose a Bayesian estimation procedure to analyze data that are facilitated by matrix-variate logistic regression. Furthermore, a modified Bayesian estimation procedure is proposed to deal with data with response misclassification. Numerical studies demonstrate satisfactory finite sample performance of the proposed methods.

The development in Chapter 5 focuses on the implementation procedures coupled with numerical studies. It is useful to develop rigorous theoretical results for the methods, which is a future project. Another interesting problem is to explore Bayesian methods to handle measurement error existing in matrix covariates; a similar problem is investigated by Fang and Yi (2020b) who focused on the frequentist framework.

References

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 669–679.

Albert, P. S., S. A. Hunsberger, and F. M. Biro (1997). Modeling repeated measures with monotonic ordinal responses and misclassification, with applications to studying maturation. *Journal of the American Statistical Association 92*, 1304–1311.

Atchadé, Y. F. (2017). On the contraction properties of some high-dimensional quasi-posterior distributions. *The Annals of Statistics 45*, 2248–2273.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association 110*, 1479–1490.

Biane, P., J. Pitman, and M. Yor (2001). Probability laws related to the Jacobi theta and Riemann zeta functions, and Brownian excursions. *Bulletin of the American Mathematical Society 38*, 435–465.

Buzas, J. S. and L. A. Stefanski (1996). Instrumental variable estimation in generalized linear measurement error models. *Journal of the American Statistical Association 91*, 999–1006.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective.* Chapman and Hall/CRC.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*, 465–480.

Carvalho, C. M. and M. West (2007). Dynamic matrix-variate graphical models. *Bayesian analysis 2*, 69–97.

Chen, Z., G. Y. Yi, and C. Wu (2011). Marginal methods for correlated binary data with misclassified responses. *Biometrika 98*, 647–662.

Chen, Z., G. Y. Yi, and C. Wu (2014). Marginal analysis of longitudinal ordinal data with misclassification in both response and covariates. *Biometrical Journal 56*, 69–85.

Choi, H. M. and J. P. Hobert (2013). The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics 7*, 2054–2064.

Cook, J. R. and L. A. Stefanski (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association 89*, 1314–1328.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B 20*, 215–242.

Dellaportas, P. and D. A. Stephens (1993). Bayesian analysis of errors-in-variables regression models. *Biometrics 51*, 1085–1095.

Dutilleu, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation 64*, 105–123.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fang, J. and G. Y. Yi (2020a). Imputation and likelihood methods for matrix-variate logistic regression with response misclassification. *Revised for The Canadian Journal of Statistics*.

Fang, J. and G. Y. Yi (2020b). Matrix-variate logistic regression with measurement error. *To be appear in Biometrika*.

Fang, J. and G. Y. Yi (2020c). Regularized matrix-variate logistic regression with response misclassification. *Submitted for Publication*.

Freedman, L. S., V. Fainberg, V. Kipnis, D. Midthune, and R. J. Carroll (2004). A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics 60*, 172–181.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing 7*, 57–68.

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

Gerlach, R. and J. Stamey (2007). Bayesian model selection for logistic regression with misclassified outcomes. *Statistical Modelling 7*, 255–273.

Gleser, L. J. (1996). Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. *Contemporary Mathematics 112*, 99–114.

Gramacy, R. B. and N. G. Polson (2012). Simulation-based regularized logistic regression. *Bayesian Analysis 7*, 567–590.

Guhaniyogi, R., S. Qamar, and D. B. Dunson (2017). Bayesian tensor regression. *The Journal of Machine Learning Research 18*, 2733–2763.

Gustafson, P. (2003). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. CRC Press.

Hoff, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Analysis 6*, 179–196.

Holmes, C. C. and L. Held (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis 1*, 145–168.

Hung, H. and C.-C. Wang (2013). Matrix variate logistic regression model with application to EEG data. *Biostatistics 14*, 189–202.

Ishwaran, H. and J. S. Rao (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics 33*, 730–773.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM review 51*, 455–500.

Ledoit, O. and M. Wolf (2004). A well conditioned estimator for large dimensional covariance matrices. *Journal of Multivariate Analysis 88*, 365–411.

Li, B., M. K. Kim, and N. Altman (2010). On dimension folding of matrix-or array-valued statistical objects. *The Annals of Statistics 38*, 1094–1121.

Li, X. (2014). *Tensor Based Statistical Models with Applications in Neuroimaging Data Analysis*. Ph.D. Thesis, North Carolina State University, United States.

Li, X., H. Zhou, and L. Li (2018). Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences 10*, 520–545.

Lindsay, B. (1982). Conditional score functions: some optimality results. *Biometrika 69*, 503–512.

Lokhorst, J. (1999). The Lasso and generalised linear models. *Honors Project. University of Adelaide, Adelaide*.

Lukas Meier, S. v. d. G. and P. Bühlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B 70*, 53–71.

Ma, Y. and R. Li (2010). Variable selection in measurement error models. *Bernoulli 16*, 274–300.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association 54*, 173–205.

Mangasarian, O., W. Street, and W. Wolberg (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research 43*, 570–577.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. Chapman and Hall/CRC, London.

McInturff, P., W. O. Johnson, D. Cowling, and I. A. Gardner (2004). Modelling risk when binary outcomes are subject to error. *Statistics in Medicine 23*, 1095–1109.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika 86*, 843–855.

Neuhaus, J. M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics 58*, 675–683.

O'Brien, S. M. and D. B. Dunson (2004). Bayesian multivariate logistic regression. *Biometrics 60*, 739–746.

Park, T. and G. Casella (2008). The Bayesian Lasso. *Journal of the American Statistical Association 103*, 681–686.

Paulino, C. D., P. Soares, and J. Neuhaus (2003). Binomial regression with misclassification. *Biometrics 59*, 670–675.

Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association 108*, 1339–1349.

Polson, N. G., J. G. Scott, and J. Windle (2014). The Bayesian bridge. *Journal of the Royal Statistical Society: Series B 76*, 713–733.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika 69*, 331–342.

Rekaya, R., K. A. Weigel, and D. Gianola (2001). Threshold model for misclassified binary responses with applications to animal breeding. *Biometrics 57*, 1123–1129.

Richardson, S. and W. R. Gilks (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology 138*, 430–442.

Rossi, P. E., G. M. Allenby, and R. McCulloch (2005). *Bayesian Statistics and Marketing.* John Wiley & Sons.

Roy, S., T. Banerjee, and T. Maiti (2005). Measurement error model for misclassified binary responses. *Statistics in Medicine 24*, 269–283.

Shen, W. and S. Ghosal ́ (2016). Adaptive Bayesian density regression for high-dimensional data. *Bernoulli 22*, 396–420.

Shevade, S. K. and S. S. Keerthi (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics 19*, 2246–2253.

Sobel, M. E. and M. A. Lindquist (2014). Causal inference for fmri time series data with systematic errors of measurement in a balanced on/off study of social evaluative threat. *Journal of the American Statistical Association 109*, 967–976.

Stefanski, L. A. and R. J. Carroll (1985). Covariate measurement error in logistic regression. *The Annals of Statistics 13*, 1335–1351.

Stefanski, L. A. and R. J. Carroll (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika 74*, 703–716.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association 82*, 528–540.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B 58*, 267–288.

Tsanas, A., M. A. Little, C. Fox, , and L. O. Ramig (2013). Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 22*, 181–190.

van der Varrt, A. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics 11*, 284–300.

Walker, S. H. and D. B. Duncan (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika 54*, 255–273.

Wang, C. and M. S. Pepe (2000). Expected estimating equations to accommodate covariate measurement error. *Journal of the Royal Statistical Society: Series B 62*, 509–524.

Wang, H., R. Li, and C.-L. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*, 553–568.

Wei, R. and S. Ghosal (2020). Contraction properties of shrinkage priors in logistic regression. *Journal of Statistical Planning and Inference 207*, 215–229.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society 50*, 1–25.

Yang, J., D. Zhang, A. Frangi, and J. Yang (2004). Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 5*, 131–137.

Ye, J. (2005). Generalized low rank approximations of matrices. *Machine Learning 61*, 167–191.

Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application.* Springer-Verlag, New York.

Yi, G. Y. and R. J. Cook (2005). Errors in the measurement of covariates. *Encyclopedia of Biostatistics 4*, 1741–1748.

Yi, G. Y., Y. Ma, and R. J. Carroll (2012). A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika 99*, 151–165.

Yi, G. Y., Y. Ma, D. Spiegelman, and R. J. Carroll (2015). Functional and structural methods with mixed measurement error and misclassification in covariates. *Journal of the American Statistical Association 110*, 681–696.

Yi, G. Y. and N. Reid (2010). A note on misspecified estimating functions. *Statistica Sinica 20*, 1749–1769.

Yi, G. Y., X. Tan, and R. Li (2015). Variable selection and inference procedures for marginal analysis of longitudinal data with missing observations and covariate measurement error. *The Canadian Journal of Statistics 43*, 498–518.

Zeger, S. L. and M. Karim (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association 86*, 79–86.

Zellner, A. and P. E. Rossi (1984). Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics 25*, 365–393.

Zhang, D. and Z.-H. Zhou (2005). (2D)2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing 69*, 224–231.

Zhang, X., L. Li, H. Zhou, and D. Shen (2014). Tensor generalized estimating equations for longitudinal imaging analysis. *arXiv preprint arXiv:1412.6592*.

Zhou, H. and L. Li (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B 76*, 463–483.

Zhou, H., L. Li, and H. Zhu (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association 108*, 540–552.

Zou, H. (2006). The Adaptive Lasso and its oracle properties. *Journal of the American statistical Association 101*, 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B 67*, 301–320.

APPENDICES

In this part, we report supplementary materials associated with Chapters 2-5, including regularity conditions, the proofs of the theoretical results, additional numerical results, and details of calculations.

Appendix A

Additional Materials for Chapter 2

A.1 Regularity Conditions

(C.1) a) $\sup_n\{(1/n)\sum_{k=1}^n \|\mathrm{vec}(x_{\mathrm{c}k})\|^2\} < \infty$, $\max_{1\le k\le n}\|x_{\mathrm{c}k}\| = o\{\max(m, \sqrt{n})\}$;

b) $\sup_n\{(1/n)\sum_{k=1}^n \|\mathrm{vec}(z_k)\|^2\} < \infty$, $\max_{1\le k\le n}\|z_k\| = o\{\max(m, \sqrt{n})\}$;

where $\|A\|$ denotes the Frobenius norm if A is a matrix and the Euclidean norm if A is a vector, and m is the number of replicates defined in the end of Section 2.3.1.

(C.2) $\sum_{k=1}^n \|x_{\mathrm{c}k}\|^2 = o(n^2)$ and $\sum_{k=1}^n \|z_k\|^2 = o(n^2)$.

(C.3) $E\{\mathrm{vec}(E_{kr})\} = 0$ and $E\{\|\mathrm{vec}(E_{kr})\|^{2+\kappa}\} < \infty$ for some constant $\kappa > 0$.

(C.4) Let

$$G_n(\theta^*) = (1/n)\sum_{k=1}^n Y_k(\alpha^{*\intercal}x_{\mathrm{c}k}\beta^* + \gamma^{*\intercal}z_k) - \log\{1+\exp(\alpha^{*\intercal}x_{\mathrm{c}k}\beta^* + \gamma^{*\intercal}z_k)\}. \quad \text{(A.1)}$$

Assume that there exists a real-valued function $G(\cdot)$ such that for any $\epsilon > 0$

$$\sup_{\theta^*\in\Theta} |G_n(\theta^*) - G(\theta^*)| \to 0 \text{ in probability as } n \to \infty;$$

$$\sup_{\theta^*:d(\theta^*,\theta)\ge\epsilon} G(\theta^*) < G(\theta),$$

where $d(a, b)$ is the distance function in a Euclidean space, say, $\mathbb{R}^d$, defined as $d(a, b) = \|a - b\|$ for $a, b \in \mathbb{R}^d$.

(C.5) There exist a positive definite matrix M_1 and constants $\delta > 0$ and $0 < N_0 < \infty$, such that $H_n(\theta^*) \geq M_1$ whenever $n \geq N_0$ and $\|\theta^* - \theta\| \leq \delta$, where the operation $\geq$ is the Loewner order, i.e., for two matrices A and B, if $A - B$ is semi-positive definite, then we write $A \geq B$.

(C.6) Define

$$\mathrm{S}_k^*(\theta^*) = \begin{pmatrix} \tilde{X}_k^{*\intercal}(\theta^*) \\ z_k \end{pmatrix} \{Y_k - p_k(\theta^*; X_k^*)\},$$

where $\tilde{X}_k^*(\theta) = (\beta^\intercal X_k^{*\intercal} C_t, \alpha^\intercal X_k^*)^\intercal$. Assume that the second derivative of $\mathrm{S}_k^*(\theta^*)$ with respect to θ^* exists and that the entries of $\mathrm{S}_k^*(\theta^*)$ are uniformly bounded by a random variable which may be a function of X_k^* and z_k, say $M_2(X_k^*, z_k)$, in a neighborhood of θ. In addition, $\mathrm{E}\{M_2(X_k^*, z_k)\} < \infty$ for all $k = 1, ..., n$.

Remark 1: In Condition (C.1) will be used to prove the approximation form of $\mathrm{S}_n^*(\theta)$, the consistency of $\mathrm{H}_n^*(\theta)$, and Theorem 2.4. Condition (C.2) is used to prove the consistency of $\mathrm{H}_n^*(\theta)$ in Appendix A.5 and Theorem 2.4 in Appendix A.10. Condition (C.4), also made by van der Varrt (1998, Theorem 5.7) and Zhou et al. (2013, Theorem 1), is used to show the consistency of the naive estimator $\hat{\theta}^*$. This assumption can be regarded as an analogue of Condition (C.1) required by Stefanski and Carroll (1985, p.1337) for logistic regression where only vector-covariates are involved. van der Varrt (1998, p.46) discussed a set of sufficient conditions, including the compactness of the parameter space, that make Condition (C.4) hold. Condition (C.5) is a regularity condition that is needed for the establishment of the asymptotic normal distribution for $H_n^{-1/2}(\theta)L_n(\theta)$ (Stefanski and Carroll 1985, p.1338). Condition (C.6) guarantees that the reminder terms of the Taylor series expansion (A.26) of $\mathrm{S}_k^*(\theta^*)$ in Appendix A.6 are bounded and ignorable when deriving the equation (2.12).

Condition (C.3) characterizes that measurement error cannot be arbitrarily large and must be bounded. This assumption immediately implies that $E\{\mathrm{vec}(\bar{U}_k)\} = 0$,

$$(1/n)\sum_{k=1}^{n} \|\bar{U}_k\|^2 = O_p(1/m) \text{ and } (1/n)\sum_{k=1}^{n} \|\bar{U}_k\| = O_p(1/m^{1/2}). \tag{A.2}$$

Conditions (C.1)-(C.5) are made in the same spirit of Stefanski and Carroll (1985), but these assumptions generalize the requirements for settings with vector-form covariates to accommodating problems with both vector-form covariates and matrix-form covariates. One may notice that the proofs of our results share the same ideas of Stefanski and Carroll

(1985) to certain extent. For instance, modifications of Lemmas 5.1 and 5.2 in Stefanski and Carroll (1985) are used in our proofs to show the relationships between $\mathrm{S}_n^*(\theta)$ and $\mathrm{Z}_n(\theta)$ and between $\mathrm{H}_n^*(\theta)$ and $\mathrm{H}_n(\theta)$. However, our derivations are a lot more technically involved where a key challenge is to figure out how to split α and β in order to establish the results for α and β separately. The presence of matrix-form covariates considerably complicates the derivations of new theoretical results.

A.2 Approximations of $p_k(\theta; X_k^*)$ and its Functions

Here we derive the first and second order Taylor expansions of $p_k(\theta; X_k^*)$ or its function to be used to find the approximation of $S_n^*(\theta)$ in Appendix A.4. To this end, we adapt the derivations for Lemma 5.2 of Stefanski and Carroll (1985).

Let $\eta(X_k^*) = \alpha^\intercal X_k^* \beta + \gamma^\intercal z_k$ and $\eta(x_{\mathrm{c}k}) = \alpha^\intercal x_{\mathrm{c}k}\beta + \gamma^\intercal z_k$ with the dependece on z_k suppressed in the symbols $\eta(X_k^*)$ and $\eta(x_{\mathrm{c}k})$. By (2.5), $X_k^* = x_{\mathrm{c}k} + \bar{U}_k$ and $E(\bar{U}_k) = 0$. Now we write $p_k(\theta; X_k^*)$ as $p_k\{\eta(X_k^*)\}$ and $p_k(\theta; x_{\mathrm{c}k})$ as $p_k\{\eta(x_{\mathrm{c}k})\}$, and consider the following four approximations.

1°. Given x_{ck} and z_k as well as a realization of X_k^*, we derive the first-order Taylor series expansion of $p_k\{\eta(X_k^*)\}$ around $\eta(x_{\mathrm{c}k})$:

$$p_k\{\eta(X_k^*)\} = p_k\{\eta(x_{\mathrm{c}k})\} + p_k^{(1)}\{\eta(x_{k,\xi})\}\{\eta(X_k^*) - \eta(x_{\mathrm{c}k})\}, \tag{A.3}$$

where $\eta(x_{k,\xi}) = \alpha^\intercal x_{k,\xi}\beta + \gamma^\intercal z_k$ with $x_{k,\xi}$ "between" X_k^* and x_{ck} in the sense that $\|x_{k,\xi} - X_k^*\| \leq \|X_k^* - x_{ck}\|$ and $\|x_{k,\xi} - x_{ck}\| \leq \|X_k^* - x_{ck}\|$.

By definition of $p_k(\cdot)$ and $\eta(\cdot)$, and $\eta(x_{k,\xi}) = \eta(x_{\mathrm{c}k}) + o_p(1)$, we write (A.3) as

$$\begin{aligned} p_k\{\eta(X_k^*)\} =& p_k\{\eta(x_{\mathrm{c}k})\} + p_k\{\eta(x_{k,\xi}, z_k)\}[1 - p_k\{\eta(x_{k,\xi})\}] \times (\alpha^\intercal X_k^*\beta - \alpha^\intercal x_{\mathrm{c}k}\beta) \\ =& p_k\{\eta(x_{\mathrm{c}k})\} + p_k\{\eta(x_{k,\xi})\}[1 - p_k\{\eta(x_{k,\xi})\}] \times \mathrm{vec}(\alpha\beta^\intercal)^\intercal \mathrm{vec}(X_k^* - x_{ck}) \\ =& p_k\{\eta(x_{\mathrm{c}k})\} + v_{1,k}\{\eta(x_{k,\xi})\} \times \mathrm{vec}(\alpha\beta^\intercal)^\intercal \mathrm{vec}(\bar{U}_k), \end{aligned} \tag{A.4}$$

where we use the fact that $a^\intercal Ab = \mathrm{vec}(ab^\intercal)^\intercal \mathrm{vec}(A)$ for a $p \times q$ matrix A, a $p \times 1$ vector a

and $q \times 1$ vector b, and $v_{1,k}(\cdot) = p_k(\cdot)\{1 - p_k(\cdot)\}$, as defined in Section 2.3.

2°. Similarly, the first-order Taylor series expansion of $p_k\{\eta(\hat{\Delta}_k)\}$ around $\eta(x_{\mathrm{c}k})$, where $\hat{\Delta}_k$ is defined in (2.22), is

$$p_k\{\eta(\hat{\Delta}_k)\} = p_k\{\eta(x_{\mathrm{c}k})\} + \mathrm{vec}(\alpha\beta^\intercal)^\intercal\{\mathrm{vec}(\bar{U}_k) + \mathrm{vec}(g_k)^\intercal\}v_{1,k}\{\eta(\Delta_{k,\xi})/m_c\}, \tag{A.5}$$

where $\eta(\Delta_{k,\xi}) = \alpha^\intercal\Delta_{k,\xi}\beta + \gamma^\intercal z_k$ with $\Delta_{k,\xi}$ "between" $\hat{\Delta}_k$ and $x_{\mathrm{c}k}$ in the sense that $\|\Delta_{k,\xi} - \hat{\Delta}_k\| \leq \|\hat{\Delta}_k - x_{ck}\|$ and $\|\Delta_{k,\xi} - x_{ck}\| \leq \|\hat{\Delta}_k - x_{ck}\|$.

Then, taking the differece of (A.4) and (A.5), we obtain that

$$\begin{aligned}\|p_k(\theta, X_k^*) - p_k(\theta, \hat{\Delta}_k)\| &= \Big\|\mathrm{vec}(\alpha\beta^\intercal)^\intercal\mathrm{vec}(\bar{U}_k)\Big[v_{1,k}\{\eta(x_{k,\xi})\} - v_{1,k}\{\eta(\Delta_{k,\xi})\}\Big] \\ &\quad -\mathrm{vec}(g_k)^\intercal\mathrm{vec}(\alpha\beta^\intercal)v_{1,k}\{\eta(\Delta_{k,\xi})/m_c\Big\| \\ &\leq \Big\|\mathrm{vec}(\alpha\beta^\intercal)^\intercal\mathrm{vec}(\bar{U}_k)\Big[v_{1,k}\{\eta(x_{k,\xi})\} - v_{1,k}\{\eta(\Delta_{k,\xi})\}\Big]\Big\| \\ &\quad + \Big\|\mathrm{vec}(g_k)^\intercal\mathrm{vec}(\alpha\beta^\intercal)v_{1,k}\{\eta(\Delta_{k,\xi})\}/m_c\Big\| \\ &\leq \Big\|\mathrm{vec}(\alpha\beta^\intercal)^\intercal\mathrm{vec}(\bar{U}_k)\Big\| + \Big\|\mathrm{vec}(g_k)^\intercal\mathrm{vec}(\alpha\beta^\intercal)/m_c\Big\|,\end{aligned} \tag{A.6}$$

where the last step is due to the boundedness of $v_{1,k}(\cdot)$ in $[0, 1]$.

3°. The first-order Taylor series expansion of $p_k\{\eta(X_k^*)\}[1 - p_k\{\eta(X_k^*)\}]$ around $\eta(x_{\mathrm{c}k})$:

$$\begin{aligned}&p_k\{\eta(X_k^*)\}[1 - p_k\{\eta(X_k^*)\}] \\ &= p_k\{\eta(x_{\mathrm{c}k})\}[1 - p_k\{\eta(x_{\mathrm{c}k})\}] + \\ &\quad \frac{\partial}{\partial\eta(x_{\mathrm{c}k})}\Big(p_k\{\eta(x_{\mathrm{c}k})\}[1 - p_k\{\eta(x_{\mathrm{c}k})\}]\Big)\Big|_{x_{\mathrm{c}k}=x_{k,\xi_2}} \times \{\eta(X_k^*) - \eta(x_{\mathrm{c}k})\} \\ &= v_{1,k}\{\eta(x_{\mathrm{c}k})\} + v_{2,k}\{\eta(x_{k,\xi_2})\} \times \mathrm{vec}(\alpha\beta^\intercal)^\intercal\mathrm{vec}(\bar{U}_k),\end{aligned} \tag{A.7}$$

where $v_{2,k}(\cdot) = p_k(\cdot)\{1-p_k(\cdot)\}\{1-2p_k(\cdot)\}$, as defined in Section 2.3, and x_{k,ξ_2} is "between" X_k^* and $x_{\mathrm{c}k}$ in the sense that $\|x_{k,\xi_2} - X_k^*\| \leq \|X_k^* - x_{ck}\|$ and $\|x_{k,\xi_2} - x_{ck}\| \leq \|X_k^* - x_{ck}\|$.

We comment that (A.7) differs from the expression obtained from directly plugging

(A.4) into $p_k\{\eta(X_k^*)\}[1-p_k\{\eta(X_k^*)\}]$, and the former expression is simpler the latter case.

4°. The first-order Taylor series expansion of $p_k\{\eta(\hat{\Delta}_k)\}[1-p_k\{\eta(\hat{\Delta}_k)\}]$ around $\eta(x_{\mathrm{c}k})$:

$$\begin{aligned}
&p_k\{\eta(\hat{\Delta}_k)\}[1-p_k\{(\hat{\Delta}_k)\}] \\
&\;= p_k\{\eta(x_{\mathrm{c}k})\}[1-p_k\{\eta(x_{\mathrm{c}k})\}]+ \\
&\quad \frac{\partial}{\partial\eta(x_{\mathrm{c}k})}\Big(p_k\{\eta(x_{\mathrm{c}k})\}[1-p_k\{\eta(x_{\mathrm{c}k})\}]\Big)\Big|_{x_{ck}=\Delta_{k,\xi 2}} \times \{\eta(\hat{\Delta}_k)-\eta(x_{\mathrm{c}k})\} \\
&\;= v_{1,k}\{\eta(x_{\mathrm{c}k})\} + v_{2,k}\{\eta(\Delta_{k,\xi 2})\} \times \mathrm{vec}(\alpha\beta^\intercal)^\intercal\{\mathrm{vec}(\bar{U}_k)+\mathrm{vec}(g_k)^\intercal/m_c\},
\end{aligned} \tag{A.8}$$

where $\Delta_{k,\xi 2}$ is "between" $\hat{\Delta}_k$ and x_{ck} in the sense that $\|\Delta_{k,\xi 2}-X_k^*\| \leq \|\hat{\Delta}_k - x_{ck}\|$ and $\|\Delta_{k,\xi 2}-x_{ck}\| \leq \|\hat{\Delta}_k - x_{ck}\|$.

5°. Furthermore, we derive the second-order Taylor series expansion of $p_k\{\eta(X_k^*)\}$ around $\eta(x_{ck})$:

$$\begin{aligned}
p_k\{\eta(X_k^*)\} =& p_k\{\eta(x_{\mathrm{c}k})\} + p_k^{(1)}\{\eta(x_{\mathrm{c}k})\} \times \{\eta(X_k^*)-\eta(x_{\mathrm{c}k})\} \\
&+ \frac{1}{2!}p_k^{(2)}\{\eta(x_{k,\xi_3})\} \times \{\eta(X_k^*)-\eta(x_{\mathrm{c}k})\}^2 \\
=& p_k\{\eta(x_{\mathrm{c}k})\} + \mathrm{vec}(\alpha\beta^\intercal)^\intercal\mathrm{vec}(\bar{U}_k)v_{1,k}\{\eta(x_{k,\xi_3})\} \\
&+ \frac{1}{2}\mathrm{vec}(\alpha\beta^\intercal)^\intercal\mathrm{vec}(\bar{U}_k)v_{2,k}\{\eta(x_{k,\xi_3})\}\mathrm{vec}(\bar{U}_k)^\intercal\mathrm{vec}(\alpha\beta^\intercal),
\end{aligned} \tag{A.9}$$

where x_{k,ξ_3} is "between" X_k^* and x_{ck} in the sense that $\|x_{k,\xi_3}-X_k^*\| \leq \|X_k^* - x_{ck}\|$ and $\|x_{k,\xi_3}-x_{ck}\| \leq \|X_k^* - x_{ck}\|$.

A.3 Proof of Theorem 1

In contrast to (2.7), we consider function (A.1) in Appendix A.1 which is identical to (2.7) with X_k^* replaced by x_{ck}.

Let

$$\begin{aligned}R_{n,1} &= \ell_n^*(\theta^*) - G_n(\theta^*) \\ &= \frac{1}{n}\sum_{k=1}^{n}\big[Y_k(\alpha^{*\intercal}X_k^*\beta^* - \alpha^{*\intercal}x_{ck}\beta^*) - \log\{1+\exp(\alpha^{*\intercal}X_k^*\beta^* + \gamma^{*\intercal}z_k)\} \\ &\quad + \log\{1+\exp(\alpha^{*\intercal}x_{ck}\beta^* + \gamma^{*\intercal}z_k)\}\big].\end{aligned}$$

Since $X_k^* = x_{ck} + \bar{U}_k$, and $E(\bar{U}_k) = 0$, then by Conditions (C.1), (C.3) and the Weak Law of Large Numbers (WLLN), $X_k^* \to x_{ck}$ in probability as $m \to \infty$. Thus, by the Continuous Mapping Theorem, as $\min(n,m) \to \infty$, $R_{n,1} = o_p(1)$. That is, $\ell_n^*(\theta^*) - G_n(\theta^*) = o_p(1)$. Then, by Condition (C.4), Theorem 1 of Zhou et al. (2013) and Theorem 5.7 of van der Varrt (1998), $\hat{\theta}^*$ is converges to θ in probability as $\min(m,n) \to \infty$.

A.4 Proof of Lemma 2.1

To show (2.13), we examine each term at a time by the following three parts.
Part I: Show that $S_{\alpha,n}^*(\theta) = \frac{1}{\sqrt{n}}Z_{\alpha,n} + (J_{\alpha,n,1} + J_{\alpha,n,2})\mathrm{vec}(\alpha\beta^\intercal) + o_p\{\max\left(\frac{1}{m}, \frac{1}{\sqrt{n}}\right)\}$.

By (2.9),

$$\mathrm{S}_{\alpha,n}^*(\theta) = \mathrm{T}_{n,\alpha,1} + \mathrm{T}_{n,\alpha,2},$$

where

$$\mathrm{T}_{n,\alpha,1} = \frac{1}{n}\sum_{k=1}^{n} C_t^\intercal x_{ck}\beta\{Y_k - p_k(\theta; X_k^*)\} \text{ and } \mathrm{T}_{n,\alpha,2} = \frac{1}{n}\sum_{k=1}^{n} C_t^\intercal \bar{U}_k\beta\{Y_k - p_k(\theta; X_k^*)\}.$$

We now separately derive the approximation of $\mathrm{T}_{n,\alpha,1}$ and $\mathrm{T}_{n,\alpha,2}$ as follows.
1. Show that $T_{n,\alpha,1} = \frac{1}{\sqrt{n}}Z_{\alpha,n} + J_{\alpha,n,1}\mathrm{vec}(\alpha\beta^\intercal) + o_p\big\{\max\left(\frac{1}{m}, \frac{1}{\sqrt{n}}\right)\big\}$**.**

By (A.9), we write $\mathrm{T}_{n,\alpha,1}$ as sum of individual terms each with one particular feature:

$$\mathrm{T}_{n,\alpha,1} = \frac{1}{\sqrt{n}}\mathrm{S}_{\alpha,\mathrm{n}} + \frac{1}{\sqrt{\mathrm{n}}}\mathrm{Q}_{\alpha,\mathrm{n},1} + \mathrm{J}_{\alpha,\mathrm{n},1}\mathrm{vec}(\alpha\beta^\intercal) + \mathrm{D}_{\alpha,\mathrm{n},1} + \mathrm{R}_{\alpha,\mathrm{n},1}, \text{(A.10)}$$

where

$$\mathrm{S}_{\alpha,\mathrm{n}} = \frac{1}{\sqrt{\mathrm{n}}}\sum_{\mathrm{k}=1}^{\mathrm{n}} \mathrm{C}_{\mathrm{t}}^{\intercal}\mathrm{x}_{\mathrm{ck}}\beta\{\mathrm{Y}_{\mathrm{k}} - \mathrm{p}_{\mathrm{k}}(\theta;\mathrm{x}_{\mathrm{ck}})\};$$

$$\mathrm{Q}_{\alpha,n,1} = -\frac{1}{\sqrt{n}}\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{ck}}\beta\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(\bar{U}_k)v_{1,k}(\theta;x_{\mathrm{ck}});$$

$$\mathrm{J}_{\alpha,n,1} = -\frac{n}{2}\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{ck}}\beta\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\frac{\Omega_0}{m_c}v_{2,k}(\theta;x_{\mathrm{ck}});$$

$$\mathrm{D}_{\alpha,n,1} = -\frac{n}{2}\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{ck}}\beta\Big[\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\Big\{\mathrm{vec}(\bar{U}_k)\mathrm{vec}(\bar{U}_k)^{\intercal} - \frac{\Omega_0}{m_c}\Big\}\mathrm{vec}(\alpha\beta^{\intercal})v_{2,k}(\theta;x_{\mathrm{ck}})\Big];$$

$$\mathrm{R}_{\alpha,n,1} = -\frac{n}{2}\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{ck}}\beta\Big[\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(\bar{U}_k)\mathrm{vec}(\bar{U}_k)^{\intercal}\mathrm{vec}(\alpha\beta^{\intercal})\{v_{2,k}(\theta;x_{k,\xi_3}) - v_{2,k}(\theta;x_{\mathrm{ck}})\}\Big].$$

Now we examine each term of (A.8) separately by the following three steps to show that the approximation form of $T_{n,\alpha,1}$ is

$$\mathrm{T}_{n,\alpha,1} = \frac{1}{\sqrt{n}}\mathrm{S}_{\alpha,\mathrm{n}} + \mathrm{J}_{\alpha,\mathrm{n},1}\mathrm{vec}(\alpha\beta^{\intercal}) + \mathrm{o_p}\Big\{\max\Big(\frac{1}{\mathrm{m}},\frac{1}{\sqrt{\mathrm{n}}}\Big)\Big\}.$$

Step1: Show that $Q_{\alpha,n,1} = o_p(1)$.

For $\mathrm{Q}_{\alpha,n,1}$, we modify the discussion of the $Q_{n,1,\sigma}$ term in Lemma 5.2 of Stefanski and Carroll (1985, p.1347) and obtain that $\mathrm{Q}_{\alpha,n,1}$ has mean zero, and

$$\begin{aligned}
Var(\mathrm{Q}_{\alpha,n,1}) &= \frac{1}{n^2}\sum_{k=1}^{n} Var\{C_t^{\intercal} x_{\mathrm{ck}}\beta\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(\bar{U}_k)v_{1,k}(\theta;x_{\mathrm{ck}})\} \\
&= \frac{1}{n^2}\sum_{k=1}^{n} v_{1,k}^2(\theta;x_{\mathrm{ck}}) \times \|C_t^{\intercal}\|^2 \times \|\beta\|^2 \times \|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\|^2 \times \|x_{\mathrm{ck}}\|^2 \times Var\{\mathrm{vec}(\bar{U}_k)\} \\
&\le \frac{1}{n^2}\sum_{k=1}^{n} \|C_t^{\intercal}\|^2 \times \|\beta\|^2 \times \|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\|^2 \times \|x_{\mathrm{ck}}\|^2 \times \|\bar{U}_k\|^2 \\
&\le \|C_t^{\intercal}\|^2 \times \|\beta\|^2 \times \|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\|^2 \times \frac{1}{n}\sum_{k=1}^{n}\|x_{\mathrm{ck}}\|^2 \times \frac{1}{n}\sum_{k=1}^{n}\|\bar{U}_k\|^2
\end{aligned}$$

where the third step comes from that $v_{1,k}(\theta;x_{\mathrm{ck}})$ is bounded between $[0,1]$, and the last step is because that

$$\frac{1}{n}\sum_{k=1}^{n}\|x_{\mathrm{c}k}\|^2 \times \frac{1}{n}\sum_{k=1}^{n}\|\bar{U}_k\|^2 = \frac{1}{n^2}\sum_{k=1}^{n}\|x_{\mathrm{c}k}\|^2 \times \|\bar{U}_k\|^2 + \frac{1}{n^2}\sum_{k=1}^{n}\sum_{j\neq k}^{n}\|x_{\mathrm{c}k}\|^2 \times \|\bar{U}_j\|^2.$$

According to Condition (C.1) and the derivation of Condition C.3, we have $\frac{1}{n}\sum_{k=1}^{n}\|x_{\mathrm{c}k}\|^2 = O(1)$ and $\frac{1}{n}\sum_{k=1}^{n}\|\bar{U}_k\|^2 = O_p(\frac{1}{m})$. As a result, $Var(\mathrm{Q}_{\alpha,n,1}) = o_p(1)$ as $\min(m,n) \to \infty$, thus $\mathrm{Q}_{\alpha,n,1} = o_p(1)$ and $\frac{1}{\sqrt{n}}\mathrm{Q}_{\alpha,n,1} = o_p(\frac{1}{\sqrt{n}})$.

Step2: Show that** $D_{\alpha,n,1} = o_p\left(\frac{1}{m}\right)$**.

To examine $\mathrm{D}_{\alpha,n,1}$, we adapt the derivations of the $D_{n,1}$ in Lemma 5.2 of Stefanski and Carroll (1985, p.1347) and obtain that

$$\begin{aligned}\|\mathrm{D}_{\alpha,n,1}\| &= \|C_t^{\intercal}\| \times \|\beta\| \times \|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\| \times \|v_{2,k}(\theta;x_{\mathrm{c}k})\| \times \frac{n}{2}\sum_{k=1}^{n}\|x_{\mathrm{c}k}\| \\ &\quad \times \left\|\mathrm{vec}(\bar{U}_k)\mathrm{vec}(\bar{U}_k)^{\intercal} - \frac{\Omega_0}{m_c}\right\| \\ &\leq \text{Constant} \times \frac{1}{m_c}\Big(\frac{1}{n}\sum_{k=1}^{n}\|x_{\mathrm{c}k}\|^2\Big)^{1/2} \times \Big(\frac{1}{n}\sum_{k=1}^{n}\|m_c\mathrm{vec}(\bar{U}_k)\mathrm{vec}(\bar{U}_k)^{\intercal} - \Omega_0\|^2\Big)^{1/2}.\end{aligned} \tag{A.11}$$

By (A.1), $\frac{1}{n}\sum_{k=1}^{n}\|x_{\mathrm{c}k}\|^2 = O(1)$, and by the Markov Inequality, we have that for any scalar $\epsilon > 0$,

$$P\Big\{\frac{1}{n}\sum_{k=1}^{n}\|m_c\mathrm{vec}(\bar{U}_k)\mathrm{vec}(\bar{U}_k)^{\intercal} - \Omega_0\|^2 > \epsilon\Big\} \leq \frac{\sum\limits_{k=1}^{n} E\Big\{\|m_c\mathrm{vec}(\bar{U}_k)\mathrm{vec}(\bar{U}_k)^{\intercal} - \Omega_0\|^2\Big\}}{n\epsilon}. \tag{A.12}$$

Now for the numerator of the right-hand-side of (A.12), we have that, by the definition of $\bar{U}_k$,

$$
\begin{aligned}
&E\Big\{\|m_c\text{vec}(\bar{U}_k)\text{vec}(\bar{U}_k)^\intercal-\Omega_0\|^2\Big\}\\
&=E\left\|\frac{n}{m(n-1)}\left\{\sum_{r=1}^{m}\text{vec}(E_{kr})-\frac{\sum\limits_{k=1}^{n}\sum\limits_{r=1}^{m}\text{vec}(E_{kr})}{n}\right\}\left\{\sum_{r=1}^{m}\text{vec}(E_{kr})-\frac{\sum\limits_{k=1}^{n}\sum\limits_{r=1}^{m}\text{vec}(E_{kr})}{n}\right\}^\intercal-\Omega_0\right\|^2\\
&=\sqrt{\frac{n}{m(n-1)}}E\left[\left\|\left\{\sum_{r=1}^{m}\text{vec}(E_{kr})-\frac{\sum\limits_{k=1}^{n}\sum\limits_{r=1}^{m}\text{vec}(E_{kr})}{n}\right\}\right.\right.\\
&\quad\left.\left.\times\left\{\sum_{r=1}^{m}\text{vec}(E_{kr})-\frac{\sum\limits_{k=1}^{n}\sum\limits_{r=1}^{m}\text{vec}(E_{kr})}{n}\right\}^\intercal-\frac{m(n-1)}{n}\Omega_0\right\|^2\right]\\
&=O_p\Big(\frac{\sqrt{n}}{\sqrt{m}\sqrt{(n-1)}}\Big)=o_p(1)
\end{aligned}
$$

as $\min(m,n)\to\infty$, where in the second last step we use the fact that

$$
E\left[\left\{\sum_{r=1}^{m}\text{vec}(E_{kr})-\frac{\sum\limits_{k=1}^{n}\sum\limits_{r=1}^{m}\text{vec}(E_{kr})}{n}\right\}\left\{\sum_{r=1}^{m}\text{vec}(E_{kr})-\frac{\sum\limits_{k=1}^{n}\sum\limits_{r=1}^{m}\text{vec}(E_{kr})}{n}\right\}^\intercal-\frac{m(n-1)}{n}\Omega_0\right]=0.
$$

Thus, by (A.12),

$$
P\Big\{\frac{1}{n}\sum_{k=1}^{n}\|m_c\text{vec}(\bar{U}_k)\text{vec}(\bar{U}_k)^\intercal-\Omega_0\|^2>\epsilon\Big\}=o_p(1)
$$

as $\min(m,n)\to\infty$. Thus (A.11) implies that $\mathrm{D}_{\alpha,n,1}=o_p(\frac{1}{m})$.

Step3: Show that $R_{\alpha,n,1}=o_p(\frac{1}{m})$.

To examine $\mathrm{R}_{\alpha,n,1}$, we first note that $v_{2,k}\{\cdot\}$ is defined in Section 2.2.3 with $p_k(\cdot)\in[0,1]$, it is readily to show that $v_{2,k}\{\cdot\}$ has the maximum value $\frac{\sqrt{3}}{18}$ and the minimum value $-\frac{\sqrt{3}}{18}$, i.e., $v_{2,k}(\cdot)\in[-\frac{\sqrt{3}}{18},\frac{\sqrt{3}}{18}]$. Thus, $|v_{2,k}(\theta;x_{k,\xi_3})-v_{2,k}(\theta;x_{ck})|<1$ because of the boundedness of $v_{2,k}(\cdot)$. Then we obtain that

$$
\begin{aligned}
\|\mathrm{R}_{\alpha,n,1}\| &\leq \frac{n}{2}\sum_{k=1}^{n}\Big\|C_t^{\intercal}x_{ck}\beta\{\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(\bar{U}_k)\mathrm{vec}(\bar{U}_k)^{\intercal}\mathrm{vec}(\alpha\beta^{\intercal})\}\\
&\quad \times |v_{2,k}(\theta;x_{k,\xi_3})-v_{2,k}(\theta;x_{ck})|\Big\|\\
&\leq \frac{n}{2}\sum_{k=1}^{n}\big\|C_t^{\intercal}x_{ck}\beta\{\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(\bar{U}_k)\mathrm{vec}(\bar{U}_k)^{\intercal}\mathrm{vec}(\alpha\beta^{\intercal})\}\big\|\\
&\leq \frac{n}{2}\|C_t^{\intercal}\|\times\|\beta\|\times\sum_{k=1}^{n}\|x_{ck}\|\times\{\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(\bar{U}_k)\}^2\\
&\leq \frac{1}{2}\|C_t^{\intercal}\|\times\|\beta\|\times\frac{1}{m_c^2}\max_{1\leq k\leq n}\|x_{ck}\|\times\frac{1}{n}\sum_{k=1}^{n}\Big\{\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}\Big(\sum_{r=1}^{m}E_{kr}\\
&\quad -\frac{1}{n}\sum_{k=1}^{n}\sum_{r=1}^{m}E_{kr}\Big)\Big\}^2\\
&\leq \frac{1}{2}\|C_t^{\intercal}\|\times\|\beta\|\times\frac{n^2}{(n-1)^2}\times o\Big(\frac{1}{m}\Big)\times\frac{1}{n}\sum_{k=1}^{n}\Big\{\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}\Big(\sum_{r=1}^{m}E_{kr}\\
&\quad -\frac{1}{n}\sum_{k=1}^{n}\sum_{r=1}^{m}E_{kr}\Big)\Big\}^2\\
&= o_p\Big(\frac{1}{m}\Big),
\end{aligned}
$$

where Condition (C.1) is used in the second last step, and the last step comes from that

$$\frac{1}{n}\sum_{k=1}^{n}\Big\{\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}\Big(\sum_{r=1}^{m}E_{kr}-\frac{1}{n}\sum_{k=1}^{n}\sum_{r=1}^{m}E_{kr}\Big)\Big\}^2=O_p(1).$$

Finally, applying the results of Steps 1-3 to (A.10), we obtain that

$$\mathrm{T}_{n,\alpha,1}=\frac{1}{\sqrt{n}}\mathrm{S}_{\alpha,n}+\mathrm{J}_{\alpha,n,1}\mathrm{vec}(\alpha\beta^{\intercal})+o_p\Big\{\max\Big(\frac{1}{m},\frac{1}{\sqrt{n}}\Big)\Big\}. \tag{A.13}$$

2. Show that $T_{n,\alpha,2}=J_{\alpha,n,2}\mathrm{vec}(\alpha\beta^{\intercal})+o_p\{\max\big(\frac{1}{m},\frac{1}{\sqrt{n}}\big)\}$**.**

By (A.4), we write $\mathrm{T}_{n,\alpha,2}$ as the sum of individual terms each with one particular feature:

$$\mathrm{T}_{n,\alpha,2} = -\frac{1}{\sqrt{n}}\mathrm{Q}_{\alpha,n,2} + \mathrm{J}_{\alpha,n,2}\mathrm{vec}(\alpha\beta^{\intercal}) + \mathrm{D}_{\alpha,n,2} + \mathrm{R}_{\alpha,n,2}, \tag{A.14}$$

where

$$\begin{aligned}
\mathrm{Q}_{\alpha,n,2} &= \frac{1}{\sqrt{n}}\sum_{k=1}^{n} C_t^{\intercal}\bar{U}_k\beta\{Y_k - p_k(\theta; x_{\mathrm{c}k})\};\\
\mathrm{J}_{\alpha,n,2} &= -\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}\Pi_\alpha\frac{\Omega_0}{m_c}v_{1,k}(\theta; x_{\mathrm{c}k});\\
\mathrm{D}_{\alpha,n,2} &= -\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}\Pi_\alpha\Big\{\mathrm{vec}(\bar{U}_k)\mathrm{vec}(\bar{U}_k)^T - \frac{\Omega_0}{m_c}\Big\}v_{1,k}(\theta; x_{\mathrm{c}k})\mathrm{vec}(\alpha\beta^{\intercal});\\
\mathrm{R}_{\alpha,n,2} &= -\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}\Pi_\alpha[\mathrm{vec}(\bar{U}_k)\mathrm{vec}(\bar{U}_k)^{\intercal}\{v_{1,k}(\theta; x_{k,\xi}) - v_{1,k}(\theta; x_{\mathrm{c}k})\}\mathrm{vec}(\alpha\beta^{\intercal})];
\end{aligned}$$

and $\Pi_\alpha = \begin{bmatrix}\beta_1 I_{(p+1)} & \beta_2 I_{(p+1)} & \cdots & \beta_q I_{(p+1)}\end{bmatrix}$ is a $(p+1)\times\{(p+1)q\}$ matrix.

Now we examine each term of (A.14) separately to show that the approximation form of $T_{n,\alpha,2}$ is

$$\mathrm{T}_{n,\alpha,2} = \mathrm{J}_{\alpha,n,2}\mathrm{vec}(\alpha\beta^{\intercal}) + o_p\Big\{\max\Big(\frac{1}{m}, \frac{1}{\sqrt{n}}\Big)\Big\}. \tag{A.15}$$

Like $\mathrm{Q}_{\alpha,n,1}$ discussed in the preceding Step 1, $\mathrm{Q}_{\alpha,n,2}$ has mean zero and $Var(\mathrm{Q}_{\alpha,n,2}) = o_p(1)$ as $\min(m,n)\to\infty$, thus $\mathrm{Q}_{\alpha,n,2} = o_p(1)$. To examine $\mathrm{D}_{\alpha,n,2}$, we adapt the derivations of $\mathrm{D}_{\alpha,n,1}$ in (A.11) to obtain that $\mathrm{D}_{\alpha,n,2} = o_p(1)$ as $\min(m,n)\to\infty$. To examine $\mathrm{R}_{\alpha,n,2}$, we note that both $v_{1,k}\{\eta(x_{k,\xi})\}$ and $v_{1,k}\{\eta(x_{\mathrm{c}k})\}$ are in $[0,1]$, thus yielding that their difference is bounded. Similar to the derivation of $\mathrm{R}_{\alpha,n,1}$, we obtain that $\mathrm{R}_{\alpha,n,2} = o_p(1)$ as $\min(m,n)\to\infty$. Thus, applying these results of $\mathrm{Q}_{\alpha,n,2}$, $\mathrm{D}_{\alpha,n,2}$ and $\mathrm{R}_{\alpha,n,2}$ to (A.14), we obtain (A.15).

Finally, combining (A.13) and (A.15) gives

$$\begin{aligned}
\mathrm{S}^*_{\alpha,n}(\theta) &= \mathrm{Z}_{\alpha,n}(\theta) + o_p\Big\{\max\Big(\frac{1}{m}, \frac{1}{\sqrt{n}}\Big)\Big\}\\
&= \frac{1}{\sqrt{n}}\mathrm{S}_{\alpha,n} + (\mathrm{J}_{\alpha,n,1} + \mathrm{J}_{\alpha,n,2})\mathrm{vec}(\alpha\beta^{\intercal}) + o_p\Big\{\max\Big(\frac{1}{m}, \frac{1}{\sqrt{n}}\Big)\Big\}.
\end{aligned} \tag{A.16}$$

Part II: Show that

$$S^*_{\beta,n}(\theta) = \frac{1}{\sqrt{n}} Z_{\beta,n} + (J_{\beta,n,1} + J_{\beta,n,2})\mathrm{vec}(\alpha\beta^\intercal) + o_p\Big\{ \max\Big(\frac{1}{m}, \frac{1}{\sqrt{n}}\Big)\Big\}.$$

By (2.9),

$$\mathrm{S}^*_{\beta,n}(\theta) = \mathrm{T}_{n,\beta,1} + \mathrm{T}_{n,\beta,2},$$

where

$$\mathrm{T}_{n,\beta,1} = \frac{1}{n}\sum_{k=1}^{n} x_{\mathrm{c}k}^\intercal \alpha\{Y_k - p_k(\theta; X_k^*)\} \text{ and } \mathrm{T}_{n,\beta,2} = \frac{1}{n}\sum_{k=1}^{n} \bar{U}_k \alpha\{Y_k - p_k(\theta; X_k^*)\},$$

Analogous to Part I, we separately examine $\mathrm{T}_{n,\beta,1}$ and $\mathrm{T}_{n,\beta,2}$ and obtain that

$$\begin{aligned} \mathrm{S}^*_{\beta,n}(\theta) &= \mathrm{Z}_{\beta,n}(\theta) + o_p\Big\{ \max\Big(\frac{1}{m}, \frac{1}{\sqrt{n}}\Big)\Big\} \\ &= \frac{1}{\sqrt{n}}\mathrm{S}_{\beta,n} + (\mathrm{J}_{\beta,n,1} + \mathrm{J}_{\beta,n,2})\mathrm{vec}(\alpha\beta^\intercal) + o_p\Big\{ \max\Big(\frac{1}{m}, \frac{1}{\sqrt{n}}\Big)\Big\}, \end{aligned} \tag{A.17}$$

where

$$\begin{aligned} \mathrm{S}_{\beta,n} &= \frac{1}{\sqrt{n}}\sum_{k=1}^{n} x_{\mathrm{c}k}^\intercal \alpha\{y_k - p_k(\theta; x_{\mathrm{c}k})\}, \\ \mathrm{J}_{\beta,n,1} &= -\frac{n}{2}\sum_{k=1}^{n} x_{\mathrm{c}k}^\intercal \alpha \mathrm{vec}(\alpha\beta^\intercal)^\intercal \frac{\Omega_0}{m_c} v_{2,k}(\theta; x_{\mathrm{c}k}), \\ \mathrm{J}_{\beta,n,2} &= -\frac{1}{n}\sum_{k=1}^{n} \Pi_\beta \frac{\Omega_0}{m_c} v_{1,k}(\theta; x_{\mathrm{c}k}), \end{aligned}$$

and

$$\Pi_\beta = \begin{bmatrix} \alpha^\intercal & & & \\ & \alpha^\intercal & & \\ & & \ddots & \\ & & & \alpha^\intercal \end{bmatrix}_{q\times\{(p+1)q\}}.$$

Part III: Show that $S^*_{\gamma,n}(\theta) = \frac{1}{\sqrt{n}} Z_{\gamma,n} + J_{\gamma,n}\mathrm{vec}(\alpha\beta^\intercal) + o_p\{\max\big(\frac{1}{m}, \frac{1}{\sqrt{n}}\big)\}.$

By (2.9) and (A.8),

$$\begin{aligned} \mathrm{S}^*_{\gamma,n}(\theta) &= \mathrm{Z}_{\gamma,n}(\theta) + o_p\Big\{\max\Big(\frac{1}{m},\frac{1}{\sqrt{n}}\Big)\Big\} \\ &= \frac{1}{\sqrt{n}}\mathrm{S}_{\gamma,n} + \mathrm{J}_{\gamma,n}\mathrm{vec}(\alpha\beta^\intercal) + o_p\Big\{\max\Big(\frac{1}{m},\frac{1}{\sqrt{n}}\Big)\Big\}, \end{aligned} \tag{A.18}$$

where

$$\begin{aligned} \mathrm{S}_{\gamma,n} &= \frac{1}{\sqrt{n}}\sum_{k=1}^{n} z_k\{Y_k - p_k(\theta; x_{ck})\}, \\ \mathrm{J}_{\gamma,n} &= -\frac{n}{2}\sum_{k=1}^{n} z_k\mathrm{vec}(\alpha\beta^\intercal)^\intercal\frac{\Omega_0}{m_c}v_{2,k}(\theta; x_{ck}). \end{aligned}$$

Combining (2.9), (A.16), (A.17) and (A.18) yields the approximation of $\mathrm{S}^*_n(\theta)$ as

$$S^*_n(\theta) = \mathrm{Z}_n(\theta) + o_p\Big\{\max\Big(\frac{1}{m},\frac{1}{\sqrt{n}}\Big)\Big\}.$$

A.5 Proof of Lemma 2.2

To examine $\mathrm{H}^*_n(\theta)$, which is defined for (2.12), we first write it in a block matrix with each block submatrix corresponding to one type of parameters, and we examine each block submatrix one at a time. To be precise, we write

$$\mathrm{H}^*_n(\theta) \triangleq -\begin{pmatrix} \mathrm{H}^*_{\alpha,\beta,n}(\theta) & \mathrm{H}^*_{\alpha\beta,\gamma,n}(\theta) \\ \mathrm{H}^{*\intercal}_{\alpha\beta,\gamma,n}(\theta) & \mathrm{H}^*_{\gamma,\gamma,n}(\theta) \end{pmatrix},$$

where

$$\begin{aligned} \mathrm{H}^*_{\alpha,\beta,n}(\theta) &= \frac{\partial\{\mathrm{S}^*_{\alpha,n}(\theta), \mathrm{S}^*_{\beta,n}(\theta)\}^\intercal}{\partial(\tilde{\alpha}^\intercal, \beta^\intercal)}, \\ \mathrm{H}^*_{\alpha\beta,\gamma,n}(\theta) &= \frac{\partial\{\mathrm{S}^*_{\alpha,n}(\theta), \mathrm{S}^*_{\beta,n}(\theta)\}^\intercal}{\partial\gamma^T}, \\ \mathrm{H}^*_{\gamma,\gamma,n}(\theta) &= \frac{\partial\{\mathrm{S}^*_{\gamma,n}(\theta)\}}{\partial\gamma^T}. \end{aligned}$$

Similarly write $\mathrm{H}_n(\theta)$ (defined for (2.14)) as

$$\mathrm{H}_n(\theta) \triangleq -\begin{pmatrix} \mathrm{H}_{\alpha,\beta,n}(\theta) & \mathrm{H}_{\alpha\beta,\gamma,n}(\theta) \\ \mathrm{H}^{\intercal}_{\alpha\beta,\gamma,n}(\theta) & \mathrm{H}_{\gamma,\gamma,n}(\theta) \end{pmatrix},$$

where

$$\begin{aligned} \mathrm{H}_{\alpha,\beta,n}(\theta) &= \frac{\partial\{\mathrm{Z}_{\alpha,n}(\theta), \mathrm{Z}_{\beta,n}(\theta)\}^{\intercal}}{\partial(\tilde{\alpha}^{\intercal}, \beta^{\intercal})}, \\ \mathrm{H}_{\alpha\beta,\gamma,n}(\theta) &= \frac{\partial\{\mathrm{Z}_{\alpha,n}(\theta), \mathrm{Z}_{\beta,n}(\theta)\}^{\intercal}}{\partial\gamma^{T}}, \\ \mathrm{H}_{\gamma,\gamma,n}(\theta) &= \frac{\partial\{\mathrm{Z}_{\gamma,n}(\theta)\}}{\partial\gamma^{T}}. \end{aligned}$$

In the following three parts, we show that each entry of $\mathrm{H}^*_n(\theta)$ differs from the corresponding part of $\mathrm{H}_n(\theta)$ with a small magnitude.

***Part I: Show that** $H^*_{\alpha,\beta,n}(\theta) - H_{\alpha,\beta,n}(\theta) = o_p(1)$ **as** $\min(m,n) \to \infty$.*

The difference between $\mathrm{H}^*_{\alpha,\beta,n}(\theta)$ and $\mathrm{H}_{\alpha,\beta,n}(\theta)$, we write it as the sum of three terms so we can look at one term at a time:

$$\mathrm{H}^*_{\alpha,\beta,n}(\theta) - \mathrm{H}_{\alpha,\beta,n}(\theta) = \mathrm{H}_{\alpha,\beta,n,1} + \mathrm{H}_{\alpha,\beta,n,2} + \mathrm{H}_{\alpha,\beta,n,3},$$

where $\mathrm{H}_{\alpha,\beta,n,1}$ includes entries depended only on $p_k(\theta; X^*_k)$ and $\tilde{U}_k(\theta) = (\beta^{\intercal}\bar{U}_k C_t, \alpha^{\intercal}\bar{U}^{\intercal}_k)$, $\mathrm{H}_{\alpha,\beta,n,2}$ includes entries depended on $p_k(\theta; X^*_k)$, $p_k(\theta; x^*_k)$ and $\tilde{x}_{\mathrm{c}k}(\theta)$, and $\mathrm{H}_{\alpha,\beta,n,3}$ contains the rest terms. The details of the three terms are given and examined as follows.

1. Show that** $H_{\alpha,\beta,n,1} = o_p(1)$**.

Let $\tilde{x}_{\mathrm{c}k}(\theta) = (\beta^{\intercal}x^{\intercal}_{\mathrm{c}k}C_t, \alpha^{\intercal}x_{\mathrm{c}k})^{\intercal}$ and $\tilde{X}^*_k(\theta) = (\beta^{\intercal}X^{*\intercal}_k C_t, \alpha^{\intercal}X^*_k)^{\intercal}$. The term $\mathrm{H}_{\alpha,\beta,n,1}$ is defined as

$$\mathrm{H}_{\alpha,\beta,n,1} = \frac{1}{n}\sum_{k=1}^{n} v_{1,k}(\theta; X^*_k)\{\tilde{X}^{*\intercal}_k(\theta)\tilde{X}^*_k(\theta) - \tilde{x}^{\intercal}_{\mathrm{c}k}(\theta)\tilde{x}_{\mathrm{c}k}(\theta)\},$$

which by (2.5), equals

$$\frac{1}{n}\sum_{k=1}^{n} v_{1,k}(\theta; X_k^*)\tilde{U}_k^{\intercal}(\theta)\tilde{U}_k(\theta) + \frac{1}{n}\sum_{k=1}^{n} v_{1,k}(\theta; X_k^*)\{\tilde{U}_k^{\intercal}(\theta)\tilde{x}_{ck}(\theta) + \tilde{x}_{ck}^{\intercal}(\theta)\tilde{U}_k(\theta)\}$$
$$\triangleq \mathbf{A.19} + \mathbf{A.20}.$$

Then following the derivation regrading Condition (C.3) in Remark, we have that by $v_{1,k}(\cdot) \in [0, 1]$,

$$\begin{aligned}
\|\mathbf{A.19}\| &\le \frac{1}{n}\sum_{k=1}^{n} \|\tilde{U}_k^{\intercal}(\theta)\tilde{U}_k(\theta)\| \le \frac{1}{n}\sum_{k=1}^{n} \|\tilde{U}_k^{\intercal}(\theta)\|^2 \\
&\le \frac{1}{n}\sum_{k=1}^{n}(\|\beta^{\intercal}\|^2 \times \|\bar{U}_k\|^2 \times \|C_t\|^2 + \|\alpha^{\intercal}\|^2 \times \|\bar{U}_k^{\intercal}\|^2) \\
&= O_p\Big(\frac{1}{m}\Big) \\
&= o_p(1)
\end{aligned}$$

as $\min(m, n) \to \infty$, where the second last step is due to (A.2).

A.20 has same structure as $\mathrm{Q}_{\alpha,n,1}$ in Appendix A.4 and we use the similar steps to derive $\mathbf{A.18} = o_p(1)$ as $\min(m, n) \to \infty$. Thus, $\mathrm{H}_{\alpha,\beta,n,1} = o_p(1)$, as $\min(m, n) \to \infty$.

2. Show that $H_{\alpha,\beta,n,2} = o_p(1)$.

The term $\mathrm{H}_{\alpha,\beta,n,2}$ is defined as

$$\mathrm{H}_{\alpha,\beta,n,2} = \frac{1}{n}\sum_{k=1}^{n}[v_{1,k}(\theta; X_k^*) - v_{1,k}(\theta; x_{ck})]\tilde{x}_{ck}^{\intercal}(\theta)\tilde{x}_{ck}(\theta)$$

Using (A.7), we obtain that

$$\mathrm{H}_{\alpha,\beta,n,2} = \frac{1}{n}\sum_{k=1}^{n} \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(\bar{U}_k)v_{2,k}(\theta; x_{k,\xi_2})\tilde{x}_{ck}^{\intercal}(\theta)\tilde{x}_{ck}(\theta). \tag{A.21}$$

By the boundedness of $v_{2,k}(\cdot)$ which we showed in Appendix A.4, Conditions (C.1), (C.2) and (C.3) together with Lemma 5.1 in Stefanski and Carroll (1985), we have that

$$
\begin{aligned}
\|\mathrm{A.21}\| &\le \frac{1}{n}\sum_{k=1}^{n} \|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(\bar{U}_k)\tilde{x}_{\mathrm{c}k}^{\intercal}(\theta)\tilde{x}_{\mathrm{c}k}(\theta)\| \\
&= \|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\| \times \frac{1}{n}\sum_{k=1}^{n} \|\mathrm{vec}(\bar{U}_k) \times (\beta^{\intercal}x_{\mathrm{c}k}^{\intercal}C_t \times C_t^{\intercal}x_{\mathrm{c}k}\beta + \alpha^{\intercal}x_{\mathrm{c}k} \times x_{\mathrm{c}k}^{\intercal}\alpha)\| \\
&\le \|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\| \times \|\beta\|^2 \times \|C_t\|^2 \times \frac{1}{n}\sum_{k=1}^{n} \|\mathrm{vec}(\bar{U}_k)\| \times \|x_{\mathrm{c}k}\|^2 \\
&\quad + \|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\| \times \|\alpha\|^2 \times \frac{1}{n}\sum_{k=1}^{n} \|\mathrm{vec}(\bar{U}_k)\| \times \|x_{\mathrm{c}k}\|^2 \\
&= o_p(1),
\end{aligned}
$$

when $\min(m,n) \to \infty$. Thus, $\mathrm{H}_{\alpha,\beta,n,2} = o_p(1)$, as $\min(m,n) \to \infty$.

3. Show that $H_{\alpha,\beta,n,3} = o_p(1)$**.**

$\mathrm{H}_{\alpha,\beta,n,3}$ contains the rest terms of $\mathrm{H}^*_{\alpha,\beta,n}(\theta) - \mathrm{H}_{\alpha,\beta,n}(\theta)$ that are not included in $\mathrm{H}_{\alpha,\beta,n,1}$ or $\mathrm{H}_{\alpha,\beta,n,2}$. It is

$$
\mathrm{H}_{\alpha,\beta,n,3} = \begin{pmatrix} 0 & \mathbf{A.22} \\ \mathbf{A.23} & 0 \end{pmatrix},
$$

where

$$
\mathbf{A.22} = \frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}X_k^*\{Y_k - p_k(\theta; X_k^*)\} - C_t^{\intercal}x_{\mathrm{c}k}\{Y_k - p_k(\theta; x_{\mathrm{c}k})\}
$$

and

$$
\mathbf{A.23} = \frac{1}{n}\sum_{k=1}^{n} X_k^{*\intercal}C_t\{Y_k - p_k(\theta; X_k^*)\} - x_{\mathrm{c}k}^{\intercal}C_t\{Y_k - p_k(\theta; x_{\mathrm{c}k})\}.
$$

Plugging (A.4) into **A.22**, we obtain that

$$
\mathbf{A.22} = -\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}x_{\mathrm{c}k}\{\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(\bar{U}_k)v_{1,k}(\theta; x_{k,\xi})\}
$$

Then

$$\begin{aligned}\|\mathbf{A.22}\| = \|\mathbf{A.21}\| &= \| -\frac{1}{n}\sum_{k=1}^{n} C_t^\intercal x_{ck}\{\mathrm{vec}(\alpha\beta^\intercal)^\intercal \mathrm{vec}(\bar{U}_k) v_{1,k}(\theta; x_{k,\xi})\| \\ &\leq \|C_t^\intercal\| \times \|\mathrm{vec}(\alpha\beta^\intercal)\| \times \frac{1}{n}\sum_{k=1}^{n} \|x_{ck}\| \times \|\mathrm{vec}(\bar{U}_k)\|.\end{aligned}$$

By Conditions (C.2) and (C.3), when $\min(n, m) \to \infty$, $\mathrm{H}_{\alpha,\beta,n,3} = o_p(1)$.

Finally, Combining the results of $\mathrm{H}_{\alpha,\beta,n,1}$, $\mathrm{H}_{\alpha,\beta,n,2}$ and $\mathrm{H}_{\alpha,\beta,n,3}$, we conclude that

$$\mathrm{H}^*_{\alpha,\beta,n}(\theta) - \mathrm{H}_{\alpha,\beta,n}(\theta) = o_p(1) \text{ as } \min(m, n) \to \infty.$$

Part II: Show that $H^*_{\alpha\beta,\gamma,n}(\theta) - H_{\alpha\beta,\gamma,n}(\theta) = o_p(1)$ ***as*** $\min(m, n) \to \infty$.

First, we write

$$\mathrm{H}^*_{\alpha\beta,\gamma,n}(\theta) - \mathrm{H}_{\alpha\beta,\gamma,n}(\theta) = \begin{pmatrix}\mathrm{H}_{\alpha,\gamma,n}\\ \mathrm{H}_{\beta,\gamma,n}\end{pmatrix}$$

to indicate the two subvectors corresponding to α and β, where

$$\mathrm{H}_{\alpha,\gamma,n} = -\frac{1}{n}\sum_{k=1}^{n} C_t^\intercal \bar{U}_k \beta v_{1,k}(\theta; X_k^*) + \frac{1}{n}\sum_{k=1}^{n} C_t^\intercal x_{ck} \beta v_{2,k}(\theta; X_k^*) z_k^\intercal$$

and

$$\mathrm{H}_{\beta,\gamma,n} = -\frac{1}{n}\sum_{k=1}^{n} \alpha^\intercal \bar{U}_k v_{1,k}(\theta; X_k^*) + \frac{1}{n}\sum_{k=1}^{n} \alpha^\intercal x_{ck} v_{2,k}(\theta; X_k^*) z_k^\intercal \triangleq \mathbf{A.24} + \mathbf{A.25}.$$

We observe that $\mathrm{H}_{\alpha,\gamma,n}$ and $\mathrm{H}_{\beta,\gamma,n}$ have a similar structure, thus we examine $\mathrm{H}_{\alpha,\gamma,n}$ only here. $\mathrm{H}_{\beta,\gamma,n}$ can be examined via the same techniques.

By Conditions (C.1), (C.3) and the boundedness of $v_{1,k}(\theta; X_k^*)$, we have

$$\begin{aligned}
\|\mathbf{A.22}\| &\leq \frac{1}{n}\sum_{k=1}^{n} \|C_t^\intercal \bar{U}_k^\intercal \beta z_k^\intercal\| \\
&\leq \|\beta\| \times \|C_t\| \times \frac{1}{n}\sum_{k=1}^{n}(\|\bar{U}_k\| \times \|z_k^\intercal\|) \\
&\leq \|\beta\| \times \|C_t\| \times \left(\frac{1}{n}\sum_{k=1}^{n}\|\bar{U}_k\|^2\right)^{1/2} \times \left(\frac{1}{n}\sum_{k=1}^{n}\|z_k^\intercal\|^2\right)^{1/2} \\
&\leq \|\beta\| \times \|C_t\| \times O_p\left(\frac{1}{\sqrt{m}}\right) \times O(1) = o_p(1),
\end{aligned}$$

where the second step and the third step are due to the Cauchy–Schwarz inequality, and the second last step is due to (A.2).

Plugging the (A.4) into **A.25**, we obtain that

$$\mathbf{A.25} = \frac{1}{n}\sum_{k=1}^{n} \text{vec}(\alpha\beta^\intercal)^\intercal \text{vec}(\bar{U}_k) p_k(\theta; x_{k,\xi})\{1 - p_k(\theta; x_{k,\xi})\}^2 C_t^\intercal x_{ck} \beta z_k^\intercal.$$

Then by the boundedness of $p_k(\theta; x_{ck})\{1 - p_k(\theta; x_{ck})\}^2$, we obtain that

$$\begin{aligned}
\|\mathbf{A.25}\| &\leq \mathrm{C}_0 \|\text{vec}(\alpha\beta^\intercal)^\intercal\| \times \|\beta\| \times \|C_t\| \times \frac{1}{n}\sum_{k=1}^{n} \|\text{vec}(\bar{U}_k)\| \times \|x_{ck}\| \times \|z_k^\intercal\| \\
&\leq \mathrm{C}_0 \|\text{vec}(\alpha\beta^\intercal)^\intercal\| \times \|\beta\| \times \|C_t\| \times \frac{1}{n}\sum_{k=1}^{n} \|\text{vec}(\bar{U}_k)\| \times \max(\|x_{ck}\|, \|z_k^\intercal\|)^2,
\end{aligned}$$

where C_0 is a bound of $p_k(\theta; x_{ck})\{1-p_k(\theta; x_{ck})\}^2$, and the first step is due to the Cauchy–Schwarz inequality.

Condition (C.1) implies that $\sum_{k=1}^{n} \max(\|x_{ck}\|, \|z_k^\intercal\|)^2 = O(n)$, and Condition (C.2) implies that $\max_{1\leq k\leq n}\{\max(\|x_{ck}\|, \|z_k^\intercal\|)^2\}$ is $o(n)$. Thus, by Lemma 5.1 of Stefanski and Carroll (1985), **A.25** is $o_p(1)$. Thus, $\mathrm{H}_{\alpha,\gamma,n} = o_p(1)$. Similarly, we can show that $\mathrm{H}_{\beta,\gamma,n} = o_p(1)$. As a result,

$$\mathrm{H}^*_{\alpha\beta,\gamma,n}(\theta) - \mathrm{H}_{\alpha\beta,\gamma,n}(\theta) = o_p(1) \text{ as } \min(m, n) \to \infty.$$

Part III: Show that $H^*_{\gamma,\gamma,n}(\theta) - H_{\gamma,\gamma,n}(\theta) = o_p(1)$ ***as*** $\min(m,n) \to \infty$.

By the definitions of $\mathrm{H}^*_{\gamma,\gamma,n}(\theta)$ and $\mathrm{H}_{\gamma,\gamma,n}(\theta)$, we obtain that

$$\mathrm{H}^*_{\gamma,\gamma,n}(\theta) - \mathrm{H}^*_{\gamma,\gamma,n}(\theta) = -\frac{1}{n}\sum_{k=1}^{n}[v_{1,k}(\theta; x^*_k) - v_{1,k}(\theta; x_{\mathrm{c}k})] z_k z_k^{\intercal}. \tag{A.26}$$

Similar to the steps for obtaining (A.21), by plugging (A.7) into (A.26), the boundedness of $v_{2,k}(\cdot)$, and Conditions (C.1), (C.2) and (C.3) together with Lemma 5.1 in Stefanski and Carroll (1985), we obtain that

$$\mathrm{H}^*_{\gamma,\gamma,n}(\theta) - \mathrm{H}^*_{\gamma,\gamma,n}(\theta) = o_p(1) \text{ as } \min(m,n) \to \infty.$$

Finally, combining the results of Parts I-III, we obtain that

$$\mathrm{H}^*_n(\theta) - \mathrm{H}_n(\theta) = o_p(1) \text{ as } \min(m,n) \to \infty.$$

A.6 Proof of (2.12)

We applied the first-order Taylor series expansion to $\mathrm{S}^*_n(\hat{\theta}^*)$ around θ, with X^*_k and z_k given, and obtain that

$$\mathrm{S}^*_n(\theta) - \mathrm{H}^*_n(\theta)(\hat{\theta}^* - \theta) + e_n = 0,$$

where e_n is the reminder term, given by:

$$e_n = \frac{1}{2!}\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial^2 \mathrm{Z}_n(\theta_m)}{\partial\theta_i \partial\theta_j}(\hat{\theta}^*_i - \theta_i)(\hat{\theta}^*_j - \theta_j) \tag{A.27}$$

with the vector $\theta_m = (\tilde{\alpha}_m^{\intercal}, \beta_m^{\intercal}, \gamma_m^{\intercal})^{\intercal}$ "between" $\hat{\theta}^*$ and θ in the sense that $\|\theta_m\|$ is between $\|\hat{\theta}^*\|$ and $\|\theta\|$, θ_i and θ_j are the ith and jth elements in θ, respectively, and d is the dimension of θ.

Similar to Parts I-III in Appendix A.4 which express $\mathrm{S}_n^*(\theta)$ as the sum of two terms based on $x_{\mathrm{c}k}$ and $\bar{U}_k$, here we apply the same process to obtain that

$$\mathrm{S}_n^*(\theta_m) = \frac{1}{\sqrt{n}}\Big\{\frac{1}{\sqrt{n}}\sum_{k=1}^{n} A_k(\theta_m) + \frac{1}{m\sqrt{n}}\sum_{k=1}^{n}\sum_{r=1}^{m} B_{kr}(\theta_m)\Big\}\{y_k - p_k(\theta_m; X_k^*)\} \tag{A.28}$$

where

$$A_k(\theta_m) = \begin{pmatrix} C_t^{\intercal} x_{\mathrm{c}k}\beta_m \\ x_{\mathrm{c}k}^{\intercal}\alpha_m \\ z_k \end{pmatrix} \text{ and } B_{kr}(\theta_m) = \begin{pmatrix} C_t^{\intercal}(E_{kr} - \frac{1}{n}\sum\limits_{k=1}^{n} E_{kr})\beta_m \\ (E_{kr} - \frac{1}{n}\sum\limits_{k=1}^{n} E_{kr})\alpha_m \\ 0 \end{pmatrix}.$$

Thus, differentiating (A.28) and taking sums gives that

$$\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial^2 \mathrm{S}_n^*(\theta_m)}{\partial\theta_i\theta_j} = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial A_k(\theta_m)\{y_k - p_k(\theta_m; X_k^*)\}}{\partial\theta_i\partial\theta_j} + \frac{1}{mn}\sum_{k=1}^{n}\sum_{r=1}^{m}\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial B_{kr}(\theta_m)\{y_k - p_k(\theta_m; X_k^*)\}}{\partial\theta_i\partial\theta_j}.$$

By Condition (C.6), we obtain that for $i, j = 1, ..., d$,

$$\frac{1}{n}\sum_{k=1}^{n}\frac{\partial A_k(\theta_m)\{y_k - p_k(\theta_m; X_k^*)\}}{\partial\theta_i\partial\theta_j} = O_p(1) = \frac{1}{\sqrt{n}}\mathrm{O}_p(\sqrt{n})$$

and

$$\frac{1}{mn}\sum_{k=1}^{n}\sum_{r=1}^{m}\frac{\partial B_{kr}(\theta_m)\{y_k - p_k(\theta_m; X_k^*)\}}{\partial\theta_i\partial\theta_j} = O_p(1) = \frac{1}{m\sqrt{n}}\mathrm{O}_p(m\sqrt{n}).$$

Thus,

$$\Big|\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial^2 \mathrm{S}_n^*(\theta_m)}{\partial\theta_i\partial\theta_j}\Big| \le \frac{1}{\sqrt{n}}\mathrm{O}_p(\sqrt{n}) + \frac{1}{m\sqrt{n}}\mathrm{O}_p(m\sqrt{n})$$

due to the triangle inequality and fixed d.

Combining this with $\hat{\theta}^* - \theta = o_p(1)$, we obtain, by (A.27), that

$$|e_n| \le o_p\{\max(1/m, 1/n^{1/2})\}.$$

Thus, (2.12) follows.

A.7 Proof of Theorem 2.3

First, we consider the terms of $\hat{\mathrm{J}}_n(\hat{\theta}^*)$ which is defined in (2.15), and show that as $\min(m,n) \to \infty$,

$$\hat{\mathrm{J}}_{\hat{\alpha}^*,n} - \mathrm{J}_{\alpha,n} = o_p\Big(\frac{1}{m}\Big), \tag{A.29}$$

$$\hat{\mathrm{J}}_{\hat{\beta}^*,n} - \mathrm{J}_{\beta,n} = o_p\Big(\frac{1}{m}\Big), \tag{A.30}$$

and

$$\hat{\mathrm{J}}_{\hat{\gamma}^*,n} - \mathrm{J}_{\gamma,n} = o_p\Big(\frac{1}{m}\Big), \tag{A.31}$$

where $\hat{\mathrm{J}}_{\hat{\alpha}^*,n}$, $\hat{\mathrm{J}}_{\hat{\beta}^*,n}$ and $\hat{\mathrm{J}}_{\hat{\gamma}^*,n}$ correspond to $\mathrm{J}_{\alpha,n}$, $\mathrm{J}_{\beta,n}$ and $\mathrm{J}_{\gamma,n}$ with x_{ck}, θ and Ω_0 replaced by X_k^*, $\hat{\theta}^*$ and $\hat{\Omega}$, respectively. Here we show (A.29). The proof of (A.30) and (A.31) is similar.

By definition,

$$\mathrm{J}_{\alpha,n} = \mathrm{J}_{\alpha,n,1} + \mathrm{J}_{\alpha,n,2} \text{ and } \hat{\mathrm{J}}_{\hat{\alpha}^*,n} = \hat{\mathrm{J}}_{\hat{\alpha}^*,n,1} + \hat{\mathrm{J}}_{\hat{\alpha}^*,n,2}.$$

We need only to prove that $\hat{\mathrm{J}}_{\hat{\alpha}^*,n,1} - \mathrm{J}_{\alpha,n,1} = o_p\Big(\frac{1}{m}\Big)$; the proof of $\hat{\mathrm{J}}_{\hat{\alpha}^*,n,2} - \mathrm{J}_{\alpha,n,2} = o_p\Big(\frac{1}{m}\Big)$ carries through in a similar manner.

Indeed,

$$
\begin{aligned}
\hat{\mathrm{J}}_{\hat{\alpha}^*,n,1} - \mathrm{J}_{\alpha,n,1} &= -(1/2n)\sum_{k=1}^{n} C_t^{\intercal} X_k^* \beta \mathrm{vec}(\hat{\alpha}^*\hat{\beta}^{*\intercal})^{\intercal}(\hat{\Omega}/m_c) v_{2,k}(\hat{\theta}^*; x_{\mathrm{c}k}) \\
&\quad - \left\{ -(1/2n)\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{c}k} \beta \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}(\Omega_0/m_c) v_{2,k}(\theta; x_{\mathrm{c}k}) \right\} \\
&= -(1/2n)\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{c}k} \beta \mathrm{vec}(\hat{\alpha}^*\hat{\beta}^{*\intercal})^{\intercal}(\hat{\Omega}/m_c) v_{2,k}(\hat{\theta}^*; x_{\mathrm{c}k}) \\
&\quad - \left\{ -(1/2n)\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{c}k} \beta \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}(\Omega_0/m_c) v_{2,k}(\theta; x_{\mathrm{c}k}) \right\} \\
&\quad - (1/2n)\sum_{k=1}^{n} C_t^{\intercal} \bar{U}_k \beta \mathrm{vec}(\hat{\alpha}^*\hat{\beta}^{*\intercal})^{\intercal}(\hat{\Omega}/m_c) v_{2,k}(\hat{\theta}^*; x_{\mathrm{c}k}) \\
&= -(1/2n)\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{c}k} \beta \mathrm{vec}(\hat{\alpha}^*\hat{\beta}^{*\intercal})^{\intercal}(\hat{\Omega}/m_c) v_{2,k}(\hat{\theta}^*; x_{\mathrm{c}k}) \\
&\quad + (1/2n)\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{c}k} \beta \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}(\Omega_0/m_c) v_{2,k}(\theta; x_{\mathrm{c}k}) + o_p\left(\frac{1}{m}\right) \\
&= \frac{1}{m_c}\Bigg\{ - (1/2n)\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{c}k} \beta \mathrm{vec}(\hat{\alpha}^*\hat{\beta}^{*\intercal})^{\intercal}\hat{\Omega} v_{2,k}(\hat{\theta}^*; x_{\mathrm{c}k}) \\
&\quad + (1/2n)\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{c}k} \beta \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\Omega_0 v_{2,k}(\theta; x_{\mathrm{c}k}) \Bigg\} + o_p\left(\frac{1}{m}\right)
\end{aligned}
\tag{A.32}
$$

where the second step is due to model (2.5); the last term in the third step is of order $o_p\left(\frac{1}{m}\right)$ using the same technique of showing the order of $Q_{\alpha,n,1}$ in Step 1 of Appendix A.4. By Theorem 2.1 and that $\hat{\Omega}$ is a $\sqrt{n}$-consistent estimator of Ω_0, (A.32) yields that $\hat{\mathrm{J}}_{\hat{\alpha}^*,n,1} - \mathrm{J}_{\alpha,n,1} = o_p\left(\frac{1}{m_c}\right) + o_p\left(\frac{1}{m}\right) = o_p\left(\frac{1}{m}\right)$.

Secondly, combining (2.15) with (2.11) gives that

$$\begin{aligned}\hat\theta_c^* &= \hat\theta^* - \hat{\mathrm{H}}_n^{-1}(\hat\theta^*)\hat{\mathrm{J}}_n(\hat\theta^*)\mathrm{vec}(\hat\alpha^*\hat\beta^{*\intercal})\\ &= \theta + \frac{1}{\sqrt{n}}\mathrm{H}_n^{-1}(\theta)\mathrm{S}_n(\theta) + \mathrm{H}_n^{-1}(\theta)\mathrm{J}_n(\theta)\mathrm{vec}(\alpha\beta^\intercal) - \hat{\mathrm{H}}_n^{-1}(\hat\theta^*)\hat{\mathrm{J}}_n(\hat\theta^*)\mathrm{vec}(\hat\alpha^*\hat\beta^{*\intercal})\\ &\quad + o_p\Big\{\max\Big(\frac{1}{m},\frac{1}{\sqrt{n}}\Big)\Big\},\end{aligned} \tag{A.33}$$

where the relevant quantities are defined in Section 2.2.

By Theorem 2.1 and the assumption in Theorem 2.3 that $\hat\Omega$ is a $\sqrt{n}$-consistent estimator of Ω_0, we obtain that by the Continuous Mapping Theorem, $\mathrm{vec}(\hat\alpha^*\hat\beta^{*\intercal}) \to \mathrm{vec}(\alpha\beta^\intercal)$ in probability and $\hat{\mathrm{H}}_n^{-1}(\hat\theta^*) - \mathrm{H}_n^{-1}(\theta) = o_p(1)$ as $\min(m,n)\to\infty$. Therefore, combining these results with (A.29), (A.30) and (A.31), we can express (A.33) as

$$\hat\theta_c^* = \theta + \frac{1}{\sqrt{n}}\mathrm{H}_n^{-1}(\theta)\mathrm{S}_n(\theta) + o_p\Big\{\max\Big(\frac{1}{m},\frac{1}{\sqrt{n}}\Big)\Big\}. \tag{A.34}$$

Finally, Stefanski and Carroll (1985) showed that under Conditions (C.1) and (C.5), $\mathrm{H}_n^{-1/2}(\theta)\mathrm{S}_n(\theta)$ has the asymptotic normal distribution whose mean is zero and covariance matrix is the identity matrix. Thus, by (A.34), we obtain Theorem 2.3(a). Applying Slutsky's theorem to (A.34) gives Theorem 2.3(b), where the asymptotic covariance of $\sqrt{n}(\hat\theta_c^* - \theta)$ is determined by that of $\mathrm{H}_n^{-1}(\theta)\mathrm{S}_n(\theta)$, which equals $I^{-1}(\theta)$,where $I(\theta) = E\{\mathrm{H}_n(\theta)\}$.

A.8 Proof of the Consistency for (2.17)

We prove that (2.17) is a $\sqrt{n}$-consistent covariance estimator for $\hat\Omega$. Indeed, we write (2.17) as

$$\hat\Omega = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{(m-1)}\sum_{r=1}^{m}\{\mathrm{vec}(X_{kr}^*) - \mathrm{vec}(\bar X_{k+}^*)\}\{\mathrm{vec}(X_{kr}^*) - \mathrm{vec}(\bar X_{k+}^*)\}^\intercal,$$

and let

$$\hat\Omega_k = \frac{1}{(m-1)}\sum_{r=1}^{m}\{\mathrm{vec}(X_{kr}^*) - \mathrm{vec}(\bar X_{k+}^*)\}\{\mathrm{vec}(X_{kr}^*) - \mathrm{vec}(\bar X_{k+}^*)\}^\intercal.$$

By the definition of X^*_{kr} and $\bar{X}^*_{k+}$, we obtain that

$$\mathrm{vec}(X^*_{kr}) - \mathrm{vec}(\bar{X}^*_{k+}) = \mathrm{vec}\left(E_{kr} - \frac{1}{m}\sum_{r=1}^{m} E_{kr}\right).$$

Thus,

$$\begin{aligned}
E(\hat{\Omega}_k) &= \frac{1}{(m-1)}\sum_{r=1}^{m} E\left\{\mathrm{vec}\left(E_{kr} - \frac{1}{m}\sum_{r=1}^{m} E_{kr}\right)\mathrm{vec}\left(E_{kr} - \frac{1}{m}\sum_{r=1}^{m} E_{kr}\right)^{\intercal}\right\} \\
&= \frac{1}{(m-1)}\sum_{r=1}^{m}\left[E\left\{\mathrm{vec}\left(E_{kr}\right)\mathrm{vec}\left(E_{kr}\right)^{\intercal}\right\} - 2E\left\{\mathrm{vec}\left(\frac{1}{m}\sum_{r=1}^{m} E_{kr}\right)\mathrm{vec}\left(E_{kr}\right)^{\intercal}\right\}\right. \\
&\quad \left. + E\left\{\mathrm{vec}\left(\frac{1}{m}\sum_{r=1}^{m} E_{kr}\right)\mathrm{vec}\left(\frac{1}{m}\sum_{r=1}^{m} E_{kr}\right)^{\intercal}\right\}\right] \\
&= \frac{1}{(m-1)}\sum_{r=1}^{m}\left[E\left\{\mathrm{vec}\left(E_{kr}\right)\mathrm{vec}\left(E_{kr}\right)^{\intercal}\right\} - \frac{2}{m}E\left\{\mathrm{vec}\left(E_{kr}\right)\mathrm{vec}\left(E_{kr}\right)^{\intercal}\right\}\right. \\
&\quad - \frac{2}{m}\sum_{j\neq r}^{m} E\left\{\mathrm{vec}\left(E_{kj}\right)\mathrm{vec}\left(E_{kr}\right)^{\intercal}\right\} + \frac{1}{m^2}E\left\{\sum_{r=1}^{m}\mathrm{vec}\left(E_{kr}\right)\mathrm{vec}\left(E_{kr}\right)^{\intercal}\right\} \\
&\quad \left. + \frac{1}{m^2}E\left\{\sum_{r=1}^{m}\sum_{j\neq r}^{m}\mathrm{vec}\left(E_{kr}\right)\mathrm{vec}\left(E_{kj}\right)^{\intercal}\right\}\right] \\
&= \frac{1}{(m-1)}\sum_{r=1}^{m}\left(\Omega_0 - \frac{2}{m}\Omega_0 + \frac{1}{m}\Omega_0\right) \\
&= \Omega_0,
\end{aligned} \tag{A.35}$$

where the third step is due to the independence assumption of E_{kr} and E_{kj} for $r \neq j$, together with $E\{\mathrm{vec}(E_{kr})\} = 0$ and the definition of Ω_0.

By Condition (C.3), $var(\hat{\Omega}_k)$ exists. Let $\hat{\sigma}_{k,ij}$ and $\sigma_{0,ij}$ denote the (i,j) element of $\hat{\Omega}_k$ and Ω_0, respectively. Then by (A.35) and the Central limit theorem, as $n \to \infty$, $\sqrt{n}(\frac{1}{n}\sum_{k=1}^{n}\hat{\sigma}_{k,ij} - \sigma_{0,ij})$ converges in distribution to a normal distribution for any (i,j) element for $\hat{\Omega}_k$. Thus, we have $\sqrt{n}(\frac{1}{n}\sum_{k=1}^{n}\hat{\sigma}_{k,ij} - \sigma_{0,ij}) = O_p(1)$ for any (i,j) element for $\hat{\Omega}_k$. Writing these in the matrix form gives that $\frac{1}{n}\sum_{k=1}^{n}\hat{\Omega}_k - \Omega_0 = O_p(1/\sqrt{n})$.

A.9 Derivation of Δ_k

Here we derive why $\Delta_k = X_k^* + (y_k - 1/2)R\alpha\beta^\intercal C/m_c$, discussed in Section 2.3.2, is a sufficient statistic for $x_{\mathrm{c}k}$. The derivations start with working out the joint distribution of Y_k and X_k^* using the model setup in Section 2.2.1 and 2.3.2. Then we work out the joint distribution of Y_k and Δ_k in order to derive the conditional distribution of Y_k given Δ_k and $\{x_{\mathrm{c}k}, z_k\}$. The details are presented in three parts.

Part I: Find $f_{\mathrm{Y,X^*}}(Y_k, X_k^* \mid x_{\mathrm{c}k})$***, the joint distribution of*** Y_k ***and*** X_k^****, given*** $\{x_{\mathrm{c}k}, z_k\}$***.***

We treat $\theta = (\alpha^\intercal, \beta^\intercal, \gamma^\intercal)^\intercal$ as given. We rewrite (2.2) as

$$f_{\mathrm{Y}}(Y_k \mid x_{\mathrm{c}k}, z_k) = h_1(x_{\mathrm{c}k}) \times \exp\{Y_k(\alpha^\intercal x_{\mathrm{c}k}\beta + \gamma^\intercal z_k)\}, \tag{A.36}$$

where $h_1(x_{\mathrm{c}k}, z_k) = \{1 + \exp(\alpha^\intercal x_{\mathrm{c}k}\beta + \gamma^\intercal z_k)\}^{-1}$.

The probability density function (2.5) of X_k^*, given x_{ck} is rewritten as

$$\begin{aligned} f_{\mathrm{X^*}}(X_k^* \mid x_{\mathrm{c}k}) &= \text{constant} \times \exp\Big[-\frac{m_c}{2}\mathrm{tr}\{C^{-1}(X_k^* - x_{\mathrm{c}k})^\intercal R^{-1}(X_k^* - x_{\mathrm{c}k})\}\Big] \\ &= h_2(x_{\mathrm{c}k}) \times \exp\Big\{\mathrm{tr}(m_c C^{-1} X_k^{*\intercal} R^{-1} x_{\mathrm{c}k}) - \frac{m_c}{2}\mathrm{tr}(C^{-1}X_k^{*\intercal}R^{-1}X_k^*)\Big\}, \end{aligned} \tag{A.37}$$

where $h_2(x_{\mathrm{c}k}) = \text{constant} \times \exp\{-\frac{1}{2}\mathrm{tr}(C^{-1}x_{\mathrm{c}k}^\intercal R^{-1} x_{\mathrm{c}k})\}$.

Combining (A.36) and (A.37), we write the joint distribution of Y_k and X_k^*, given $\{x_{\mathrm{c}k}, z_k\}$, as

$$\begin{aligned} f_{\mathrm{Y,X^*}}(Y_k, X_k^* \mid x_{\mathrm{c}k}) = h_3(x_{\mathrm{c}k}, z_k) &\times \exp\{Y_k(\alpha^\intercal x_{\mathrm{c}k}\beta + \gamma^\intercal z_k)\} \\ &\times \exp\Big\{\mathrm{tr}(m_c C^{-1} X_k^{*\intercal} R^{-1} x_{\mathrm{c}k}) - \frac{m_c}{2}\mathrm{tr}(C^{-1}X_k^{*\intercal}R^{-1}X_k^*)\Big\}, \end{aligned} \tag{A.38}$$

where $h_3(x_{\mathrm{c}k}, z_k) = h_1(x_{\mathrm{c}k}, z_k) \times h_2(x_{\mathrm{c}k})$.

Part II: Find $f_{\mathrm{Y},\Delta}(y_k, \Delta_k \mid x_{ck}, z_k)$***, the joint probability density/mass function for*** Y_k ***and*** Δ_k***.***

In the following we want to work out the joint distribution of Y_k and Δ_k, given $\{x_{ck}, z_k\}$, using (A.38). Because Y_k is binary, our discussion here is slightly different from the setting of Stefanski and Carroll (1987, p.4) who applied the variable transformation method to find the joint distribution of *continuous* variables Y_k and Δ_k. We consider the joint cumulative distribution function of $\{Y_k, X_k^*\}$, given $\{x_{ck}, z_k\}$. For any (y, δ),

$$\begin{aligned} pr(Y_k = y, \Delta_k \le \delta \mid \{x_{ck}, z_k\}) &= pr\Big(Y_k = y, X_k^* + \Big(y_k - \frac{1}{2}\Big)R\alpha\beta^{\intercal}C\Big/m_c \le \delta \mid \{x_{ck}, z_k\}\Big) \\ &= pr\Big(Y_k = y, X_k^* \le \delta - \Big(y_k - \frac{1}{2}\Big)R\alpha\beta^{\intercal}C\Big/m_c \mid \{x_{ck}, z_k\}\Big) \\ &= \int_{-\infty}^{\delta-(y_k-\frac{1}{2})R\alpha\beta^{\intercal}C/m_c} f_{\mathrm{Y,X^*}}(y, x_k^*| \, x_{ck}, z_k)dx_k^*, \end{aligned} \tag{A.39}$$

where the integrand is determined by (A.38).

Thus, the joint probability density/mass function for Y_k and Δ_k is given by the derivative of (A.39) with respect to δ. That is,

$$\begin{aligned} f_{\mathrm{Y},\Delta}(y_k, \Delta_k \mid x_{ck}, z_k) &= f_{\mathrm{Y,X^*}}\Big\{y_k, \Delta_k - \Big(y_k - \frac{1}{2}\Big)R\alpha\beta^{\intercal}C\Big/m_c \mid x_{ck}, z_k\Big\} \\ &\quad \times \frac{d}{d\Delta_k}\Big\{\Delta_k - \Big(y_k - \frac{1}{2}\Big)R\alpha\beta^{\intercal}C\Big/m_c\Big\} \\ &= f_{\mathrm{Y,X^*}}\Big\{y_k, \Delta_k - \Big(y_k - \frac{1}{2}\Big)R\alpha\beta^{\intercal}C\Big/m_c \mid x_{ck}, z_k\Big\} \\ &= h_3(x_{ck}, z_k) \times \exp\{Y_k(\alpha^{\intercal}x_{ck}\beta + \gamma^{\intercal}z_k)\} \\ &\quad \times \exp\big[\mathrm{tr}\{C^{-1}(m_c\Delta_k - y_kR\alpha\beta^{\intercal}C + \frac{1}{2}R\alpha\beta^{\intercal}C)^{\intercal}R^{-1}x_{ck}\} \\ &\quad - \frac{1}{2}\mathrm{tr}\{C^{-1}(m_c\Delta_k - y_kR\alpha\beta^{\intercal}C \\ &\quad + \frac{1}{2}R\alpha\beta^{\intercal}C)^{\intercal}R^{-1}(\Delta_k - y_kR\alpha\beta^{\intercal}C/m_c + \frac{1}{2m_c}R\alpha\beta^{\intercal}C)\}\big] \end{aligned} \tag{A.40}$$

where we purposefully use upper case letters in the arguments to emphasize the random variables to which the distribution corresponds, and the last step comes from plugging in (A.38).

Part III: Show that Δ_k can be treated as a sufficient statistic of x_{ck}.

To simplify (A.40), we individually examine each term using the following matrix or vector identities:

$$\begin{aligned}
&\mathrm{tr}(\beta\alpha^\intercal x_{\mathrm{c}k}) = \mathrm{vec}(\alpha\beta^\intercal)^\intercal \mathrm{vec}(x_{\mathrm{c}k}) = \alpha^\intercal x_{\mathrm{c}k}\beta;\\
&\mathrm{tr}(\beta\alpha^T\Delta_k) = \mathrm{vec}(\alpha\beta^\intercal)^\intercal \mathrm{vec}(\Delta_k) = \alpha^\intercal\Delta_k\beta;\\
&\mathrm{tr}(C^{-1}\Delta_k^\intercal\alpha\beta^\intercal C) = \mathrm{tr}(CC^{-1}\Delta_k^\intercal\alpha\beta^\intercal) = \mathrm{tr}(\Delta_k^\intercal\alpha\beta^\intercal) = \alpha^\intercal\Delta_k\beta;
\end{aligned}$$

which are obtained from direct calculations and the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for two square matrice A and B which have the same dimension.

The first term of (A.40) is simplified as

$$\begin{aligned}
&\mathrm{tr}[C^{-1}(m_c\Delta_k - Y_kR\alpha\beta^\intercal C)^\intercal R^{-1}x_{\mathrm{c}k}] + \frac{1}{2}\mathrm{tr}\{C^{-1}(R\alpha\beta^\intercal C)^\intercal R^{-1}x_{\mathrm{c}k}\}\\
&= \mathrm{tr}(m_cC^{-1}\Delta_k^\intercal R^{-1}x_{\mathrm{c}k} - Y_k\beta\alpha^\intercal x_{\mathrm{c}k}) + \frac{1}{2}\mathrm{tr}(\beta\alpha^\intercal x_{\mathrm{c}k})\\
&= \mathrm{tr}(m_cC^{-1}\Delta_k^\intercal R^{-1}x_{\mathrm{c}k}) - Y_k\alpha^\intercal x_{\mathrm{c}k}\beta + \frac{1}{2}\alpha^\intercal x_{\mathrm{c}k}\beta,
\end{aligned} \tag{A.41}$$

and the second term of (A.40) becomes

$$\begin{aligned}
&\frac{1}{2}\mathrm{tr}\{C^{-1}(m_c\Delta_k - Y_kR\alpha\beta^\intercal C + \frac{1}{2}R\alpha\beta^\intercal C)^\intercal R^{-1}(\Delta_k - Y_kR\alpha\beta^\intercal C/m_c + \frac{1}{2m_c}R\alpha\beta^\intercal C)\}\\
&= \frac{1}{2}\mathrm{tr}[C^{-1}\{m_c\Delta_k^\intercal - Y_k(C\beta\alpha^\intercal R) + \frac{1}{2}(C\beta\alpha^\intercal R)\}R^{-1}(\Delta_k - Y_kR\alpha\beta^\intercal C/m_c + \frac{1}{2m_c}R\alpha\beta^\intercal C)]\\
&= \frac{1}{2}\mathrm{tr}\{(m_cC^{-1}\Delta_k^\intercal R^{-1} - Y_k\beta\alpha^\intercal + \frac{1}{2}\beta\alpha^\intercal)(\Delta_k - Y_kR\alpha\beta^\intercal C/m_c + \frac{1}{2m_c}R\alpha\beta^\intercal C)\}\\
&= \frac{1}{2}\mathrm{tr}(m_cC^{-1}\Delta_k^\intercal R^{-1}\Delta_k - Y_k\beta\alpha^\intercal\Delta_k - Y_kC^{-1}\Delta_k^\intercal\alpha\beta^\intercal C\\
&\quad + \frac{1}{2}\beta\alpha^\intercal\Delta_k^\intercal + \frac{1}{4m_c}\beta\alpha^\intercal R\alpha\beta^\intercal C + \frac{1}{2}C^{-1}\Delta_k^\intercal\alpha\beta^\intercal C)\\
&= \frac{1}{2}\mathrm{tr}(m_cC^{-1}\Delta_k^\intercal R^{-1}\Delta_k + \frac{1}{4m_c}\beta\alpha^\intercal R\alpha\beta^\intercal C) - Y_k\alpha^\intercal\Delta_k\beta + \frac{1}{2}\alpha^\intercal\Delta_k\beta,
\end{aligned} \tag{A.42}$$

where we use the fact that $Y_k^2 = Y_k$ for the binary variable Y_k taking value 0 or 1.

Then plugging (A.41) and (A.42) into (A.40) gives

$$\begin{aligned}
f_{\mathrm{Y},\Delta}(Y_k, \Delta_k \mid x_{\mathrm{c}k}, z_k) &= h_3(x_{\mathrm{c}k}, z_k) \times \exp\Big\{ -\frac{1}{2}\mathrm{tr}(m_c C^{-1}\Delta_k^{\intercal} R^{-1}\Delta_k + \frac{1}{4m_c}\beta\alpha^{\intercal} R\alpha\beta^{\intercal} C) \\
&\quad + Y_k\alpha^{\intercal}\Delta_k\beta - \frac{1}{2}\alpha^{\intercal}\Delta_k\beta + \frac{1}{2}\alpha^{\intercal} x_{\mathrm{c}k}\beta + \mathrm{tr}(m_c C^{-1}\Delta_k^{\intercal} R^{-1} x_{\mathrm{c}k}) \\
&\quad - Y_k\alpha^{\intercal} x_{\mathrm{c}k}\beta + Y_k\alpha^{\intercal} x_{\mathrm{c}k}\beta\Big\} \times \exp(Y_k\gamma^{\intercal} z_k) \\
&= h_4(\Delta_k, x_{\mathrm{c}k}, z_k) \times \exp(Y_k\alpha^{\intercal}\Delta_k\beta) \times \exp(Y_k\gamma^{\intercal} z_k) \\
&\quad \times \exp\Big\{ -\frac{1}{2}\mathrm{tr}(m_c C^{-1}\Delta_k^{\intercal} R^{-1}\Delta_k\beta + \frac{1}{4m_c}\beta\alpha^{\intercal} R\alpha\beta^{\intercal} C) \\
&\quad + \mathrm{tr}(m_c C^{-1}\Delta_k^{\intercal} R^{-1} x_{\mathrm{c}k})\Big\} \\
&= h_4(\Delta_k, x_{\mathrm{c}k}, z_k) \times \exp\big\{Y_k(\alpha^{\intercal}\Delta_k\beta + \gamma^{\intercal} z_k)\big\},
\end{aligned} \tag{A.43}$$

where $h_4(\Delta_k, x_{ck}, z_k) = h_3(\Delta_k, x_{ck}, z_k) \times \exp\{-\frac{1}{2}\mathrm{tr}(m_c C^{-1}\Delta_k^{\intercal} R^{-1}\Delta_k\beta + \frac{1}{4m_c}\beta\alpha^{\intercal} R\alpha\beta^{\intercal} C) + m_c C^{-1}\Delta_k^{\intercal} R^{-1} x_{ck} - \frac{1}{2}\alpha^{\intercal}\Delta_k\beta + \frac{1}{2}\alpha^{\intercal} x_{ck}\beta\}$.

Noting that (A.43) can be expressed as the product of two functions each involving Y_k or Δ_k alone, together with other variables, we obtain the conditional distribution of Y_k given Δ_k as well as $\{x_{\mathrm{c}k}, z_k\}$,

$$pr(Y_k = y_k \mid \Delta_k, x_{\mathrm{c}k}, z_k) = \mathfrak{C} \times \exp\{y_k(\alpha^{\intercal}\Delta_k\beta + \gamma^{\intercal} z_k)\}, \tag{A.44}$$

where

$$\mathfrak{C} = \frac{1}{h_4(\Delta_k, x_{\mathrm{c}k}, z_k)\{1 + \exp(\alpha^{\intercal}\Delta_k\beta + \gamma^{\intercal} z_k)\}}.$$

Non-involvement of x_{ck} in the right-hand side of (A.40) shows that Δ_k can be treated as a sufficient statistic of $x_{\mathrm{c}k}$.

Remark:

The preceding derivations basically focus on verifying the "sufficiency" for Δ_k which is given before hand. In contrast, a simple way to find a sufficient statistic of x_{ck} by directly applying the Factorization Theorem to the joint distribution (A.38). Specifically, by treating x_{ck} as an unknown parameter and θ as a given constant, we write (A.38) as

$$
\begin{aligned}
f_{Y,X^*}(Y_k, X_k^*|x_{ck}) &\propto \exp\{Y_k(\alpha^\intercal x_{ck}\beta)\} \times \exp\{\mathrm{tr}(m_c C^{-1} X_k^{*\intercal} R^{-1} x_{ck}\} \\
&= \exp[\mathrm{tr}\{(Y_k \alpha\beta^\intercal + m_c R^{-1} X_k^* C^{-1})^\intercal x_{ck}\}] \\
&= \exp\left[\mathrm{tr}\left[\left\{R^{-1}\left(\frac{1}{m_c} Y_k R\alpha\beta^\intercal C + X_k^*\right) C^{-1}\right\}^\intercal\right] x_{ck}\right],
\end{aligned}
$$

implying that $\frac{1}{m_c} Y_k R\alpha\beta^\intercal C + X_k^*$ is a sufficient statistic for x_{ck}. Write $\Delta_k^* = \frac{1}{m_c} Y_k R\alpha\beta^\intercal C + X_k^*$.

Note that the difference between Δ_k^* and Δ_k is a constant, indicating the equivalence of them. While Δ_k^* may be used in the same way as Δ_k in Section 2.3.2 to derive a consistent estimator of θ, in our development we choose to use $\Delta_k = X_k^* + (Y_k - 1/2)R\alpha\beta^\intercal C/m_c$; this quantity can be regarded as "symmetric" around the surrogated measurement X_k^* because the centered version $Y_k - 1/2$ for the binary variable Y_k is either $1/2$ or $-1/2$. In addition, Δ_k shares similarity to the sufficient statistics considered by Stefanski and Carroll (1985).

A.10 Proof of Theorem 2.4

Corresponding to $\mathrm{S}_n^*(\theta)$ and $\mathrm{H}_n^*(\theta)$ in (2.12), we let $\mathrm{S}_n(\theta, \hat{\Delta}_k) = (\mathrm{S}^\intercal_{\hat{\alpha}_s^*,n}(\theta, \hat{\Delta}_k), \mathrm{S}^\intercal_{\hat{\beta}_s^*,n}(\theta, \hat{\Delta}_k), \mathrm{S}^\intercal_{\hat{\gamma}_s^*,n}(\theta, \hat{\Delta}_k))^\intercal$ and $\mathrm{H}_{n,s}(\theta, \hat{\Delta}_k)$ be $\mathrm{S}_n^*(\theta)$ and $\mathrm{H}_n^*(\theta)$, respectively, with X_k^* replaced by $\hat{\Delta}_k$, where $\hat{\Delta}_k$ is defined by (2.22). To show Theorem 2.4, we examine $\mathrm{S}_n(\theta, \hat{\Delta}_k)$ and $\mathrm{H}_{n,s}(\theta, \hat{\Delta}_k)$ separately in the following two parts using similar techniques to those in Appendix A.4 and A.5.

Part I: Show that $S_n(\theta, \hat{\Delta}_k) = \frac{1}{\sqrt{n}} S_n(\theta) + o_p\{\max\left(\frac{1}{m}, \frac{1}{\sqrt{n}}\right)\}$.

Since $\mathrm{S}_{\hat{\alpha}_s^*,n}(\theta, \hat{\Delta}_k)$, $\mathrm{S}_{\hat{\beta}_s^*,n}(\theta, \hat{\Delta}_k)$ and $\mathrm{S}_{\hat{\gamma}_s^*,n}(\theta, \hat{\Delta}_k)$ have similar structures, here we provide only the examination of $\mathrm{S}_{\hat{\alpha}_s^*,n}(\theta, \hat{\Delta}_k)$; the rest two terms can be shown similarly.

By the definition of $\mathrm{S}_n(\theta, \hat{\Delta}_k)$ and (2.9) as well as (2.22), we have that

$$
\begin{aligned}
\mathrm{S}_{\hat{\alpha}_s^*,n}(\theta,\hat{\Delta}_k) &= \frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}\hat{\Delta}_k\beta\{Y_k - p_k(\theta;\hat{\Delta}_k)\} \\
&= \frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}\Big(X_k^* + g_k/m_c\Big)\beta\{Y_k - p_k(\theta;\hat{\Delta}_k)\} \\
&= \frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}\Big(X_k^* + g_k/m_c\Big)\beta\{Y_k - p_k(\theta;X_k^*) + p_k(\theta;X_k^*) - p_k(\theta;\hat{\Delta}_k)\} \\
&\triangleq \mathrm{S}_{\alpha,n}^*(\theta) + \mathrm{W}_{n1} + \mathrm{W}_{n2} + \mathrm{W}_{n3},
\end{aligned}
\tag{A.45}
$$

where by (2.5), we set

$$
\begin{aligned}
\mathrm{W}_{n1} &= \frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}(\bar{U}_k + g_k/m_c)\beta\{p_k(\theta;X_k^*) - p_k(\theta;\hat{\Delta}_k)\}, \\
\mathrm{W}_{n2} &= \frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal} g_k\beta\{Y_k - p_k(\theta;X_k^*)\}/m_c,
\end{aligned}
$$

and

$$
\mathrm{W}_{n3} = \frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal} x_{ck}\beta\{p_k(\theta;X_k^*) - p_k(\theta;\hat{\Delta}_k)\}.
$$

To examine $\mathrm{S}_{\hat{\alpha}_s^*,n}(\theta,\hat{\Delta}_k)$, it sufficies to check W_{n1}, W_{n2} and W_{n3} individually. In the following, we examine W_{n1}, W_{n2} and W_{n3} separately and show they are of order $o_p(1/m)$. Before doing so, we introduce two expressions.

Replacing X_k^* with x_{ck} in W_{n2}, we define

$$
\mathrm{b}_2 = \frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}\{Y_k - p_k(\theta;x_{ck})\}g_k\beta/m_c,
$$

and we define

$$
\mathrm{b}_3 = -\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal} x_{ck}\beta\mathrm{vec}(g_k)^{\intercal}\mathrm{vec}(\alpha\beta^{\intercal})v_{1,k}(\theta;x_{ck})/m_c.
$$

By the fact that $\hat{\theta}^* - \theta = o_p(1)$, and $n^{1/2}(\hat{C} \otimes \hat{R} - \Omega_0) = O_p(1)$, we obtain that

$$\begin{aligned}
\mathrm{b}_2 &= \frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}\Big[\Big(y_k - \frac{1}{2}\Big)\{Y_k - p_k(\theta; x_{ck})\}\Big] R\alpha\beta^{\intercal}C\beta/m_c + o_p(1) \\
&= \frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}\Big[\Big(y_k - \frac{1}{2}\Big)\{Y_k - p_k(\theta; x_{ck})\}\Big] \Pi_\alpha\Omega_0\mathrm{vec}(\alpha\beta^{\intercal})/m_c + o_p(1),
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{b}_3 &= -\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}x_{ck}\beta\Big(y_k - \frac{1}{2}\Big)\mathrm{vec}(R\alpha\beta^{\intercal}C)^{\intercal}\mathrm{vec}(\alpha\beta^{\intercal})v_{1,k}(\theta; x_{ck})/m_c + o_p(1) \\
&= -\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}x_{ck}\beta\Big(y_k - \frac{1}{2}\Big)\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\Omega_0\mathrm{vec}(\alpha\beta^{\intercal})v_{1,k}(\theta; x_{ck})/m_c + o_p(1),
\end{aligned}$$

where Π_α is defined in (A.14).

Furthermore, using $\mathrm{J}_{\alpha,n,1}$ and $\mathrm{J}_{\alpha,n,2}$, which are defined in Section 2.2.3, with b_2 and b_3, we obtain that

$$\begin{aligned}
\mathrm{b}_2 = &\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}\Pi_\alpha\frac{\Omega_0}{m_c}\Big[\Big(y_k - \frac{1}{2}\Big)\{Y_k - p_k(\theta; x_{ck})\}\Big]\mathrm{vec}(\alpha\beta^{\intercal}) - \mathrm{J}_{\alpha,n,2}\mathrm{vec}(\alpha\beta^{\intercal}) \\
&+ \mathrm{J}_{\alpha,n,2}\mathrm{vec}(\alpha\beta^{\intercal}) + o_p(1) \\
= &\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}\Pi_\alpha\frac{\Omega_0}{m_c}\Big\{Y_k^2 - Y_k p_k(\theta; x_{ck}) - \frac{1}{2}Y_k + \frac{1}{2}p_k(\theta; x_{ck})\Big\}\mathrm{vec}(\alpha\beta^{\intercal}) \\
&- \mathrm{J}_{\alpha,n,2}\mathrm{vec}(\alpha\beta^{\intercal}) + \mathrm{J}_{\alpha,n,2}\mathrm{vec}(\alpha\beta^{\intercal}) + o_p(1) \\
= &\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}\Pi_\alpha\frac{\Omega_0}{m_c}\Big[Y_k\{1 - p_k(\theta; x_{ck})\} - \frac{1}{2}\{Y_k - p_k(\theta; x_{ck})\} \\
&- v_{1,k}(\theta; x_{ck})\Big]\mathrm{vec}(\alpha\beta^{\intercal}) - \mathrm{J}_{\alpha,n,2}\mathrm{vec}(\alpha\beta^{\intercal}) + o_p(1) \\
= &-\mathrm{J}_{\alpha,n,2}\mathrm{vec}(\alpha\beta^{\intercal}) + o_p(1)
\end{aligned} \tag{A.46}$$

where the last step is due to $E(Y_k|x_{ck}, z_k) = p_k(\theta; x_{ck})$, and

$$
\begin{aligned}
\mathrm{b}_3 &= -\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{c}k}\beta \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\frac{\Omega_0}{m_c}\mathrm{vec}(\alpha\beta^{\intercal})\Big(y_k - \frac{1}{2}\Big)v_{1,k}(\theta; x_{\mathrm{c}k}) \\
&\quad + \mathrm{J}_{\alpha,n,1}\mathrm{vec}(\alpha\beta^{\intercal}) - \mathrm{J}_{\alpha,n,1}\mathrm{vec}(\alpha\beta^{\intercal}) + o_p(1) \\
&= -\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{c}k}\beta \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\frac{\Omega_0}{m_c}\mathrm{vec}(\alpha\beta^{\intercal})\Big(y_k - \frac{1}{2}\Big)v_{1,k}(\theta; x_{\mathrm{c}k}) \\
&\quad - \frac{n}{2}\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{c}k}\beta \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}(\Omega_0/m_c)v_{2,k}(\theta; x_{ck}) - \mathrm{J}_{\alpha,n,1}\mathrm{vec}(\alpha\beta^{\intercal}) + o_p(1) \\
&= -\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{c}k}\beta \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\frac{\Omega_0}{m_c}\mathrm{vec}(\alpha\beta^{\intercal})\Big[\Big(y_k - \frac{1}{2}\Big)v_{1,k}(\theta; x_{ck}) \\
&\quad - v_{1,k}(\theta; x_{\mathrm{c}k})\Big\{p_k(\theta; x_{ck}) - \frac{1}{2}\Big\}\Big] - \mathrm{J}_{\alpha,n,1}\mathrm{vec}(\alpha\beta^{\intercal}) + o_p(1) \\
&= -\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal} x_{\mathrm{c}k}\beta \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\frac{\Omega_0}{m_c}\mathrm{vec}(\alpha\beta^{\intercal})\Big\{Y_k - p_k(\theta; x_{\mathrm{c}k})\Big\}v_{1,k}(\theta; x_{\mathrm{c}k}) \\
&\quad - \mathrm{J}_{\alpha,n,1}\mathrm{vec}(\alpha\beta^{\intercal}) + o_p(1) \\
&= -\mathrm{J}_{\alpha,n,1}\mathrm{vec}(\alpha\beta^{\intercal}) + o_p(1),
\end{aligned}
\tag{A.47}
$$

where the third step is because of the definition $v_{1,k}(\cdot)$, given in Section 2.2.3, and the last step is due to $E(Y_k|x_{\mathrm{c}k}, z_k) = p_k(\theta; x_{\mathrm{c}k})$.

Now we examine W_{n2}, W_{n3} and W_{n1} by the following three steps:

<u>Step1:</u> Show that $\| W_{n2} - b_2 \| = o_p\big(\frac{1}{m}\big)$.

$$
\begin{aligned}
\|\mathrm{W}_{n2} - b_2\| &= \Big\| \frac{1}{n} \sum_{k=1}^{n} C_t^{\intercal} g_k \{ p_k(\theta; x_{\mathrm{c}k}) - p_k(\theta; X_k^*) \} / m_c \Big\| \\
&\leq \frac{1}{m_c n} \|C_t^{\intercal}\| \sum_{k=1}^{n} \big\| p_k(\theta; x_{\mathrm{c}k}) - p_k(\theta; X_k^*) \big\| \times \|g_k\| \\
&\leq \frac{1}{m_c n} \|C_t^{\intercal}\| \sum_{k=1}^{n} \big| p_k(\theta; x_{\mathrm{c}k}) - p_k(\theta; X_k^*) \big| \times \|g_k\| \\
&\leq \|C_t^{\intercal}\| \times \frac{1}{m_c n} \sum_{k=1}^{n} \|g_k\| \\
&\leq \frac{1}{m_c} \|C_t^{\intercal}\| \times \Big(\frac{1}{n} \sum_{k=1}^{n} \|g_k\|^2 \Big)^{1/2} \\
&= o_p\Big(\frac{1}{m}\Big),
\end{aligned}
$$

where the second step is due to the Cauchy–Schwarz inequality; and the fourth step is because that between $p_k(\theta; x_{\mathrm{c}k})$ and $p_k(\theta; X_k^*)$ the absolute value of the difference is bounded between $[0, 1]$; and the fifth step is due to the Cauchy–Schwarz inequality; and the last step is due to the assumption that $\sum_{k=1}^{n} \|g_k\|^2 = O_p(n)$, and the definition $m_c = \frac{nm}{n-1}$.

Step2: ***Show that*** $\| W_{n3} - b_3 \| = o_p\left(\frac{1}{m}\right)$.

$$
\begin{aligned}
\|\mathrm{W}_{n3} - b_3\| &= \bigg\| \frac{1}{n} \sum_{k=1}^{n} C_t^\intercal x_{ck} \beta \Big[\{ p_k(\theta; X_k^*) - p_k(\theta; \hat{\Delta}_k) \} \\
&\quad + \mathrm{vec}(g_k)^\intercal \mathrm{vec}(\alpha\beta^\intercal) v_{1,k}(\theta; x_{ck})/m_c \Big] \bigg\| \\
&\le \|C_t^\intercal\| \times \|\beta\| \times \frac{1}{n} \sum_{k=1}^{n} \|x_{ck}\| \times \|p_k(\theta; X_k^*) - p_k(\theta; \hat{\Delta}_k)\| \\
&\quad + \|C_t^\intercal\| \times \|\beta\| \times \frac{1}{n} \sum_{k=1}^{n} \|x_{ck}\| \times \Big\| \mathrm{vec}(g_k)^\intercal \mathrm{vec}(\alpha\beta^\intercal) v_{1,k}(\theta; x_{ck})/m_c \Big\| \\
&\le \|C_t^\intercal\| \times \|\beta\| \times \frac{1}{n} \sum_{k=1}^{n} \|x_{ck}\| \times \|p_k(\theta; X_k^*) - p_k(\theta; \hat{\Delta}_k)\| \\
&\quad + \|C_t^\intercal\| \times \|\beta\| \times \frac{1}{n} \sum_{k=1}^{n} \|x_{ck}\| \times \Big\| \mathrm{vec}(g_k)^\intercal \mathrm{vec}(\alpha\beta^\intercal)/m_c \Big\|,
\end{aligned}
\tag{A.48}
$$

where the last second step is due to the Cauchy–Schwarz inequality, and the last step by that $v_{1,k}(\cdot)$ is bounded by 0 and 1.

Then plugging (A.6) into (A.48), we obtain that

$$
\begin{aligned}
(\text{A.48}) &\le 2\|C_t^\intercal\| \times \|\beta\| \times \|\mathrm{vec}(\alpha\beta^\intercal)\| \times \frac{1}{n} \sum_{k=1}^{n} \{ \|x_{ck}\| \times \|\mathrm{vec}(\bar{U}_k)\| \} + \|x_{ck}\| \times \|\mathrm{vec}(g_k)\|/m_c \} \\
&= 2\|C_t^\intercal\| \times \|\beta\| \times \|\mathrm{vec}(\alpha\beta^\intercal)\| \times \frac{1}{n} \sum_{k=1}^{n} \{ \|x_{ck}\| \times \|\mathrm{vec}(\bar{U}_k)\| \} \\
&\quad + 2\|C_t^\intercal\| \times \|\beta\| \times \|\mathrm{vec}(\alpha\beta^\intercal)\| \times \frac{1}{n} \sum_{k=1}^{n} \big\{ \|x_{ck}\| \times \|\mathrm{vec}(g_k)\|/m_c \big\}.
\end{aligned}
\tag{A.49}
$$

By Conditions (C.2) and (C.3), the first term of (A.49) is $o_p\left(\frac{1}{m}\right)$; by Condition (C.2), definition of m_c, and the assumption $\sum_{k=1}^{n} \|g_k\|^2 = O_p(n)$, the second term of (A.49) is $o_p\left(\frac{1}{m}\right)$. As a result, we obtain that $\|\mathrm{W}_{n3} - b_3\| = o_p\left(\frac{1}{m}\right)$.

Step3: ***Show that*** $W_{n1} = o_p\big(\frac{1}{m}\big)$.

$$
\begin{aligned}
\|\mathrm{W}_{n1}\| &= \Big\|\frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}\Big(\bar{U}_k + \frac{g_k}{m_c}\Big)\beta\{p_k(\theta; X_k^*) - p_k(\theta; \hat{\Delta}_k)\}\Big\| \\
&\le \|C_t^{\intercal}\| \times \|\beta\| \times \Big\|\frac{1}{n}\sum_{k=1}^{n}\Big(\bar{U}_k + \frac{g_k}{m_c}\Big) \times \{p_k(\theta; X_k^*) - p_k(\theta; \hat{\Delta}_k)\}\Big\| \\
&= \|C_t^{\intercal}\| \times \|\beta\| \times \Big\|\frac{1}{n}\sum_{k=1}^{n}\Big(\bar{U}_k + \frac{g_k}{m_c}\Big) \times \Big(\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(\bar{U}_k)\Big[v_{1,k}(\theta; x_{k,\xi}) - v_{1,k}(\theta; \Delta_{k,\xi})\Big] \\
&\quad - \mathrm{vec}(g_k)^{\intercal}\mathrm{vec}(\alpha\beta^{\intercal})/m_c\Big)\Big\| \\
&\le \|C_t^{\intercal}\| \times \|\beta\| \times \Big\|\frac{1}{n}\sum_{k=1}^{n}\bar{U}_k\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(\bar{U}_k) \times \Big[v_{1,k}(\theta; x_{k,\xi}) - v_{1,k}(\theta; \Delta_{k,\xi})\Big]\Big\| \\
&\quad + \frac{\|C_t^{\intercal}\| \times \|\beta\| \times \|\mathrm{vec}(\alpha\beta^{\intercal})\|}{nm_c} \times \sum_{k=1}^{n}\Big\{\Big\|\mathrm{vec}(\bar{U}_k)\Big\| \times \Big\|\mathrm{vec}(g_k)\Big\| \\
&\quad \times \Big\|v_{1,k}(\theta; x_{k,\xi}) - v_{1,k}(\theta; \Delta_{k,\xi})\Big\|\Big\} + \|C_t^{\intercal}\| \times \|\beta\| \times \|\mathrm{vec}(\alpha\beta^{\intercal})\| \\
&\quad \times \frac{1}{n}\sum_{k=1}^{n}\left\{\frac{\Big\|\mathrm{vec}(g_k)\Big\|^2}{m_c^2} + \Big\|\mathrm{vec}(\bar{U}_k)\Big\| \times \frac{\Big\|\mathrm{vec}(g_k)\Big\|}{m_c}\right\} \\
&\le \|C_t^{\intercal}\| \times \|\beta\| \times \|\mathrm{vec}(\alpha\beta^{\intercal})\| \times \max\nolimits^2_{1\le k\le n}\|\bar{U}_k\| \times \Big\|\frac{1}{n}\sum_{k=1}^{n} v_{1,k}(\theta; x_{k,\xi}) - v_{1,k}(\theta; \Delta_{k,\xi})\Big\| \\
&\quad + \|C_t^{\intercal}\| \times \|\beta\| \times \|\mathrm{vec}(\alpha\beta^{\intercal})\| \times \frac{1}{n}\sum_{k=1}^{n}\left\{\frac{\Big\|\mathrm{vec}(g_k)\Big\|^2}{m_c^2} + 2\Big\|\mathrm{vec}(\bar{U}_k)\Big\| \times \frac{\Big\|\mathrm{vec}(g_k)\Big\|}{m_c}\right\} \\
&= o_p\Big(\frac{1}{m}\Big),
\end{aligned}
\tag{A.50}
$$

where the second step and the fourth step are due to the Cauchy–Schwarz inequality, the third step is due to plug in the difference between (A.4) and (A.5); in fifth step, we apply the facts that $\max_{1\le k\le n}\|\bar{U}_k\| = O_p(\frac{1}{\sqrt{m}})$ and $\frac{1}{n}\sum_{k=1}^{n} v_{1,k}(\theta; x_{k,\xi}) - v_{1,k}(\theta; \Delta_{k,\xi}) = o_p(1)$; in last step using (A.2) and $\sum_{k=1}^{n}\|g_k\|^2 = O_p(n)$ again, we obtain that (A.47) $= o_p\big(\frac{1}{m}\big)$.

Thus, using the results in Steps 1-3 and by (A.45), we obtain that

$$\begin{aligned}
\mathrm{S}_{\hat{\alpha}_s^*,n}(\theta,\hat{\Delta}_k) &= \mathrm{S}^*_{\alpha,n}(\theta) + b_2 + b_3 + o_p\Big\{\max\Big(\frac{1}{m},\frac{1}{\sqrt{n}}\Big)\Big\} \\
&= \frac{1}{\sqrt{n}}\mathrm{S}_{\alpha,n} + (\mathrm{J}_{\alpha,n,1}+\mathrm{J}_{\alpha,n,2})\mathrm{vec}(\alpha\beta^{\intercal}) + b_2 + b_3 + o_p\Big\{\max\Big(\frac{1}{m},\frac{1}{\sqrt{n}}\Big)\Big\} \\
&= \frac{1}{\sqrt{n}}\mathrm{S}_{\alpha,n} + o_p\Big\{\max\Big(\frac{1}{m},\frac{1}{\sqrt{n}}\Big)\Big\},
\end{aligned}$$

where in the second step, we plug in (A.16) to instead of $\mathrm{S}^*_{\alpha,n}(\theta)$ directly, and third step is due to (A.45) and (A.47).

Part II: Show that $H_{n,s}(\theta,\hat{\Delta}_k) - H_n(\theta) = o_p(\mathit{1})$.

Similar to the treatment of $\mathrm{H}^*_n(\theta)$ in Appendix A.5, we write $\mathrm{H}_{n,s}(\theta,\hat{\Delta}_k)$ as

$$\mathrm{H}_{n,s}(\theta,\hat{\Delta}_k) \triangleq -\begin{pmatrix} \mathrm{H}_{\alpha,\beta,n,s}(\theta,\hat{\Delta}_k) & \mathrm{H}_{\alpha\beta,\gamma,n,s}(\theta,\hat{\Delta}_k) \\ \mathrm{H}^{\intercal}_{\alpha\beta,\gamma,n,s}(\theta,\hat{\Delta}_k) & \mathrm{H}_{\gamma,\gamma,n,s}(\theta,\hat{\Delta}_k) \end{pmatrix}$$

so that each element is identical to the corresponding part of $\mathrm{H}^*_n(\theta)$ with X^*_k replaced by $\hat{\Delta}_k$.

To show that $\mathrm{H}_{n,s}(\theta,\hat{\Delta}_k) - \mathrm{H}_n(\theta) = o_p(1)$, it suffices to show that

$$\mathrm{H}_{\alpha,\beta,n,s}(\theta,\hat{\Delta}_k) - \mathrm{H}_{\alpha,\beta,n}(\theta) = o_p(1), \tag{A.51}$$

as well as $\mathrm{H}_{\alpha\beta,\gamma,n,s}(\theta,\hat{\Delta}_k) - \mathrm{H}_{\alpha\beta,\gamma,n,s}(\theta) = o_p(1)$ and $\mathrm{H}_{\gamma,\gamma,n,s}(\theta,\hat{\Delta}_k) - \mathrm{H}_{\gamma,\gamma,n,s}(\theta) = o_p(1)$. Here we show only (A.51) using same techniques in Appendix A.6; the other two expressions can be shown similarly.

Let

$$\mathrm{H}_{n,\hat{\alpha}_s^*}(\theta,\hat{\Delta}_k) = \frac{\partial\{\mathrm{S}_{\hat{\alpha}_s^*,n}(\theta,\hat{\Delta}_k),\mathrm{S}_{\hat{\beta}_s^*,n}(\theta,\hat{\Delta}_k)\}^{\intercal}}{\partial\tilde{\alpha}^{\intercal}}$$

and

$$\mathrm{H}_{n,\hat{\beta}_s^*}(\theta,\hat{\Delta}_k) = \frac{\partial\{\mathrm{S}_{\hat{\alpha}_s^*,n}(\theta,\hat{\Delta}_k),\mathrm{S}_{\hat{\beta}_s^*,n}(\theta,\hat{\Delta}_k)\}^{\intercal}}{\partial\beta^{\intercal}},$$

and let

$$\mathrm{H}_{n,\alpha}(\theta, x_{\mathrm{c}k}) = \frac{\partial\{\mathrm{Z}_{\alpha,n}(\theta), \mathrm{Z}_{\beta,n}(\theta)\}^{\intercal}}{\partial\tilde{\alpha}^{\intercal}}$$

and

$$\mathrm{H}_{n,\beta}(\theta, x_{\mathrm{c}k}) = \frac{\partial\{\mathrm{Z}_{\alpha,n}(\theta), \mathrm{Z}_{\beta,n}(\theta)\}^{\intercal}}{\partial\beta^{\intercal}}.$$

Then

$$\mathrm{H}_{\alpha,\beta,n,s}(\theta, \hat{\Delta}_k) = \Big(\mathrm{H}_{n,\hat{\alpha}_s^*}(\theta, \hat{\Delta}_k) \quad \mathrm{H}_{n,\hat{\beta}_s^*}(\theta, \hat{\Delta}_k)\Big)$$

and

$$\mathrm{H}_{\alpha,\beta,n}(\theta, x_{\mathrm{c}k}) = \Big(\mathrm{H}_{n,\alpha}(\theta, x_{\mathrm{c}k}) \quad \mathrm{H}_{n,\beta}(\theta, x_{\mathrm{c}k})\Big).$$

To compare $\mathrm{H}_{\alpha,\beta,n,s}(\theta, \hat{\Delta}_k)$ and $\mathrm{H}_{\alpha,\beta,n}(\theta, x_{\mathrm{c}k})$, it suffices to compare $\mathrm{H}_{n,\hat{\alpha}_s^*}(\theta, \hat{\Delta}_k)$ and $\mathrm{H}_{n,\alpha}(\theta, x_{\mathrm{c}k})$ and to compare $\mathrm{H}_{n,\hat{\beta}_s^*}(\theta, \hat{\Delta}_k)$ and $\mathrm{H}_{n,\beta}(\theta, x_{\mathrm{c}k})$ separately. Due to the similarity in comparison, we examine only the difference between $\mathrm{H}_{n,\hat{\alpha}_s^*}(\theta, \hat{\Delta}_k)$ and $\mathrm{H}_{n,\alpha}(\theta, x_{\mathrm{c}k})$ here.

We now write

$$\mathrm{H}_{n,\hat{\alpha}_s^*}(\theta, \hat{\Delta}_k) - \mathrm{H}_{n,\alpha}(\theta, x_{\mathrm{c}k}) = \mathrm{H}_{s,\alpha,1}(\theta) + \mathrm{H}_{s,\alpha,2}(\theta) + \mathrm{H}_{s,\alpha,3}(\theta), \tag{A.52}$$

where

$$\begin{aligned}
\mathrm{H}_{s,\alpha,1}(\theta) &= \frac{1}{n}\sum_{k=1}^{n}\Big\{(C_t^{\intercal}\hat{\Delta}_k\beta)^{\intercal}C_t^{\intercal}\hat{\Delta}_k\beta - (C_t^{\intercal}x_{\mathrm{c}k}\beta)^{\intercal}C_t^{\intercal}x_{\mathrm{c}k}\beta\Big\}v_{1,k}(\theta;\hat{\Delta}_k),\\
\mathrm{H}_{s,\alpha,2}(\theta) &= \frac{1}{n}\sum_{k=1}^{n}(C_t^{\intercal}x_{\mathrm{c}k}\beta)^{\intercal}C_t^{\intercal}x_{\mathrm{c}k}\beta \times v_{1,k}(\theta;\hat{\Delta}_k) - v_{1,k}(\theta;x_{\mathrm{c}k})\big],\\
\mathrm{H}_{s,\alpha,3}(\theta) &= \left(0_{p\times p} \quad \tfrac{1}{n}\sum_{k=1}^{n}\Big[C^{\intercal}\hat{\Delta}_k\{Y_k - p_k(\theta;\hat{\Delta}_k)\} - C^{\intercal}x_{\mathrm{c}k}\{Y_k - p_k(\theta;x_{\mathrm{c}k})\}\Big]\right),
\end{aligned}$$

where $0_{p\times p}$ represents the $p \times p$ zero matrix.

In the following three steps, we show that all the terms in (A.52) are $o_p(1)$ as $\max(m, n) \to \infty$.

Step1: ***Show that*** $\|H_{s,\alpha,1}(\theta)\| = o_p(1)$ *when* $\min(m,n) \to \infty$.

$$\begin{aligned}
\|\mathrm{H}_{s,\alpha,1}(\theta)\| \leq & \Big\|\frac{1}{n}\sum_{k=1}^{n}\Big[\Big\{C_t^{\intercal}\Big(x_{\mathrm{c}k} + \bar{U}_k + \frac{g_k}{m_c}\Big)\beta\Big\}^{\intercal} C_t^{\intercal}\Big(x_{\mathrm{c}k} + \bar{U}_k + \frac{g_k}{m_c}\Big)\beta \\
& - (C_t^{\intercal} x_{\mathrm{c}k}\beta)^{\intercal} C_t^{\intercal} x_{\mathrm{c}k}\beta\Big]\Big\| \\
= & \Big\|\frac{1}{n}\sum_{k=1}^{n}\Big[\Big(C_t^{\intercal} x_{\mathrm{c}k}\beta\Big)^{\intercal} C_t^{\intercal}\Big(\bar{U}_k + \frac{g_k}{m_c}\Big)\beta + \Big\{C_t^{\intercal}\Big(\bar{U}_k + \frac{g_k}{m_c}\Big)\beta\Big\}^{\intercal} C_t^{\intercal} x_{\mathrm{c}k}\beta \\
& + \Big\{C_t^{\intercal}\Big(\bar{U}_k + \frac{g_k}{m_c} Big)\beta\Big\}^{\intercal} C_t^{\intercal}\Big(\bar{U}_k + \frac{g_k}{m_c}\Big)\beta\Big]\Big\| \\
\leq & \frac{2}{n}\|C_t^{\intercal}\|^2 \times \|\beta\|^2 \times \sum_{k=1}^{n}\Big\{\|x_{\mathrm{c}k}\| \times \Big(\|\bar{U}_k\| + \Big\|\frac{g_k}{m_c}\Big\|\Big) + \|\bar{U}_k\|^2 + \Big\|\frac{g_k}{m_c}\Big\|^2 \\
& + \|\bar{U}_k\| \times \Big\|\frac{g_k}{m_c}\Big\|\Big\} \\
\leq & 2\|C_t^{\intercal}\|^2 \times \|\beta\|^2 \times \frac{1}{n}\sum_{k=1}^{n}\Big(\|x_{\mathrm{c}k}\| \times \|\bar{U}_k\| + \|\bar{U}_k\|^2 + \Big\|\frac{g_k}{m_c}\Big\|^2\Big) \\
& + \frac{2}{m_c}\|C_t^{\intercal}\|^2 \times \|\beta\|^2 \times \Big(\frac{1}{n}\sum_{k=1}^{n}\|x_{\mathrm{c}k}\|^2\Big)^{1/2} \times \Big(\frac{1}{n}\sum_{k=1}^{n}\|g_k\|^2\Big)^{1/2} \\
& + \frac{2}{m_c}\|C_t^{\intercal}\|^2 \times \|\beta\|^2 \times \Big(\frac{1}{n}\sum_{k=1}^{n}\|\bar{U}_k\|^2\Big)^{1/2} \times \Big(\frac{1}{n}\sum_{k=1}^{n}\|g_k\|^2\Big)^{1/2},
\end{aligned}$$

where the first step is due to the definition of $\hat{\Delta}_k$ and the boundedness of $v_{1,k}(\cdot)$ which is between $[0,1]$; and the last two steps are due to the Cauchy–Schwarz inequality. By Conditions (C.2), (C.3), (A.2) and $\sum_{k=1}^{n}\|g_k\|^2 = O_p(n)$, we obtain that $\|\mathrm{H}_{s,\alpha,1}(\theta)\| = o_p(1)$ when $\min(m,n) \to \infty$.

Step2: ***Show that*** $\|H_{s,\alpha,2}(\theta)\| = o_p(1)$ *as* $\min(m,n) \to \infty$.

Plugging (A.8) into $\mathrm{H}_{s,\alpha,2}(\theta)$, we obtain that

$$\begin{aligned}
\|\mathrm{H}_{s,\alpha,2}(\theta)\| &= \Big\|\frac{1}{n}\sum_{k=1}^{n}(C_t^{\intercal}x_{ck}\beta)^{\intercal}C_t^{\intercal}x_{ck}\beta \times v_{2,k}(\theta;\Delta_{k,\xi 2}) \times \mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\Big\{\mathrm{vec}(\bar{U}_k)+\mathrm{vec}(g_k)^{\intercal}/m_c\Big\}\Big\| \\
&\leq \|C_t^{\intercal}\|^2 \times \|\beta\|^2 \times \|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\| \times \frac{1}{n}\sum_{k=1}^{n}\left(\|\bar{U}_k\|+\left\|\frac{g_k}{m_c}\right\|\right) \times \|x_{ck}\|^2 \\
&= \|C_t^{\intercal}\|^2 \times \|\beta\|^2 \times \|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\| \times \frac{1}{n}\sum_{k=1}^{n}\|\bar{U}_k\| \times \|x_{ck}\|^2 \\
&\quad + \|C_t^{\intercal}\|^2 \times \|\beta\|^2 \times \|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\| \times \frac{1}{n}\sum_{k=1}^{n}\left\|\frac{g_k}{m_c}\right\| \times \|x_{ck}\|^2 \\
&\leq \|C_t^{\intercal}\|^2 \times \|\beta\|^2 \times \|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\| \times \frac{1}{n}\sum_{k=1}^{n}\|\bar{U}_k\| \times \|x_{ck}\|^2 \\
&\quad + \left(\|C_t^{\intercal}\|^2 \times \|\beta\|^2 \times \frac{\|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\|}{m_c}\right)(\max_{1\leq k\leq n}\|x_{ck}\|) \\
&\quad \times \left(\frac{1}{n}\sum_{k=1}^{n}\|x_{ck}\|^2\right)^{1/2} \times \left(\frac{1}{n}\sum_{k=1}^{n}\|g_k\|^2\right)^{1/2},
\end{aligned} \tag{A.53}$$

where the second step is due to the boundedness of $p_k(\cdot)$ and $v_{2,k}(\cdot)$, and the Cauchy–Schwarz inequality; and the last step is because of the Cauchy–Schwarz inequality. By Conditions (C.2), (C.3) and Lemma 5.1 in Stefanski and Carroll (1985), the first product term of (A.53) is $o_p(1)$ as $\min(m,n)\to\infty$. Using Conditions (C.1), (C.2) and $\sum_{k=1}^{n}\|g_k\|^2 = O_p(n)$, the second product term of (A.53) is $o_p(1)$ as $\min(m,n)\to\infty$. Thus, $\|\mathrm{H}_{s,\alpha,2}(\theta)\| = o_p(1)$ as $\min(m,n)\to\infty$.

Step3: Show that $\|H_{s,\alpha,3}(\theta)\| = o_p(1)$.

Since, in Appendix A.5, we obtain $\|\mathbf{A.22}\| = o_p(1)$, as a result, we have that

$$\frac{1}{n}\sum_{k=1}^{n}C_t^{\intercal}x_{ck}\{Y_k - p_k(\theta;x_{ck})\} = \frac{1}{n}\sum_{k=1}^{n}C_t^{\intercal}X_k^*\{Y_k - p_k(\theta;X_k^*)\} + o_p(1). \tag{A.54}$$

Plugging (A.54) into $\mathrm{H}_{s,\alpha,3}(\theta)$, we obtain that

$$\begin{aligned}
\mathrm{H}_{s,\alpha,3}(\theta) &= \left(0_{p\times p} \quad \tfrac{1}{n}\sum_{k=1}^{n}\left[C_t^{\intercal}\hat{\Delta}_k\{Y_k - p_k(\theta;\hat{\Delta}_k)\} - C_t^{\intercal}x_{\mathrm{c}k}\{Y_k - p_k(\theta;x_{\mathrm{c}k})\}\right]\right)\\
&= \left(0_{p\times p} \quad \tfrac{1}{n}\sum_{k=1}^{n}\left[C_t^{\intercal}\hat{\Delta}_k\{Y_k - p_k(\theta;\hat{\Delta}_k)\} - C_t^{\intercal}X_k^*\{Y_k - p_k(\theta;X_k^*)\}\right] + o_p(1)\right)\\
&= \left(0_{p\times p} \quad M_s + o_p(1)\right),
\end{aligned} \tag{A.55}$$

where by plugging (A.4) and (A.5) into (A.55), we obtain that

$$\begin{aligned}
M_s &= \frac{1}{n}\sum_{k=1}^{n}\left[C_t^{\intercal}\hat{\Delta}_k\{Y_k - p_k(\theta;\hat{\Delta}_k)\} - C_t^{\intercal}X_k^*\{Y_k - p_k(\theta;X_k^*)\}\right]\\
&= \frac{1}{n}\sum_{k=1}^{n} C_t^{\intercal}X_k^*\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(\bar{\mathbf{U}}_k)[v_{1,k}(\theta;x_{k,\xi}) - v_{1,k}(\theta;\Delta_{k,\xi})]\\
&\quad - \frac{1}{m_c n}\sum_{k=1}^{n} C_t^{\intercal}X_k^*\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(g_k)v_{1,k}(\theta;\Delta_{k,\xi})\\
&\quad - \frac{1}{m_c n}\sum_{k=1}^{n} C_t^{\intercal}g_k\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(g_k)v_{1,k}(\theta;\Delta_{k,\xi}) - \frac{1}{m_c n}\sum_{k=1}^{n} C_t^{\intercal}g_k\{Y_k - p_k(\theta;x_{\mathrm{c}k})\}.
\end{aligned}$$

By Condition (C.3), the first term of M_s is $o_p(1)$; and the last term of M_s is $o_p(1)$ is due to $E(Y_k|x_{\mathrm{c}k}, z_k) = p_k(\theta;x_{\mathrm{c}k})$. Now it remains to examine the middle term of M_s. Let

$$A_s = -\frac{1}{m_c n}\sum_{k=1}^{n} C_t^{\intercal}g_k\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(g_k)v_{1,k}(\theta;\Delta_{k,\xi})$$

and

$$B_s = -\frac{1}{m_c n}\sum_{k=1}^{n} C_t^{\intercal}X_k^*\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(g_k)v_{1,k}(\theta;\Delta_{k,\xi}).$$

Then,

$$\begin{aligned}
\|A_s\| &= \left\| -\frac{1}{m_c n}\sum_{k=1}^{n} C_t^{\intercal}g_k\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\mathrm{vec}(g_k)v_{1,k}(\theta;\Delta_{k,\xi})\right\|\\
&\le \|C_t^{\intercal}\| \times \|\mathrm{vec}(\alpha\beta^{\intercal})^{\intercal}\| \times \frac{1}{m_c n}\sum_{k=1}^{n}\|g_k\|^2,
\end{aligned}$$

By the assumption that $\sum_{k=1}^{n}\|g_k\|^2 = O_p(n)$ and the boundedness of $v_{1,k}(\cdot)$, we obtain that $\|A_s\| = o_p(1)$ as $\min(m, n) \to \infty$.

Similarly,

$$\begin{aligned}\|B_s\| &= \Big\| - \frac{1}{m_c n}\sum_{k=1}^{n} C_t^\intercal X_k^* \mathrm{vec}(\alpha\beta^\intercal)^\intercal \mathrm{vec}(g_k) v_{1,k}(\theta; \Delta_{k,\xi})\Big\| \\ &\le (\|C_t^\intercal\| \times \|\mathrm{vec}(\alpha\beta^\intercal)^\intercal\|/m_c) \times \Big(\frac{1}{n}\sum_{k=1}^{n}\|X_k^*\|^2\Big)^{1/2}\Big(\frac{1}{n}\sum_{k=1}^{n}\|g_k\|^2\Big)^{1/2}.\end{aligned}$$

By Conditions (C.2) and (C.3) in Appendix A.1, we know that $\frac{1}{n}\sum_{k=1}^{n}\|x_{\mathrm{c}k}\|^2 = O(1)$, and $\frac{1}{n}\sum_{k=1}^{n}\|\bar{U}_k\|^2 = o_p(1)$ when $\min(m, n) \to \infty$. Consequently,

$$\Big(\frac{1}{n}\sum_{k=1}^{n}\|X_k^*\|^2\Big)^{1/2} \le \Big(\frac{1}{n}\sum_{k=1}^{n}\|x_{\mathrm{c}k}\|^2 + 2\|x_{\mathrm{c}k}\| \times \|\bar{U}_k\| + \|\bar{U}_k\|^2\Big)^{1/2} = O_p(1).$$

By the assumption $\sum_{k=1}^{n}\|g_k\|^2 = O_p(n)$, we obtain that $\|B_s\| = o_p(1)$. As a result, $\|\mathrm{H}_{s,\alpha,3}(\theta)\| = o_p(1)$ when $\min(m, n) \to \infty$.

Combining the results of Steps 1-3, we obtain that $\mathrm{H}_{n,\hat{\alpha}_s^*}(\theta, \hat{\Delta}_k) - \mathrm{H}_{n,\alpha}(\theta, x_{\mathrm{c}k}) = o_p(1)$. Following the same steps as Steps 1-3, we obtain that $\mathrm{H}_{n,\hat{\beta}_s^*}(\theta, \hat{\Delta}_k) - \mathrm{H}_{n,\beta}(\theta, x_{\mathrm{c}k}) = o_p(1)$. These two results show (A.51), and thus by the comments after (A.51), we obtain that

$$\mathrm{H}_{n,s}(\theta, \hat{\Delta}_k) - \mathrm{H}_n(\theta) = o_p(1).$$

Finally, using the results we showed in Part I and Part II and following the same steps in Appendix A.7, we can show Theorem 2.4(a) by Conditions (C.1), (C.5), the Continuous Mapping Theorem and the asymptotic normal distribution of $\mathrm{H}_n^{-1/2}(\theta)\mathrm{S}_n(\theta)$ (Stefanski and Carroll, 1985). Theorem 2.4(b) is obtained by applying Slutsky's theorem.

A.11 Additional Simulation Results for Sections 2.4.1-2.4.2

The following tables record simulation results for different settings **described in Section 2.4**.

- Table A.1 records the simulation results for the *row* effects with $p_x = 5$, E_{kr} generated from the matrix normal distribution, and $n = 1000$;
- Table A.2 records the simulation results for the *column* effects and covariate effects with $p_x = 5$, E_{kr} generated from the matrix normal distribution, and $n = 1000$;
- Table A.3 records the simulation results for the *row* effects with $p_x = 10$, $\sigma = 0.25$, E_{kr} generated from the matrix normal distribution, and $n = 1000$;
- Table A.4 records the simulation results for the *row* effects with $p_x = 10$, $\sigma = 0.5$, E_{kr} generated from the matrix normal distribution, and $n = 1000$;
- Table A.5 records the simulation results for the *row* effects with $p_x = 10$, $\sigma = 0.75$, E_{kr} generated from the matrix normal distribution, and $n = 1000$;
- Table A.6 records the simulation results for the *column* effects and covariate effects with $p_x = 10$, $\sigma = 0.25$, E_{kr} generated from the matrix normal distribution, and $n = 1000$;
- Table A.7 records the simulation results for the *column* effects and covariate effects with $p_x = 10$, $\sigma = 0.5$, E_{kr} generated from the matrix normal distribution, and $n = 1000$;
- Table A.8 records the simulation results for the *column* effects and covariate effects with $p_x = 10$, $\sigma = 0.75$, E_{kr} generated from the matrix normal distribution, and $n = 1000$;
- Table A.9 records the simulation results for the *row* effects with $p_x = 20$, E_{kr} generated from matrix normal distribution, and $n = 1000$;
- Table A.10 records the simulation results for the *column* effects and covariate effects with $p_x = 20$, E_{kr} generated from matrix normal distribution, and $n = 1000$;
- Table A.11 records the simulation results for the *row* effects with $p_x = 5$, and E_{kr} generated from the matrix t-distribution;

- Table A.12 records the simulation results for the *column* effects and covariate effects with $p_x = 5$, and E_{kr} generated from the matrix t-distribution;

- Table A.13 records the simulation results for the *row* effects with $p_x = 20$, E_{kr} generated from the matrix normal distribution, and $n = 2000$;

- Table A.14 records the simulation results for the *column* effects and covariate effects with $p_x = 20$, E_{kr} generated from the matrix normal distribution, and $n = 2000$;

Table A.1: Simulation results for the *row* parameters with $p_x = 5$, E_{kr} is generated from the matrix normal distribution, and $n = 1000$

Parameter	σ	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
			Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
α_1	0.25	2	1.192	0.062	0.063	94.4	0.004	1.232	0.063	0.062	93.8	0.004	1.206	0.063	0.061	89.6	0.004
		5	1.423	0.060	0.060	94.4	0.004	1.434	0.060	0.060	94.0	0.004	1.437	0.060	0.060	92.4	0.004
		10	1.442	0.060	0.060	93.8	0.004	1.448	0.060	0.060	93.8	0.004	1.442	0.060	0.059	93.0	0.004
	0.5	2	0.987	0.071	0.072	94.4	0.005	1.109	0.074	0.069	92.4	0.006	1.013	0.073	0.067	92.0	0.005
		5	1.398	0.064	0.064	95.2	0.004	1.424	0.064	0.063	94.2	0.004	1.444	0.064	0.062	94.6	0.004
		10	1.495	0.063	0.062	94.6	0.004	1.529	0.063	0.061	94.2	0.004	1.501	0.063	0.061	94.4	0.004
	0.75	2	0.864	0.083	0.083	94.2	0.007	0.969	0.088	0.077	91.0	0.008	0.849	0.086	0.074	90.8	0.007
		5	1.385	0.069	0.070	95.0	0.005	1.412	0.071	0.068	94.6	0.005	1.450	0.071	0.066	94.0	0.005
		10	1.576	0.066	0.065	94.6	0.004	1.669	0.067	0.064	94.4	0.005	1.599	0.067	0.063	94.0	0.005
α_3	0.25	2	0.488	0.078	0.079	94.4	0.006	0.534	0.079	0.078	94.2	0.006	0.520	0.079	0.077	91.8	0.006
		5	0.348	0.079	0.076	94.4	0.006	0.362	0.079	0.076	94.2	0.006	0.361	0.079	0.075	91.6	0.006
		10	0.404	0.078	0.075	95.0	0.006	0.412	0.078	0.075	95.0	0.006	0.410	0.078	0.075	92.6	0.006
	0.5	2	0.574	0.087	0.090	95.8	0.008	0.690	0.090	0.087	93.6	0.008	0.645	0.090	0.085	93.4	0.008
		5	0.317	0.084	0.082	93.6	0.007	0.331	0.086	0.080	93.0	0.007	0.344	0.086	0.079	93.2	0.007
		10	0.414	0.081	0.078	93.2	0.007	0.449	0.081	0.077	93.4	0.007	0.438	0.081	0.077	93.2	0.007
	0.75	2	0.735	0.100	0.106	96.4	0.010	0.928	0.107	0.098	93.6	0.011	0.808	0.106	0.094	93.0	0.011
		5	0.348	0.091	0.089	94.2	0.008	0.314	0.095	0.086	92.2	0.009	0.373	0.095	0.084	92.0	0.009
		10	0.422	0.085	0.082	92.8	0.007	0.489	0.086	0.081	93.8	0.007	0.468	0.086	0.079	93.6	0.007
α_4	0.25	2	0.500	0.079	0.079	95.4	0.006	0.517	0.080	0.078	94.6	0.006	0.497	0.080	0.077	91.6	0.006
		5	0.232	0.077	0.076	94.2	0.006	0.225	0.077	0.076	93.6	0.006	0.223	0.077	0.075	92.4	0.006
		10	0.330	0.077	0.075	94.8	0.006	0.329	0.077	0.075	94.6	0.006	0.326	0.077	0.075	91.6	0.006
	0.5	2	0.796	0.090	0.091	94.6	0.008	0.948	0.093	0.087	93.2	0.009	0.910	0.092	0.085	92.8	0.009
		5	0.216	0.082	0.081	94.2	0.007	0.177	0.083	0.080	93.4	0.007	0.180	0.083	0.078	92.6	0.007
		10	0.409	0.081	0.078	94.6	0.006	0.410	0.081	0.077	94.0	0.007	0.399	0.081	0.077	93.4	0.007
	0.75	2	1.105	0.106	0.106	95.6	0.011	1.318	0.112	0.098	92.0	0.013	1.337	0.109	0.095	91.6	0.012
		5	0.253	0.088	0.089	94.2	0.008	0.151	0.091	0.085	92.2	0.008	0.190	0.092	0.084	91.2	0.008
		10	0.511	0.085	0.083	95.4	0.007	0.518	0.087	0.081	94.4	0.008	0.501	0.087	0.079	93.6	0.008
α_5	0.25	2	-0.075	0.078	0.079	93.8	0.006	-0.044	0.079	0.078	94.0	0.009	-0.041	0.079	0.077	91.2	0.006
		5	-0.110	0.077	0.076	95.0	0.006	-0.100	0.077	0.076	94.6	0.007	-0.093	0.077	0.075	92.6	0.006
		10	-0.040	0.077	0.075	93.8	0.006	-0.034	0.077	0.075	94.0	0.006	-0.035	0.077	0.075	93.4	0.006
	0.5	2	-0.087	0.089	0.090	94.8	0.008	0.014	0.093	0.087	92.8	0.012	0.015	0.091	0.085	92.8	0.008
		5	-0.124	0.082	0.081	95.2	0.007	-0.103	0.082	0.080	95.0	0.008	-0.066	0.083	0.079	93.6	0.007
		10	0.020	0.080	0.078	94.4	0.006	0.050	0.080	0.077	94.8	0.007	0.038	0.080	0.077	94.4	0.006
	0.75	2	-0.059	0.103	0.105	96.0	0.011	0.033	0.110	0.097	90.8	0.012	0.041	0.108	0.094	90.8	0.012
		5	-0.100	0.089	0.089	95.2	0.008	-0.093	0.092	0.086	92.6	0.008	0.000	0.092	0.084	92.2	0.008
		10	0.084	0.084	0.082	94.8	0.007	0.160	0.086	0.081	94.4	0.007	0.129	0.086	0.080	93.6	0.007

Table A.2: Simulation results for the *column* parameters and covariate parameters with $p_x = 5$, E_{kr} is generated from the matrix normal distribution, and $n = 1000$

Parameter	σ	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
			Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
β_1	0.25	2	-9.082	0.089	0.086	76.0	0.016	1.021	0.107	0.104	94.3	0.012	1.345	0.108	0.097	93.0	0.012
		5	-2.766	0.098	0.092	89.5	0.010	2.016	0.106	0.100	93.0	0.012	2.090	0.106	0.097	91.8	0.012
		10	-0.656	0.098	0.093	94.0	0.010	1.866	0.103	0.097	93.8	0.011	1.891	0.103	0.096	93.0	0.011
	0.5	2	-30.079	0.071	0.070	2.4	0.095	-8.581	0.106	0.108	83.8	0.019	-7.138	0.111	0.094	80.0	0.017
		5	-14.297	0.087	0.082	53.8	0.028	-0.299	0.112	0.107	93.6	0.013	0.305	0.114	0.098	90.0	0.013
		10	-7.178	0.093	0.088	80.6	0.014	1.318	0.108	0.102	94.8	0.012	1.553	0.109	0.097	93.4	0.012
	0.75	2	-48.331	0.057	0.056	0.0	0.237	-24.022	0.095	0.096	29.6	0.067	-22.216	0.102	0.081	27.8	0.060
		5	-27.863	0.075	0.071	4.0	0.083	-7.012	0.111	0.107	87.0	0.017	-5.632	0.116	0.094	82.6	0.017
		10	-16.210	0.085	0.080	49.8	0.033	-1.343	0.111	0.106	93.6	0.013	-0.605	0.114	0.096	90.4	0.013
β_2	0.25	2	-9.744	0.062	0.061	84.8	0.006	0.259	0.072	0.070	95.3	0.005	0.591	0.072	0.068	94.0	0.005
		5	-3.367	0.068	0.064	90.5	0.005	1.397	0.072	0.068	91.8	0.005	1.479	0.073	0.067	91.5	0.005
		10	-1.050	0.069	0.065	94.3	0.005	1.466	0.071	0.067	95.3	0.005	1.491	0.071	0.067	95.0	0.005
	0.5	2	-30.756	0.051	0.052	17.8	0.026	-9.485	0.071	0.072	87.4	0.007	-8.020	0.074	0.067	86.2	0.007
		5	-15.170	0.061	0.059	72.0	0.010	-1.251	0.076	0.072	91.2	0.006	-0.573	0.077	0.068	89.2	0.006
		10	-7.738	0.064	0.062	86.0	0.006	0.728	0.073	0.070	95.2	0.005	0.964	0.073	0.068	94.8	0.005
	0.75	2	-48.876	0.043	0.043	0.0	0.062	-24.866	0.066	0.065	50.6	0.020	-23.071	0.071	0.060	47.2	0.018
		5	-28.757	0.055	0.052	23.0	0.024	-8.107	0.076	0.071	88.6	0.007	-6.552	0.080	0.067	85.2	0.007
		10	-16.656	0.059	0.058	67.0	0.010	-1.807	0.075	0.072	93.2	0.006	-1.068	0.077	0.067	91.2	0.006
β_3	0.25	2	-8.872	0.091	0.086	74.8	0.016	1.295	0.109	0.104	94.0	0.012	1.628	0.110	0.097	92.8	0.012
		5	-2.564	0.099	0.092	90.0	0.010	2.246	0.108	0.100	93.3	0.012	2.322	0.108	0.097	92.5	0.012
		10	-0.361	0.101	0.093	92.5	0.010	2.176	0.105	0.097	93.3	0.011	2.202	0.106	0.096	92.8	0.011
	0.5	2	-29.961	0.070	0.070	2.6	0.095	-8.350	0.105	0.108	83.0	0.018	-6.855	0.111	0.094	81.0	0.017
		5	-14.260	0.087	0.082	56.4	0.028	-0.197	0.113	0.107	93.8	0.013	0.416	0.115	0.098	90.6	0.013
		10	-6.912	0.093	0.088	84.2	0.014	1.632	0.109	0.102	94.2	0.012	1.861	0.110	0.097	92.8	0.012
	0.75	2	-48.355	0.056	0.057	0.0	0.237	-23.991	0.094	0.099	30.4	0.066	-22.143	0.101	0.082	27.6	0.059
		5	-27.939	0.075	0.071	3.2	0.084	-6.981	0.112	0.107	88.2	0.018	-5.577	0.118	0.094	82.2	0.017
		10	-15.961	0.086	0.081	47.6	0.033	-0.992	0.113	0.108	93.6	0.013	-0.274	0.115	0.098	92.4	0.013
β_4	0.25	2	-8.910	0.085	0.086	76.8	0.015	1.228	0.101	0.103	97.3	0.011	1.559	0.102	0.097	95.8	0.011
		5	-2.595	0.093	0.091	93.0	0.009	2.192	0.101	0.099	94.8	0.011	2.272	0.101	0.096	93.8	0.011
		10	-0.472	0.094	0.093	93.5	0.009	2.051	0.098	0.097	95.0	0.010	2.078	0.099	0.096	94.8	0.010
	0.5	2	-29.879	0.068	0.070	1.8	0.094	-8.249	0.101	0.108	84.2	0.017	-6.769	0.106	0.094	81.4	0.016
		5	-14.119	0.086	0.082	55.6	0.027	-0.097	0.110	0.106	93.6	0.012	0.545	0.112	0.097	90.2	0.013
		10	-7.024	0.091	0.088	82.8	0.013	1.463	0.105	0.102	94.4	0.011	1.697	0.106	0.097	92.6	0.011
	0.75	2	-48.241	0.054	0.056	0.0	0.236	-23.813	0.090	0.098	30.4	0.065	-21.962	0.096	0.083	26.8	0.058
		5	-27.700	0.076	0.070	4.0	0.082	-6.787	0.112	0.104	86.8	0.017	-5.339	0.116	0.091	82.8	0.016
		10	-16.030	0.084	0.079	45.8	0.033	-1.153	0.110	0.105	92.8	0.012	-0.438	0.111	0.095	90.0	0.012
β_5	0.25	2	-9.053	0.091	0.086	76.3	0.016	1.099	0.108	0.103	93.0	0.011	1.435	0.109	0.097	91.0	0.012
		5	-2.578	0.097	0.092	90.0	0.010	2.227	0.106	0.100	95.5	0.011	2.294	0.106	0.097	94.0	0.012
		10	-0.552	0.098	0.093	93.3	0.009	1.977	0.102	0.098	94.0	0.011	2.001	0.102	0.096	94.0	0.011
	0.5	2	-30.188	0.070	0.070	2.6	0.096	-8.621	0.105	0.108	83.4	0.018	-7.101	0.110	0.094	80.2	0.017
		5	-14.281	0.087	0.082	54.2	0.028	-0.239	0.112	0.107	93.4	0.013	0.369	0.114	0.098	91.2	0.013
		10	-7.259	0.091	0.088	81.2	0.014	1.254	0.106	0.102	94.8	0.011	1.481	0.107	0.097	92.6	0.012
	0.75	2	-48.523	0.056	0.056	0.0	0.239	-24.226	0.093	0.096	26.2	0.067	-22.379	0.100	0.081	24.2	0.060
		5	-27.888	0.076	0.071	4.4	0.084	-6.986	0.113	0.107	87.6	0.018	-5.598	0.118	0.094	83.8	0.017
		10	-16.300	0.084	0.081	44.2	0.034	-1.398	0.109	0.108	94.6	0.012	-0.680	0.111	0.097	91.6	0.012
γ	0.25	2	-5.747	0.110	0.109	94.3	0.013	1.877	0.121	0.121	95.8	0.015	2.285	0.122	0.120	95.0	0.015
		5	-1.171	0.114	0.113	94.3	0.013	2.454	0.119	0.118	93.8	0.014	2.555	0.119	0.117	93.5	0.014
		10	0.363	0.115	0.114	94.5	0.013	2.277	0.118	0.117	93.5	0.013	2.306	0.118	0.116	93.5	0.013
	0.5	2	-20.889	0.099	0.099	80.2	0.021	-5.330	0.124	0.127	95.6	0.016	-3.505	0.129	0.125	94.4	0.017
		5	-9.686	0.106	0.107	91.0	0.014	0.655	0.121	0.123	95.0	0.015	1.363	0.123	0.122	94.4	0.015
		10	-4.921	0.109	0.110	94.4	0.012	1.381	0.117	0.120	94.2	0.014	1.603	0.118	0.119	94.0	0.014
	0.75	2	-33.506	0.089	0.090	53.0	0.036	-16.864	0.118	0.127	91.0	0.021	-14.352	0.126	0.123	90.2	0.021
		5	-19.413	0.099	0.100	83.6	0.019	-4.445	0.123	0.126	94.4	0.016	-2.740	0.127	0.124	93.8	0.016
		10	-11.602	0.104	0.106	91.6	0.014	-0.771	0.120	0.124	94.6	0.014	-0.082	0.120	0.122	94.0	0.015

Table A.3: Simulation results for the *row* parameters with $p_x = 10$, $\sigma = 0.25$, E_{kr} generated is from the matrix normal distribution, and $n = 1000$

Parameter	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
		Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
α_1	2	1.767	0.071	0.069	93.6	0.005	1.766	0.075	0.068	92.4	0.006	1.588	0.074	0.069	93.8	0.006
	5	1.353	0.065	0.064	94.8	0.004	1.306	0.067	0.063	93.4	0.004	1.318	0.067	0.064	94.0	0.004
	10	1.368	0.063	0.061	94.2	0.004	1.369	0.063	0.061	93.6	0.004	1.369	0.063	0.061	93.6	0.004
α_3	2	1.179	0.090	0.088	95.0	0.008	1.008	0.094	0.087	92.6	0.009	0.954	0.095	0.088	92.0	0.009
	5	1.187	0.084	0.081	92.8	0.007	1.158	0.086	0.080	91.8	0.008	1.133	0.087	0.080	92.4	0.008
	10	1.138	0.081	0.078	93.2	0.007	1.109	0.082	0.078	93.2	0.007	1.107	0.082	0.078	93.2	0.007
α_4	2	0.640	0.091	0.087	93.4	0.008	0.495	0.096	0.086	92.0	0.009	0.502	0.097	0.087	92.6	0.009
	5	0.766	0.085	0.081	93.0	0.007	0.780	0.088	0.080	92.4	0.008	0.765	0.088	0.080	92.2	0.007
	10	0.686	0.079	0.078	94.4	0.006	0.672	0.080	0.077	93.6	0.006	0.673	0.080	0.077	92.8	0.006
α_5	2	0.768	0.088	0.087	95.0	0.008	0.667	0.094	0.086	93.2	0.009	0.604	0.094	0.087	92.8	0.009
	5	0.855	0.082	0.080	93.4	0.007	0.909	0.086	0.080	90.6	0.008	0.874	0.086	0.080	91.0	0.007
	10	0.679	0.078	0.078	95.0	0.006	0.689	0.080	0.077	94.6	0.006	0.680	0.080	0.077	95.0	0.007
α_6	2	1.186	0.088	0.087	94.2	0.008	0.984	0.091	0.086	91.6	0.008	0.866	0.092	0.087	92.4	0.008
	5	1.332	0.084	0.081	93.8	0.007	1.316	0.087	0.080	92.8	0.008	1.317	0.087	0.080	92.6	0.008
	10	1.044	0.080	0.078	94.0	0.006	1.011	0.081	0.078	93.2	0.007	1.008	0.081	0.078	93.0	0.007
α_7	2	1.289	0.088	0.087	95.0	0.008	1.096	0.094	0.086	91.8	0.009	0.982	0.094	0.087	92.4	0.009
	5	1.353	0.082	0.081	94.0	0.007	1.359	0.084	0.081	93.2	0.007	1.392	0.085	0.081	93.0	0.007
	10	1.293	0.079	0.078	94.2	0.006	1.272	0.080	0.077	93.6	0.007	1.273	0.080	0.077	93.4	0.007
α_8	2	1.154	0.086	0.087	95.4	0.008	1.101	0.091	0.086	94.0	0.008	1.025	0.092	0.087	93.2	0.009
	5	1.120	0.081	0.081	95.0	0.007	1.170	0.083	0.081	94.0	0.007	1.158	0.083	0.081	93.6	0.007
	10	1.047	0.078	0.078	94.2	0.006	1.060	0.079	0.078	93.2	0.006	1.070	0.080	0.078	93.8	0.006
α_9	2	1.213	0.094	0.088	93.6	0.009	1.143	0.099	0.087	91.2	0.010	1.037	0.098	0.088	91.6	0.010
	5	1.007	0.086	0.080	93.2	0.007	0.978	0.087	0.080	91.2	0.008	0.978	0.087	0.080	92.0	0.008
	10	0.852	0.084	0.078	92.8	0.007	0.815	0.084	0.078	92.8	0.007	0.817	0.085	0.078	92.4	0.007
α_{10}	2	1.071	0.092	0.087	93.6	0.009	0.957	0.097	0.086	90.4	0.009	0.911	0.097	0.087	91.4	0.009
	5	1.030	0.088	0.081	91.8	0.008	1.100	0.091	0.081	90.6	0.008	1.059	0.090	0.081	90.0	0.008
	10	0.881	0.079	0.077	94.2	0.006	0.867	0.081	0.077	94.6	0.007	0.866	0.081	0.077	93.8	0.007

Table A.4: Simulation results for the *row* parameters with $p_x = 10$, $\sigma = 0.5$, E_{kr} is generated from the matrix normal distribution, and $n = 1000$

Parameter	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
		Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
α_1	2	0.113	0.091	0.088	93.4	0.008	-0.359	0.104	0.085	89.2	0.011	-0.455	0.102	0.087	88.4	0.010
	5	-0.492	0.077	0.073	94.0	0.006	-1.053	0.084	0.073	90.8	0.007	-0.891	0.083	0.074	92.0	0.007
	10	-0.466	0.074	0.067	91.6	0.006	-0.720	0.077	0.066	89.4	0.006	-0.633	0.075	0.067	91.8	0.006
α_3	2	0.491	0.121	0.111	93.8	0.015	0.256	0.133	0.108	89.8	0.018	0.672	0.133	0.112	89.6	0.018
	5	0.946	0.098	0.094	95.6	0.010	0.936	0.104	0.093	92.2	0.011	1.036	0.106	0.095	91.8	0.011
	10	0.965	0.087	0.085	93.2	0.008	0.949	0.090	0.085	92.0	0.008	1.019	0.092	0.085	92.0	0.009
α_4	2	0.382	0.116	0.111	93.0	0.014	0.139	0.123	0.107	91.2	0.015	0.194	0.122	0.111	93.6	0.015
	5	0.138	0.099	0.093	91.8	0.010	0.062	0.108	0.092	88.2	0.012	0.118	0.109	0.094	86.6	0.012
	10	0.449	0.086	0.085	93.4	0.008	0.472	0.091	0.085	90.8	0.008	0.493	0.092	0.085	92.8	0.008
α_5	2	0.527	0.113	0.112	94.6	0.013	-0.145	0.125	0.108	89.0	0.016	0.284	0.122	0.111	91.2	0.015
	5	0.561	0.098	0.093	93.2	0.010	0.700	0.110	0.092	88.6	0.012	0.694	0.108	0.094	89.0	0.012
	10	0.117	0.084	0.085	94.6	0.007	0.137	0.089	0.084	94.4	0.008	0.303	0.089	0.085	94.8	0.008
α_6	2	1.042	0.106	0.111	94.6	0.011	0.890	0.111	0.108	94.0	0.012	1.067	0.112	0.111	92.6	0.013
	5	1.783	0.102	0.095	95.0	0.011	1.807	0.109	0.094	90.0	0.012	1.934	0.112	0.096	88.2	0.013
	10	0.959	0.085	0.086	96.0	0.007	0.865	0.088	0.085	94.4	0.008	0.749	0.089	0.086	94.0	0.008
α_7	2	0.501	0.108	0.111	95.0	0.012	0.513	0.118	0.108	91.2	0.014	0.473	0.120	0.111	90.2	0.015
	5	0.881	0.099	0.094	93.6	0.010	1.027	0.107	0.093	93.2	0.012	1.174	0.107	0.096	93.2	0.012
	10	0.187	0.087	0.085	94.0	0.008	0.228	0.090	0.085	92.8	0.008	0.280	0.090	0.086	92.4	0.008
α_8	2	0.133	0.108	0.110	95.6	0.012	0.094	0.119	0.107	88.6	0.014	0.164	0.117	0.110	93.2	0.014
	5	0.320	0.092	0.094	95.8	0.008	0.621	0.102	0.093	94.4	0.010	0.646	0.101	0.095	95.0	0.010
	10	-0.259	0.082	0.085	95.2	0.007	-0.098	0.086	0.084	93.8	0.007	-0.039	0.088	0.085	94.8	0.008
α_9	2	1.497	0.133	0.112	87.0	0.018	1.488	0.144	0.109	84.6	0.021	1.989	0.142	0.114	88.2	0.021
	5	1.249	0.108	0.094	90.8	0.012	1.473	0.115	0.092	87.4	0.013	1.267	0.114	0.095	87.6	0.013
	10	0.676	0.085	0.085	94.4	0.007	0.568	0.087	0.084	94.2	0.008	0.608	0.089	0.085	94.4	0.008
α_{10}	2	0.182	0.114	0.112	94.6	0.013	0.029	0.122	0.109	91.0	0.015	0.360	0.120	0.113	91.6	0.014
	5	0.694	0.103	0.094	92.6	0.011	1.011	0.111	0.094	90.2	0.013	0.866	0.110	0.096	90.6	0.012
	10	0.239	0.084	0.085	94.6	0.007	0.260	0.088	0.084	91.6	0.008	0.224	0.090	0.085	92.6	0.008

Table A.5: Simulation results for the *row* parameters with $p_x = 10$, $\sigma = 0.75$, E_{kr} is generated from the matrix normal distribution, and $n = 1000$

Parameter	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
		Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
α_1	2	2.654	0.107	0.109	94.8	0.012	3.355	0.120	0.103	90.6	0.015	3.025	0.120	0.103	89.6	0.015
	5	1.504	0.087	0.086	95.4	0.008	1.515	0.100	0.085	90.6	0.010	1.319	0.101	0.087	92.0	0.010
	10	1.211	0.076	0.076	94.2	0.006	1.441	0.082	0.075	92.2	0.007	1.311	0.081	0.077	92.0	0.007
α_3	2	2.002	0.145	0.140	94.6	0.021	1.911	0.160	0.132	89.6	0.026	2.108	0.159	0.133	90.0	0.026
	5	1.210	0.107	0.110	96.0	0.012	1.138	0.117	0.108	92.0	0.014	1.420	0.119	0.111	92.2	0.014
	10	1.267	0.098	0.096	93.6	0.010	1.154	0.104	0.095	91.6	0.011	1.186	0.103	0.097	93.6	0.011
α_4	2	1.603	0.146	0.139	94.0	0.022	1.409	0.166	0.132	89.4	0.028	1.405	0.164	0.133	90.2	0.027
	5	0.645	0.114	0.110	93.6	0.013	0.898	0.127	0.108	90.8	0.016	0.973	0.128	0.111	90.2	0.016
	10	0.526	0.093	0.096	96.0	0.009	0.575	0.102	0.095	93.4	0.010	0.560	0.101	0.097	93.4	0.010
α_5	2	1.667	0.141	0.139	94.2	0.020	1.373	0.162	0.132	89.0	0.026	1.229	0.158	0.133	89.6	0.025
	5	0.832	0.111	0.110	94.4	0.012	0.948	0.128	0.108	89.6	0.017	1.121	0.126	0.111	90.0	0.016
	10	0.481	0.095	0.096	95.2	0.009	0.452	0.105	0.095	91.8	0.011	0.597	0.107	0.097	92.4	0.011
α_6	2	2.235	0.142	0.139	95.0	0.021	1.836	0.156	0.132	92.2	0.025	2.068	0.156	0.134	91.8	0.025
	5	1.862	0.117	0.111	93.2	0.014	1.940	0.132	0.108	90.8	0.018	2.348	0.133	0.112	88.6	0.018
	10	1.062	0.098	0.096	95.8	0.010	1.006	0.106	0.095	92.6	0.011	0.960	0.104	0.097	92.6	0.011
α_7	2	2.045	0.142	0.139	95.6	0.021	1.947	0.162	0.132	90.0	0.027	1.701	0.153	0.132	91.6	0.024
	5	1.292	0.111	0.110	95.2	0.012	1.323	0.120	0.108	91.4	0.015	1.658	0.122	0.111	92.8	0.015
	10	1.201	0.099	0.096	93.6	0.010	1.203	0.108	0.094	90.4	0.012	1.042	0.107	0.096	90.0	0.012
α_8	2	1.854	0.141	0.139	94.8	0.020	1.855	0.157	0.131	89.0	0.025	1.772	0.156	0.132	89.4	0.024
	5	0.883	0.115	0.110	94.2	0.013	1.163	0.127	0.108	91.2	0.016	1.216	0.123	0.111	92.6	0.015
	10	0.843	0.098	0.097	95.6	0.010	1.014	0.106	0.096	92.0	0.011	1.075	0.108	0.098	92.6	0.012
α_9	2	1.968	0.139	0.140	95.0	0.020	2.044	0.155	0.132	89.4	0.024	2.067	0.149	0.133	91.6	0.023
	5	1.075	0.116	0.110	94.2	0.013	1.209	0.125	0.108	92.0	0.016	1.275	0.127	0.110	90.2	0.016
	10	0.764	0.100	0.096	93.4	0.010	0.675	0.106	0.095	91.4	0.011	0.724	0.107	0.097	92.2	0.012
α_{10}	2	1.908	0.137	0.139	96.2	0.019	1.979	0.157	0.132	90.6	0.025	1.926	0.155	0.134	90.4	0.024
	5	1.000	0.116	0.110	93.8	0.014	1.340	0.131	0.108	90.4	0.017	1.379	0.130	0.112	91.2	0.017
	10	0.854	0.093	0.095	95.4	0.009	0.800	0.101	0.094	93.0	0.010	0.646	0.100	0.096	94.2	0.010

Table A.6: Simulation results for the *column* parameters and covariate parameters with $p_x = 10$, $\sigma = 0.25$, E_{kr} is generated from the matrix normal distribution, and $n = 1000$

Parameter	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
		Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
β_1	2	-22.270	0.086	0.086	28.2	0.058	0.646	0.144	0.154	95.8	0.018	5.362	0.172	0.138	89.8	0.027
	5	-7.806	0.115	0.103	80.8	0.019	6.956	0.158	0.146	93.8	0.028	8.594	0.169	0.134	88.4	0.033
	10	-0.734	0.123	0.112	91.2	0.014	8.384	0.151	0.138	93.0	0.028	8.911	0.154	0.129	89.8	0.030
β_2	2	-22.057	0.056	0.058	49.8	0.015	0.874	0.084	0.092	97.2	0.007	5.481	0.095	0.087	93.6	0.009
	5	-7.560	0.069	0.067	86.4	0.006	7.222	0.090	0.089	95.0	0.009	8.812	0.095	0.084	91.4	0.011
	10	-0.509	0.077	0.072	91.6	0.006	8.691	0.092	0.085	93.2	0.010	9.176	0.094	0.082	91.2	0.011
β_3	2	-21.780	0.088	0.087	28.4	0.057	1.264	0.144	0.155	96.4	0.019	5.909	0.167	0.138	90.4	0.028
	5	-7.475	0.113	0.104	81.4	0.018	7.318	0.157	0.147	94.2	0.027	8.926	0.167	0.134	86.8	0.032
	10	-0.351	0.122	0.113	94.2	0.014	8.822	0.149	0.139	93.2	0.028	9.298	0.152	0.131	90.2	0.030
β_4	2	-21.881	0.086	0.087	30.6	0.057	1.106	0.143	0.155	97.4	0.019	5.827	0.169	0.138	90.0	0.028
	5	-7.299	0.113	0.104	81.4	0.018	7.530	0.156	0.147	94.8	0.028	9.197	0.167	0.134	88.0	0.034
	10	-0.270	0.120	0.113	92.8	0.014	8.905	0.148	0.139	92.6	0.028	9.407	0.151	0.131	90.0	0.030
β_5	2	-22.028	0.092	0.087	29.4	0.058	0.999	0.150	0.155	95.4	0.020	5.605	0.175	0.139	89.8	0.028
	5	-7.541	0.113	0.104	81.6	0.018	7.271	0.156	0.147	95.6	0.026	8.895	0.167	0.135	88.6	0.032
	10	-0.411	0.121	0.113	92.2	0.014	8.767	0.148	0.139	93.0	0.028	9.260	0.151	0.131	89.6	0.029
β_6	2	-21.878	0.087	0.087	31.6	0.057	1.045	0.142	0.154	95.4	0.018	5.653	0.167	0.138	90.0	0.027
	5	-7.520	0.112	0.103	81.4	0.018	7.231	0.153	0.146	95.0	0.027	8.817	0.163	0.134	87.4	0.032
	10	-0.274	0.120	0.113	93.0	0.014	8.870	0.147	0.139	93.2	0.028	9.361	0.150	0.130	89.4	0.030
β_7	2	-22.354	0.063	0.059	48.0	0.017	0.507	0.094	0.093	92.8	0.008	4.904	0.104	0.087	90.2	0.011
	5	-7.928	0.075	0.067	84.0	0.007	6.762	0.096	0.088	94.2	0.010	8.331	0.101	0.084	88.8	0.011
	10	-0.905	0.082	0.072	90.4	0.007	8.206	0.096	0.085	91.6	0.011	8.700	0.098	0.081	89.2	0.011
β_8	2	-22.098	0.087	0.087	31.0	0.058	0.894	0.145	0.155	95.6	0.018	5.563	0.172	0.138	90.2	0.027
	5	-7.715	0.107	0.103	81.2	0.017	6.984	0.148	0.146	95.4	0.025	8.600	0.159	0.133	89.8	0.030
	10	-0.396	0.117	0.113	93.0	0.013	8.798	0.144	0.139	94.0	0.027	9.294	0.148	0.130	90.0	0.028
β_9	2	-22.361	0.089	0.086	30.6	0.059	0.510	0.144	0.154	95.6	0.019	5.276	0.169	0.138	90.6	0.028
	5	-7.864	0.112	0.103	80.0	0.019	6.866	0.153	0.145	94.8	0.026	8.556	0.165	0.133	87.6	0.032
	10	-0.868	0.119	0.112	91.2	0.014	8.241	0.145	0.138	92.8	0.027	8.762	0.149	0.129	89.4	0.029
β_{10}	2	-22.086	0.090	0.087	29.8	0.058	0.943	0.148	0.155	95.4	0.019	5.561	0.172	0.139	91.2	0.027
	5	-7.738	0.113	0.103	81.4	0.019	7.009	0.155	0.146	95.0	0.026	8.596	0.164	0.133	88.0	0.031
	10	-0.443	0.126	0.113	92.2	0.015	8.753	0.154	0.139	92.0	0.029	9.252	0.157	0.131	87.6	0.031
γ	2	-18.586	0.141	0.144	90.4	0.029	2.541	0.188	0.193	95.8	0.035	7.041	0.204	0.200	94.2	0.043
	5	-6.714	0.159	0.157	94.2	0.026	6.544	0.192	0.186	95.4	0.038	8.039	0.197	0.187	95.4	0.040
	10	0.243	0.166	0.164	94.0	0.028	8.672	0.185	0.181	94.2	0.036	9.080	0.187	0.181	94.0	0.037

Table A.7: Simulation results for the *column* parameters and covariate parameters with $p_x = 10$, $\sigma = 0.5$, E_{kr} is generated from the matrix normal distribution, and $n = 1000$

Parameter	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
		Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
β_1	2	-52.666	0.056	0.057	0.0	0.280	-28.457	0.103	0.121	30.8	0.093	-21.440	0.141	0.113	42.6	0.066
	5	-32.473	0.082	0.076	5.0	0.112	-7.924	0.144	0.148	84.0	0.027	-1.934	0.183	0.134	84.4	0.029
	10	-18.415	0.097	0.091	43.0	0.042	2.192	0.151	0.151	96.0	0.020	5.662	0.171	0.135	89.2	0.028
β_2	2	-52.429	0.040	0.042	0.0	0.071	-28.072	0.067	0.076	49.6	0.025	-21.224	0.085	0.076	60.4	0.019
	5	-32.211	0.052	0.053	17.4	0.030	-7.476	0.085	0.089	88.6	0.009	-1.494	0.102	0.086	88.4	0.009
	10	-18.340	0.064	0.061	60.2	0.013	2.495	0.095	0.092	95.4	0.008	5.874	0.105	0.087	91.6	0.010
β_3	2	-52.294	0.058	0.057	0.0	0.279	-27.878	0.107	0.122	34.4	0.092	-20.906	0.144	0.114	46.2	0.067
	5	-32.270	0.080	0.076	2.6	0.111	-7.693	0.139	0.148	84.6	0.025	-1.752	0.172	0.133	83.8	0.026
	10	-18.175	0.094	0.091	45.6	0.042	2.496	0.147	0.152	97.0	0.020	5.826	0.164	0.135	91.0	0.027
β_4	2	-52.379	0.056	0.057	0.0	0.279	-28.192	0.104	0.121	32.8	0.094	-21.047	0.148	0.114	43.0	0.068
	5	-32.030	0.081	0.076	4.2	0.109	-7.417	0.141	0.148	85.4	0.025	-1.264	0.177	0.133	84.2	0.027
	10	-18.075	0.094	0.091	46.4	0.041	2.580	0.147	0.152	96.6	0.019	6.024	0.166	0.136	89.8	0.026
β_5	2	-52.587	0.060	0.057	0.0	0.284	-28.256	0.111	0.122	31.4	0.096	-21.159	0.155	0.114	43.2	0.071
	5	-32.334	0.080	0.076	3.2	0.111	-7.729	0.141	0.148	86.6	0.025	-1.675	0.180	0.134	85.2	0.026
	10	-18.244	0.095	0.091	44.4	0.041	2.463	0.149	0.152	95.6	0.019	5.906	0.167	0.136	90.6	0.026
β_6	2	-52.342	0.059	0.057	0.0	0.278	-28.041	0.108	0.122	35.4	0.092	-21.189	0.140	0.113	46.0	0.066
	5	-32.359	0.081	0.076	4.6	0.110	-7.809	0.141	0.148	86.0	0.025	-1.868	0.176	0.133	86.6	0.027
	10	-18.040	0.094	0.091	46.0	0.042	2.638	0.147	0.152	97.0	0.020	6.077	0.167	0.135	89.4	0.028
β_7	2	-52.702	0.045	0.042	0.0	0.070	-28.498	0.077	0.076	47.2	0.026	-22.055	0.093	0.075	58.2	0.020
	5	-32.531	0.057	0.053	18.4	0.028	-8.047	0.091	0.089	85.8	0.009	-2.233	0.107	0.086	86.4	0.010
	10	-18.582	0.069	0.061	60.0	0.013	2.036	0.099	0.092	93.0	0.009	5.435	0.111	0.086	89.4	0.012
β_8	2	-52.643	0.057	0.057	0.0	0.282	-28.411	0.106	0.121	33.0	0.094	-21.462	0.147	0.113	42.2	0.067
	5	-32.484	0.077	0.076	3.6	0.110	-8.022	0.136	0.147	84.6	0.023	-2.009	0.174	0.133	83.6	0.024
	10	-18.253	0.091	0.091	43.6	0.040	2.505	0.143	0.152	97.0	0.018	5.971	0.162	0.135	92.4	0.026
β_9	2	-52.724	0.059	0.057	0.0	0.284	-28.556	0.108	0.121	32.4	0.098	-21.404	0.147	0.113	45.2	0.070
	5	-32.494	0.081	0.076	3.6	0.113	-7.928	0.139	0.147	85.6	0.026	-1.832	0.176	0.133	85.6	0.028
	10	-18.595	0.094	0.090	44.8	0.043	1.964	0.144	0.151	96.2	0.020	5.412	0.163	0.134	90.0	0.028
β_{10}	2	-52.535	0.058	0.057	0.0	0.281	-28.187	0.107	0.122	35.4	0.093	-21.184	0.147	0.114	45.0	0.066
	5	-32.471	0.080	0.076	4.2	0.112	-7.952	0.140	0.147	86.2	0.026	-2.083	0.174	0.132	82.8	0.026
	10	-18.220	0.099	0.091	44.2	0.043	2.571	0.156	0.153	95.4	0.022	6.015	0.174	0.136	89.6	0.030
γ	2	-45.479	0.112	0.115	47.4	0.066	-24.041	0.168	0.176	89.2	0.046	-16.506	0.200	0.194	91.4	0.051
	5	-28.972	0.134	0.134	79.6	0.038	-7.191	0.195	0.190	94.4	0.034	-1.634	0.220	0.202	92.6	0.039
	10	-16.182	0.148	0.147	88.2	0.028	2.534	0.194	0.190	95.0	0.036	5.755	0.209	0.196	94.0	0.042

Table A.8: Simulation results for the *column* parameters and covariate parameters with $p_x = 10$, $\sigma = 0.75$, E_{kr} is generated from the matrix normal distribution, and $n = 1000$

Parameter	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
		Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
β_1	2	-69.007	0.043	0.043	0.0	0.478	-49.632	0.081	0.088	0.4	0.254	-45.330	0.098	0.081	1.8	0.217
	5	-50.538	0.063	0.059	0.0	0.261	-26.463	0.118	0.124	37.8	0.086	-19.865	0.159	0.115	47.4	0.064
	10	-34.915	0.077	0.073	2.2	0.130	-10.144	0.136	0.145	80.8	0.028	-4.092	0.171	0.130	81.4	0.026
β_2	2	-68.763	0.032	0.033	0.0	0.120	-49.206	0.057	0.058	5.2	0.065	-45.125	0.065	0.058	9.4	0.056
	5	-50.317	0.042	0.043	0.0	0.066	-26.009	0.073	0.077	54.2	0.023	-19.267	0.091	0.077	66.2	0.018
	10	-35.040	0.052	0.052	12.0	0.034	-10.077	0.087	0.088	85.2	0.010	-4.023	0.106	0.085	85.8	0.009
β_3	2	-68.699	0.044	0.043	0.0	0.475	-49.108	0.083	0.089	0.8	0.250	-44.886	0.103	0.081	2.4	0.215
	5	-50.409	0.061	0.059	0.0	0.260	-26.336	0.113	0.124	40.4	0.085	-19.708	0.149	0.114	49.8	0.062
	10	-34.754	0.074	0.074	1.2	0.129	-9.959	0.130	0.145	83.4	0.027	-4.090	0.160	0.130	83.8	0.024
β_4	2	-68.712	0.043	0.043	0.0	0.477	-49.304	0.081	0.088	0.8	0.254	-45.066	0.103	0.081	2.0	0.218
	5	-50.134	0.061	0.059	0.0	0.258	-25.961	0.114	0.124	41.0	0.083	-19.119	0.155	0.115	48.2	0.060
	10	-34.648	0.074	0.074	1.8	0.130	-9.862	0.132	0.145	85.2	0.027	-3.891	0.164	0.130	84.4	0.024
β_5	2	-68.995	0.046	0.043	0.0	0.481	-49.603	0.085	0.088	1.2	0.257	-45.263	0.107	0.081	3.0	0.220
	5	-50.448	0.061	0.059	0.0	0.261	-26.332	0.113	0.124	39.0	0.085	-19.579	0.158	0.116	46.6	0.063
	10	-34.851	0.075	0.073	1.8	0.130	-10.035	0.133	0.145	82.0	0.027	-3.966	0.165	0.130	82.8	0.024
β_6	2	-68.722	0.046	0.043	0.0	0.475	-49.215	0.085	0.089	1.0	0.251	-45.056	0.100	0.081	0.8	0.216
	5	-50.546	0.062	0.059	0.0	0.261	-26.542	0.114	0.124	41.4	0.086	-19.957	0.153	0.115	48.4	0.064
	10	-34.618	0.075	0.074	1.2	0.128	-9.769	0.132	0.145	83.0	0.027	-3.676	0.168	0.130	83.4	0.025
β_7	2	-68.983	0.035	0.033	0.0	0.120	-49.594	0.062	0.058	5.2	0.066	-45.641	0.073	0.058	9.4	0.058
	5	-50.577	0.046	0.043	0.6	0.066	-26.571	0.078	0.077	51.0	0.024	-19.855	0.098	0.077	60.4	0.019
	10	-35.054	0.057	0.051	13.0	0.035	-10.301	0.092	0.088	83.4	0.011	-4.267	0.113	0.085	83.6	0.012
β_8	2	-69.019	0.043	0.043	0.0	0.479	-49.631	0.081	0.088	0.4	0.254	-45.488	0.101	0.080	1.6	0.219
	5	-50.574	0.059	0.059	0.0	0.261	-26.606	0.110	0.123	37.8	0.085	-19.904	0.155	0.115	46.0	0.063
	10	-34.895	0.073	0.074	1.2	0.130	-10.029	0.129	0.145	83.2	0.027	-3.949	0.160	0.130	84.0	0.023
β_9	2	-68.999	0.045	0.043	0.0	0.479	-49.675	0.084	0.088	0.8	0.255	-45.331	0.104	0.081	2.2	0.219
	5	-50.546	0.063	0.059	0.0	0.262	-26.426	0.115	0.124	39.4	0.086	-19.641	0.153	0.115	48.4	0.063
	10	-35.113	0.076	0.073	1.4	0.132	-10.385	0.131	0.144	81.0	0.028	-4.367	0.162	0.129	83.0	0.025
β_{10}	2	-68.909	0.043	0.043	0.0	0.477	-49.391	0.081	0.089	1.0	0.252	-45.131	0.101	0.081	2.6	0.216
	5	-50.558	0.061	0.059	0.0	0.260	-26.523	0.114	0.124	40.0	0.085	-19.998	0.152	0.114	47.2	0.063
	10	-34.833	0.078	0.074	2.0	0.130	-9.871	0.140	0.145	82.8	0.029	-3.753	0.176	0.131	82.4	0.027
γ	2	-59.099	0.095	0.098	15.8	0.098	-43.031	0.144	0.154	69.0	0.069	-37.993	0.169	0.167	76.4	0.066
	5	-44.715	0.116	0.117	52.0	0.064	-24.008	0.178	0.177	86.2	0.046	-17.515	0.218	0.196	89.0	0.051
	10	-31.124	0.132	0.131	77.6	0.042	-9.332	0.189	0.187	91.8	0.038	-3.616	0.218	0.201	92.6	0.047

Table A.9: Simulation results for the *row* parameters with $p_x = 20$, E_{kr} is generated from matrix normal distribution, and $n = 1000$

Parameter	σ	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
			Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
α_1	0.25	2	0.002	0.065	0.065	93.6	0.004	-0.006	0.067	0.064	92.8	0.005	-0.026	0.067	0.063	93.2	0.004
		5	-0.203	0.064	0.062	92.0	0.004	-0.183	0.065	0.061	92.0	0.004	-0.199	0.065	0.061	91.6	0.004
		10	-0.299	0.064	0.061	93.8	0.004	-0.288	0.064	0.060	93.8	0.004	-0.298	0.064	0.060	93.8	0.004
	0.5	2	0.377	0.076	0.076	93.4	0.006	0.295	0.082	0.073	91.0	0.007	0.463	0.079	0.073	91.0	0.006
		5	-0.180	0.069	0.067	93.6	0.005	-0.194	0.073	0.066	91.0	0.005	-0.144	0.072	0.065	90.0	0.005
		10	-0.369	0.067	0.064	93.6	0.004	-0.372	0.068	0.063	92.6	0.005	-0.388	0.068	0.062	92.4	0.005
	0.75	2	0.741	0.089	0.090	94.8	0.008	0.661	0.099	0.084	89.8	0.010	0.927	0.094	0.083	89.8	0.009
		5	-0.068	0.077	0.075	93.6	0.006	-0.181	0.084	0.072	89.2	0.007	-0.030	0.083	0.072	89.8	0.007
		10	-0.405	0.071	0.068	93.8	0.005	-0.470	0.074	0.067	91.4	0.006	-0.436	0.073	0.066	91.2	0.005
α_5	0.25	2	0.666	0.067	0.065	95.4	0.004	0.622	0.068	0.064	94.8	0.005	0.630	0.068	0.063	94.0	0.005
		5	0.864	0.066	0.062	94.4	0.004	0.832	0.066	0.061	93.8	0.004	0.863	0.066	0.061	93.4	0.004
		10	0.777	0.065	0.061	92.8	0.004	0.772	0.065	0.060	92.4	0.004	0.779	0.065	0.060	92.2	0.004
	0.5	2	0.772	0.079	0.076	94.8	0.006	0.668	0.083	0.073	92.2	0.007	0.847	0.083	0.072	91.4	0.007
		5	1.173	0.072	0.067	93.8	0.005	1.080	0.075	0.066	91.4	0.006	1.203	0.076	0.066	90.8	0.006
		10	0.875	0.069	0.064	92.8	0.005	0.862	0.071	0.063	92.2	0.005	0.883	0.071	0.062	91.4	0.005
	0.75	2	0.906	0.096	0.090	94.0	0.009	0.490	0.103	0.084	89.4	0.011	1.070	0.101	0.082	89.2	0.010
		5	1.545	0.080	0.075	93.8	0.007	1.459	0.088	0.072	90.0	0.008	1.579	0.087	0.072	89.4	0.008
		10	0.974	0.073	0.068	92.8	0.005	0.897	0.078	0.067	90.2	0.006	0.978	0.078	0.066	89.6	0.006
α_{15}	0.25	2	-0.437	0.066	0.065	94.2	0.004	-0.440	0.067	0.064	93.6	0.005	-0.382	0.067	0.063	93.6	0.005
		5	-0.031	0.064	0.062	92.6	0.004	0.003	0.065	0.061	92.0	0.004	0.050	0.065	0.061	91.0	0.004
		10	-0.366	0.062	0.061	93.6	0.004	-0.362	0.062	0.061	93.8	0.004	-0.343	0.062	0.060	93.8	0.004
	0.5	2	-0.489	0.076	0.076	93.8	0.006	-0.590	0.081	0.073	91.2	0.007	-0.507	0.081	0.072	91.2	0.007
		5	0.203	0.070	0.067	92.6	0.005	0.345	0.073	0.066	91.2	0.005	0.447	0.073	0.066	90.6	0.005
		10	-0.367	0.064	0.064	94.0	0.004	-0.343	0.065	0.063	93.4	0.004	-0.294	0.065	0.063	93.4	0.004
	0.75	2	-0.386	0.091	0.090	93.6	0.008	-0.773	0.098	0.084	89.6	0.010	-0.636	0.098	0.082	90.2	0.010
		5	0.398	0.077	0.075	92.8	0.006	0.679	0.084	0.072	90.2	0.007	0.717	0.084	0.072	88.6	0.007
		10	-0.377	0.068	0.068	94.4	0.005	-0.294	0.071	0.067	92.8	0.005	-0.225	0.071	0.066	92.4	0.005
α_{17}	0.25	2	0.249	0.068	0.065	93.0	0.005	0.184	0.069	0.063	91.6	0.005	0.105	0.069	0.063	91.8	0.005
		5	0.474	0.066	0.062	93.2	0.004	0.526	0.066	0.062	92.8	0.004	0.490	0.066	0.061	92.8	0.004
		10	0.505	0.065	0.061	93.8	0.004	0.524	0.065	0.061	93.6	0.004	0.496	0.065	0.061	93.4	0.004
	0.5	2	-0.051	0.081	0.076	94.2	0.006	-0.483	0.085	0.073	90.6	0.007	-0.486	0.084	0.072	89.6	0.007
		5	0.401	0.072	0.068	93.0	0.005	0.463	0.074	0.066	91.8	0.005	0.415	0.074	0.066	91.0	0.005
		10	0.581	0.068	0.064	94.0	0.005	0.617	0.069	0.063	93.0	0.005	0.509	0.068	0.063	92.4	0.005
	0.75	2	-0.232	0.096	0.090	92.4	0.009	-0.984	0.105	0.084	88.0	0.011	-0.821	0.102	0.082	87.0	0.010
		5	0.356	0.080	0.075	93.0	0.006	0.429	0.085	0.072	88.8	0.007	0.336	0.084	0.072	89.4	0.007
		10	0.698	0.073	0.068	93.8	0.005	0.682	0.075	0.067	90.8	0.006	0.503	0.074	0.066	91.2	0.006

Table A.10: Simulation results for the *column* parameters and covariate parameters with $p_x = 20$, E_{kr} is generated from matrix normal distribution, and $n = 1000$

Parameter	σ	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
			Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
β_1	0.25	2	-5.318	0.068	0.065	89.4	0.005	11.888	0.090	0.085	93.2	0.012	13.599	0.095	0.080	87.6	0.014
		5	4.042	0.074	0.070	94.4	0.006	12.894	0.086	0.080	90.0	0.012	13.225	0.086	0.077	87.8	0.012
		10	8.187	0.077	0.072	91.8	0.008	13.122	0.084	0.078	88.6	0.011	13.189	0.084	0.076	87.8	0.011
	0.5	2	-32.826	0.052	0.051	13.2	0.030	-4.330	0.085	0.085	92.6	0.008	1.835	0.100	0.080	88.8	0.010
		5	-13.194	0.064	0.061	75.4	0.008	8.560	0.090	0.086	95.2	0.010	11.434	0.095	0.081	88.8	0.012
		10	-2.197	0.070	0.066	92.4	0.005	12.799	0.089	0.084	89.8	0.012	13.935	0.092	0.079	86.6	0.013
	0.75	2	-52.880	0.042	0.041	0.0	0.072	-25.355	0.073	0.073	54.2	0.021	-19.527	0.087	0.069	58.6	0.017
		5	-30.474	0.056	0.052	20.8	0.026	-2.713	0.088	0.086	91.2	0.008	2.820	0.098	0.080	89.0	0.010
		10	-14.951	0.062	0.060	70.6	0.009	8.347	0.090	0.088	94.8	0.010	11.715	0.098	0.081	87.0	0.013
β_7	0.25	2	-6.380	0.070	0.065	88.4	0.006	10.599	0.093	0.084	91.6	0.011	12.269	0.097	0.080	88.0	0.013
		5	2.891	0.074	0.070	93.4	0.006	11.604	0.085	0.079	90.4	0.011	11.913	0.086	0.077	87.8	0.011
		10	6.492	0.076	0.072	92.6	0.007	11.308	0.083	0.077	90.2	0.010	11.365	0.083	0.076	88.6	0.010
	0.5	2	-33.324	0.054	0.051	14.0	0.031	-4.973	0.090	0.085	89.2	0.009	1.437	0.107	0.081	86.2	0.012
		5	-13.841	0.064	0.061	73.8	0.009	7.740	0.089	0.086	94.4	0.009	10.451	0.095	0.080	89.0	0.012
		10	-3.777	0.069	0.066	91.8	0.005	10.830	0.087	0.083	92.0	0.010	11.898	0.090	0.079	88.0	0.012
	0.75	2	-53.136	0.043	0.041	0.0	0.072	-25.644	0.077	0.073	52.6	0.022	-19.221	0.095	0.069	59.2	0.018
		5	-30.787	0.054	0.052	20.0	0.027	-3.091	0.086	0.086	93.4	0.008	2.215	0.097	0.080	89.0	0.010
		10	-16.325	0.061	0.060	68.2	0.010	6.393	0.087	0.087	95.8	0.009	9.594	0.096	0.081	89.4	0.011
β_{18}	0.25	2	-5.482	0.069	0.065	89.2	0.006	11.856	0.092	0.085	91.8	0.012	13.478	0.096	0.080	86.8	0.014
		5	3.740	0.072	0.070	93.6	0.006	12.599	0.083	0.080	90.6	0.011	12.899	0.084	0.077	88.4	0.011
		10	7.642	0.076	0.073	91.8	0.007	12.559	0.082	0.078	89.2	0.011	12.609	0.082	0.076	88.4	0.011
	0.5	2	-32.966	0.054	0.051	15.0	0.030	-4.212	0.090	0.086	90.2	0.009	1.946	0.108	0.081	87.2	0.012
		5	-13.472	0.062	0.061	74.8	0.008	8.308	0.086	0.087	95.4	0.009	11.191	0.092	0.081	89.4	0.012
		10	-2.843	0.069	0.067	93.0	0.005	12.067	0.088	0.084	91.8	0.011	13.156	0.090	0.080	87.2	0.013
	0.75	2	-53.059	0.043	0.041	0.2	0.072	-25.376	0.077	0.074	52.4	0.022	-19.330	0.095	0.069	58.6	0.018
		5	-30.767	0.053	0.052	16.8	0.026	-3.101	0.083	0.085	92.8	0.007	2.634	0.095	0.080	91.0	0.009
		10	-15.689	0.062	0.060	68.4	0.010	7.400	0.090	0.088	94.4	0.009	10.690	0.097	0.082	89.4	0.012
γ	0.25	2	-2.627	0.141	0.128	92.0	0.020	11.683	0.170	0.153	92.2	0.032	13.283	0.176	0.152	90.6	0.035
		5	5.026	0.147	0.134	93.2	0.022	12.518	0.162	0.146	92.4	0.030	12.766	0.162	0.145	91.4	0.030
		10	8.488	0.152	0.136	92.0	0.025	12.661	0.161	0.143	91.0	0.030	12.690	0.161	0.142	90.8	0.030
	0.5	2	-24.324	0.117	0.110	78.6	0.028	-1.557	0.166	0.158	93.8	0.027	4.628	0.192	0.163	91.6	0.038
		5	-9.146	0.131	0.123	93.6	0.019	8.871	0.167	0.156	93.4	0.030	11.595	0.176	0.156	91.2	0.034
		10	-0.090	0.144	0.129	92.8	0.021	12.340	0.169	0.151	92.0	0.032	13.367	0.172	0.150	91.0	0.034
	0.75	2	-39.361	0.100	0.097	48.6	0.049	-18.064	0.148	0.149	89.0	0.030	-11.826	0.177	0.155	89.2	0.035
		5	-22.959	0.116	0.111	80.2	0.027	-0.878	0.161	0.157	96.2	0.026	4.557	0.180	0.162	93.0	0.033
		10	-10.420	0.133	0.121	90.6	0.020	8.353	0.172	0.157	93.2	0.031	11.517	0.181	0.157	91.2	0.036

Table A.11: Simulation results for the *row* parameters with $p_x = 5$ and E_{kr} is generated from the matrix t-distribution

Parameter	σ	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
			Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
α_1	0.25	2	0.800	0.063	0.063	95.2	0.004	0.851	0.064	0.062	94.4	0.004	-57.518	6.518	106.238	93.8	42.570
		5	0.977	0.061	0.060	94.6	0.004	0.964	0.061	0.060	94.6	0.004	0.933	0.061	0.060	94.2	0.004
		10	0.908	0.061	0.059	93.4	0.004	0.887	0.061	0.059	93.6	0.004	0.925	0.062	0.059	93.6	0.004
	0.5	2	0.834	0.071	0.072	96.2	0.005	0.892	0.083	0.069	91.8	0.007	-10.776	1.328	0.367	92.4	1.767
		5	1.110	0.065	0.064	95.8	0.004	1.027	0.066	0.063	93.6	0.004	-10.047	1.234	7.733	93.0	1.525
		10	0.961	0.064	0.062	94.4	0.004	0.876	0.064	0.061	94.4	0.004	-8.662	1.064	0.237	94.0	1.135
	0.75	2	1.266	0.083	0.086	96.4	0.007	1.640	0.139	0.109	86.8	0.019	-80.595	6.507	75.014	90.0	42.508
		5	1.285	0.071	0.071	94.8	0.005	1.097	0.074	0.068	92.4	0.006	-17.377	1.580	27.154	92.0	2.503
		10	1.066	0.068	0.065	94.6	0.005	0.887	0.070	0.063	93.4	0.005	6.977	0.683	0.490	93.2	0.468
α_3	0.25	2	0.451	0.082	0.079	92.8	0.007	0.356	0.082	0.078	92.2	0.007	-32.012	7.242	117.808	92.0	52.547
		5	0.348	0.077	0.076	94.6	0.006	0.339	0.077	0.076	95.0	0.006	0.341	0.077	0.075	94.4	0.006
		10	0.280	0.077	0.075	93.8	0.006	0.278	0.077	0.075	93.8	0.006	0.253	0.077	0.075	93.8	0.006
	0.5	2	0.642	0.095	0.093	93.2	0.009	0.274	0.106	0.088	89.8	0.011	-6.398	1.555	0.441	90.8	2.421
		5	0.401	0.081	0.082	94.6	0.007	0.331	0.082	0.080	95.0	0.007	-14.692	3.354	21.097	94.2	11.274
		10	0.236	0.080	0.078	93.6	0.006	0.220	0.081	0.077	92.8	0.006	-3.376	0.803	0.187	92.2	0.647
	0.75	2	0.919	0.111	0.110	93.8	0.012	0.095	0.175	0.118	87.2	0.031	-44.541	7.075	83.471	89.2	50.255
		5	0.565	0.088	0.089	94.8	0.008	0.391	0.091	0.086	93.6	0.008	-37.394	6.355	108.463	91.6	40.530
		10	0.212	0.085	0.082	93.2	0.007	0.160	0.086	0.080	91.8	0.007	2.201	0.470	0.335	92.0	0.221
α_4	0.25	2	0.297	0.082	0.079	93.4	0.007	0.249	0.084	0.078	93.6	0.007	-40.525	9.128	148.741	93.4	83.490
		5	0.231	0.078	0.076	93.6	0.006	0.219	0.079	0.075	94.0	0.006	0.185	0.079	0.075	93.8	0.006
		10	0.208	0.078	0.075	92.8	0.006	0.214	0.078	0.075	93.0	0.006	0.276	0.080	0.075	92.6	0.006
	0.5	2	0.464	0.092	0.092	95.0	0.009	0.261	0.115	0.088	92.0	0.013	-3.865	0.990	0.286	91.0	0.982
		5	0.287	0.083	0.081	93.6	0.007	0.221	0.085	0.079	93.6	0.007	-3.914	0.934	5.783	93.0	0.873
		10	0.184	0.081	0.078	93.4	0.007	0.200	0.082	0.077	93.0	0.007	-14.752	3.333	0.725	92.0	11.129
	0.75	2	0.621	0.106	0.109	96.4	0.011	-0.013	0.209	0.139	87.8	0.044	-48.374	7.884	94.127	89.4	62.395
		5	0.401	0.091	0.089	94.2	0.008	0.226	0.096	0.086	91.6	0.009	-23.895	5.121	94.935	92.2	26.283
		10	0.194	0.085	0.082	92.8	0.007	0.207	0.089	0.080	92.2	0.008	21.815	4.841	3.370	91.4	23.482
α_5	0.25	2	0.525	0.081	0.080	96.4	0.007	0.463	0.082	0.078	94.4	0.007	-13.274	3.069	49.558	95.0	9.437
		5	0.395	0.076	0.076	96.2	0.006	0.404	0.076	0.076	96.2	0.006	0.426	0.076	0.075	96.0	0.006
		10	0.331	0.075	0.075	95.8	0.006	0.334	0.075	0.075	95.4	0.006	0.319	0.075	0.075	95.4	0.006
	0.5	2	0.692	0.096	0.094	95.4	0.009	0.503	0.120	0.088	90.6	0.014	2.029	0.348	0.146	90.8	0.121
		5	0.483	0.081	0.081	95.4	0.007	0.512	0.083	0.080	94.6	0.007	-19.266	4.424	28.006	93.6	19.613
		10	0.330	0.078	0.078	95.8	0.006	0.333	0.078	0.077	95.0	0.006	2.758	0.555	0.155	94.6	0.308
	0.75	2	0.738	0.112	0.112	95.2	0.013	0.155	0.245	0.148	87.8	0.060	-38.500	7.562	94.497	88.8	57.327
		5	0.639	0.089	0.089	96.0	0.008	0.698	0.096	0.086	93.0	0.009	-74.278	13.108	231.923	93.0	172.360
		10	0.390	0.083	0.082	95.0	0.007	0.397	0.083	0.080	93.8	0.007	3.302	0.674	0.493	92.8	0.455

Table A.12: Simulation results for the *column* parameters and covariate parameters with $p_x = 5$, and E_{kr} is generated from the matrix t-distribution

Parameter	σ	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
			Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
β_1	0.25	2	-7.712	0.097	0.089	81.4	0.015	2.657	0.114	0.109	93.6	0.014	4.197	0.299	12.283	89.8	0.091
		5	-1.664	0.098	0.093	92.4	0.010	3.164	0.106	0.101	93.2	0.012	3.362	0.106	0.098	92.6	0.012
		10	0.610	0.099	0.094	93.4	0.010	3.164	0.103	0.098	93.8	0.012	3.305	0.105	0.097	93.4	0.012
	0.5	2	-27.410	0.084	0.077	9.4	0.082	-4.163	0.128	0.147	90.8	0.018	-1.953	0.353	0.111	83.6	0.125
		5	-13.034	0.089	0.084	63.1	0.025	1.224	0.113	0.110	94.4	0.013	1.824	0.125	0.101	90.2	0.016
		10	-5.924	0.093	0.089	85.8	0.012	2.708	0.108	0.104	94.4	0.012	2.691	0.134	0.098	92.0	0.019
	0.75	2	-44.550	0.071	0.067	0.2	0.203	-15.745	0.189	13.524	66.9	0.060	-16.907	0.194	0.145	53.9	0.066
		5	-26.147	0.078	0.075	9.0	0.075	-4.423	0.115	0.117	90.6	0.015	-3.833	0.136	0.102	85.8	0.020
		10	-14.652	0.086	0.082	53.7	0.029	0.662	0.114	0.110	94.4	0.013	1.228	0.128	0.100	90.2	0.016
β_2	0.25	2	-8.190	0.061	0.062	87.4	0.005	1.998	0.071	0.072	94.0	0.005	3.399	0.134	4.437	92.8	0.018
		5	-2.173	0.063	0.065	94.4	0.004	2.582	0.067	0.069	95.4	0.005	2.767	0.067	0.068	94.8	0.005
		10	0.109	0.064	0.066	94.6	0.004	2.642	0.066	0.068	95.0	0.005	2.748	0.066	0.067	94.8	0.005
	0.5	2	-27.621	0.055	0.055	30.9	0.022	-4.760	0.080	0.085	90.6	0.007	0.344	0.441	0.075	88.8	0.194
		5	-13.387	0.058	0.060	78.0	0.008	0.663	0.072	0.074	96.2	0.005	1.196	0.076	0.070	93.2	0.006
		10	-6.280	0.061	0.063	89.8	0.005	2.304	0.069	0.070	95.2	0.005	2.369	0.074	0.068	94.6	0.006
	0.75	2	-44.606	0.048	0.048	0.8	0.052	-16.581	0.091	4.349	74.7	0.015	-16.567	0.126	0.091	67.5	0.023
		5	-26.308	0.053	0.055	32.5	0.020	-4.911	0.075	0.077	92.6	0.006	-4.184	0.084	0.072	90.2	0.008
		10	-14.895	0.058	0.059	72.5	0.009	0.351	0.074	0.073	94.6	0.005	0.885	0.078	0.069	93.2	0.006
β_3	0.25	2	-8.125	0.093	0.089	81.2	0.015	2.215	0.110	0.108	95.0	0.013	3.043	0.193	7.308	92.0	0.038
		5	-1.990	0.096	0.092	93.0	0.010	2.806	0.103	0.100	94.6	0.011	3.031	0.103	0.098	93.0	0.011
		10	0.250	0.099	0.094	94.0	0.010	2.784	0.103	0.098	93.8	0.011	2.943	0.106	0.097	92.2	0.012
	0.5	2	-27.809	0.079	0.076	7.8	0.084	-4.622	0.120	0.131	89.8	0.016	-4.177	0.135	0.109	85.2	0.020
		5	-13.294	0.085	0.084	60.3	0.025	0.856	0.107	0.110	95.4	0.011	1.484	0.119	0.101	91.6	0.014
		10	-6.295	0.094	0.089	84.8	0.013	2.263	0.108	0.103	94.8	0.012	2.239	0.143	0.098	92.0	0.021
	0.75	2	-44.931	0.066	0.066	0.0	0.206	-16.508	0.135	7.287	65.7	0.045	-18.277	0.156	0.129	50.5	0.058
		5	-26.303	0.076	0.075	8.4	0.075	-4.733	0.109	0.117	92.0	0.014	-4.035	0.132	0.103	86.4	0.019
		10	-15.017	0.087	0.082	51.1	0.030	0.162	0.113	0.109	95.0	0.013	0.866	0.129	0.099	90.0	0.017
β_4	0.25	2	-7.965	0.089	0.089	81.2	0.014	2.366	0.105	0.109	96.2	0.012	3.126	0.174	6.468	92.8	0.031
		5	-2.015	0.091	0.092	93.2	0.009	2.766	0.099	0.100	96.0	0.011	3.003	0.098	0.098	94.8	0.011
		10	0.405	0.094	0.094	95.8	0.009	2.957	0.098	0.098	96.6	0.010	3.109	0.100	0.097	95.8	0.011
	0.5	2	-27.664	0.077	0.076	7.0	0.082	-4.530	0.119	0.148	91.2	0.016	-6.099	0.400	0.108	85.8	0.163
		5	-13.444	0.081	0.084	60.3	0.025	0.609	0.104	0.109	96.6	0.011	1.262	0.114	0.100	93.8	0.013
		10	-6.063	0.089	0.089	86.0	0.012	2.583	0.105	0.104	97.2	0.012	2.462	0.146	0.099	95.0	0.022
	0.75	2	-44.797	0.066	0.067	0.2	0.205	-16.167	0.181	13.330	65.5	0.059	-18.528	0.193	0.127	50.3	0.072
		5	-26.525	0.072	0.075	6.6	0.076	-5.175	0.106	0.116	92.2	0.014	-4.451	0.126	0.103	88.8	0.018
		10	-14.757	0.084	0.082	52.7	0.029	0.596	0.114	0.110	97.0	0.013	1.191	0.127	0.100	93.4	0.016
β_5	0.25	2	-7.912	0.091	0.089	82.4	0.015	2.465	0.106	0.108	95.4	0.012	3.795	0.250	9.859	92.8	0.064
		5	-1.691	0.094	0.093	94.4	0.009	3.136	0.101	0.101	94.6	0.011	3.409	0.102	0.098	93.2	0.012
		10	0.577	0.095	0.094	95.4	0.009	3.144	0.100	0.099	95.0	0.011	3.253	0.101	0.097	94.2	0.011
	0.5	2	-27.650	0.080	0.076	7.0	0.083	-4.411	0.121	0.134	91.0	0.017	-2.982	0.262	0.112	84.8	0.070
		5	-13.066	0.084	0.084	61.7	0.024	1.165	0.107	0.110	95.4	0.011	1.915	0.124	0.102	92.2	0.016
		10	-5.942	0.091	0.089	86.0	0.012	2.747	0.105	0.104	95.0	0.012	2.790	0.122	0.099	92.6	0.016
	0.75	2	-44.794	0.068	0.066	0.0	0.205	-16.168	0.147	9.086	65.9	0.048	-17.895	0.166	0.118	51.3	0.059
		5	-26.134	0.075	0.076	9.8	0.074	-4.449	0.108	0.117	92.4	0.014	-3.751	0.134	0.110	88.8	0.019
		10	-14.687	0.085	0.082	54.5	0.029	0.735	0.113	0.110	94.0	0.013	1.277	0.126	0.100	90.6	0.016
γ	0.25	2	-5.858	0.112	0.110	92.6	0.013	1.638	0.122	0.122	95.2	0.015	2.940	0.165	5.709	93.4	0.028
		5	-1.226	0.112	0.113	94.2	0.013	2.335	0.116	0.118	94.6	0.014	2.520	0.117	0.118	94.2	0.014
		10	0.814	0.114	0.114	94.4	0.013	2.718	0.117	0.117	94.4	0.014	2.793	0.117	0.116	94.4	0.014
	0.5	2	-20.298	0.103	0.100	81.4	0.021	-4.354	0.129	0.134	94.0	0.017	-3.651	0.139	0.130	91.6	0.020
		5	-9.749	0.105	0.107	91.4	0.013	0.474	0.119	0.124	95.0	0.014	1.092	0.120	0.123	94.8	0.014
		10	-3.979	0.111	0.111	93.6	0.013	2.393	0.120	0.121	94.4	0.015	2.635	0.120	0.119	94.2	0.015
	0.75	2	-32.267	0.095	0.091	55.5	0.035	-13.815	0.131	3.989	91.4	0.022	-14.373	0.144	0.154	88.0	0.026
		5	-19.159	0.099	0.101	83.8	0.019	-4.042	0.120	0.129	95.0	0.015	-3.225	0.123	0.168	94.4	0.015
		10	-10.350	0.107	0.106	91.4	0.014	0.751	0.124	0.125	95.0	0.015	1.438	0.126	0.123	94.0	0.016

Table A.13: Simulation results for the *row* parameters with $p_x = 20$, E_{kr} is generated from the matrix normal distribution, and $n = 2000$

Parameter	σ	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
			Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
α_1	0.25	2	0.113	0.047	0.046	93.6	0.002	0.182	0.048	0.045	93.6	0.002	0.157	0.047	0.045	93.6	0.002
		5	-0.027	0.044	0.044	93.2	0.002	-0.005	0.045	0.043	92.8	0.002	0.000	0.045	0.043	92.4	0.002
		10	-0.076	0.045	0.043	93.4	0.002	-0.064	0.045	0.043	93.2	0.002	-0.073	0.045	0.043	93.0	0.002
	0.5	2	0.301	0.054	0.054	94.2	0.003	0.421	0.058	0.052	90.8	0.003	0.441	0.058	0.051	91.8	0.003
		5	-0.004	0.048	0.047	93.6	0.002	0.027	0.050	0.047	92.4	0.002	0.091	0.050	0.046	92.4	0.002
		10	-0.077	0.047	0.045	93.4	0.002	-0.057	0.048	0.045	92.8	0.002	-0.096	0.048	0.044	93.2	0.002
	0.75	2	0.419	0.063	0.064	94.8	0.004	0.522	0.070	0.059	89.0	0.005	0.707	0.069	0.058	90.2	0.005
		5	0.033	0.053	0.053	93.4	0.003	0.014	0.057	0.051	91.2	0.003	0.155	0.057	0.051	92.0	0.003
		10	-0.056	0.051	0.048	92.8	0.003	-0.071	0.053	0.047	91.2	0.003	-0.146	0.052	0.047	91.2	0.003
α_5	0.25	2	0.222	0.047	0.046	94.8	0.002	0.090	0.048	0.045	94.0	0.002	0.084	0.049	0.045	93.0	0.002
		5	0.074	0.045	0.044	94.0	0.002	0.032	0.045	0.043	93.6	0.002	0.030	0.045	0.043	93.4	0.002
		10	-0.120	0.044	0.043	93.6	0.002	-0.148	0.044	0.043	93.6	0.002	-0.142	0.044	0.043	93.4	0.002
	0.5	2	0.633	0.055	0.054	95.4	0.003	0.336	0.060	0.052	91.6	0.004	0.318	0.060	0.052	90.4	0.004
		5	0.266	0.049	0.048	94.6	0.002	0.189	0.050	0.047	92.8	0.002	0.152	0.050	0.046	92.2	0.003
		10	-0.144	0.046	0.045	93.8	0.002	-0.239	0.046	0.045	93.2	0.002	-0.220	0.047	0.044	93.0	0.002
	0.75	2	1.012	0.064	0.064	95.2	0.004	0.598	0.071	0.060	89.2	0.005	0.621	0.071	0.058	89.6	0.005
		5	0.462	0.054	0.053	93.8	0.003	0.410	0.057	0.051	91.8	0.003	0.365	0.057	0.051	91.6	0.003
		10	-0.121	0.049	0.048	94.6	0.002	-0.271	0.050	0.047	92.6	0.003	-0.240	0.050	0.047	92.2	0.003
α_{15}	0.25	2	0.153	0.048	0.046	93.0	0.002	0.160	0.048	0.045	93.6	0.002	0.140	0.048	0.045	93.6	0.002
		5	0.032	0.045	0.044	94.4	0.002	0.051	0.045	0.043	94.2	0.002	0.045	0.045	0.043	94.2	0.002
		10	-0.247	0.045	0.043	93.6	0.002	-0.245	0.045	0.043	93.4	0.002	-0.246	0.045	0.043	93.4	0.002
	0.5	2	0.304	0.057	0.054	92.2	0.003	0.331	0.059	0.052	91.0	0.004	0.254	0.058	0.051	91.6	0.003
		5	-0.024	0.048	0.048	93.8	0.002	0.018	0.049	0.047	92.8	0.002	0.019	0.049	0.046	93.2	0.002
		10	-0.514	0.047	0.045	93.8	0.002	-0.545	0.048	0.045	93.2	0.002	-0.540	0.048	0.044	92.8	0.002
	0.75	2	0.452	0.068	0.064	92.0	0.005	0.426	0.073	0.059	88.2	0.005	0.497	0.070	0.058	90.2	0.005
		5	-0.086	0.054	0.053	93.6	0.003	-0.070	0.056	0.051	91.4	0.003	0.013	0.055	0.051	92.0	0.003
		10	-0.711	0.050	0.048	95.0	0.003	-0.834	0.052	0.047	92.0	0.003	-0.817	0.052	0.047	91.6	0.003
α_{17}	0.25	2	0.242	0.044	0.046	94.8	0.002	0.139	0.045	0.045	94.8	0.002	0.155	0.045	0.045	94.6	0.002
		5	0.290	0.044	0.044	94.8	0.002	0.276	0.044	0.044	94.0	0.002	0.272	0.044	0.043	94.0	0.002
		10	0.152	0.043	0.043	95.4	0.002	0.149	0.043	0.043	95.6	0.002	0.145	0.043	0.043	95.6	0.002
	0.5	2	0.495	0.051	0.054	96.2	0.003	0.287	0.055	0.052	94.4	0.003	0.296	0.054	0.052	93.4	0.003
		5	0.511	0.048	0.048	93.6	0.002	0.436	0.049	0.047	93.0	0.002	0.446	0.049	0.046	93.0	0.002
		10	0.187	0.045	0.045	95.2	0.002	0.172	0.046	0.045	94.8	0.002	0.150	0.046	0.044	94.4	0.002
	0.75	2	0.704	0.060	0.064	96.4	0.004	0.556	0.066	0.060	93.6	0.004	0.561	0.064	0.058	92.8	0.004
		5	0.794	0.054	0.053	93.8	0.003	0.621	0.057	0.051	91.8	0.003	0.677	0.056	0.051	91.8	0.003
		10	0.229	0.049	0.048	95.0	0.002	0.163	0.050	0.047	93.6	0.002	0.133	0.050	0.047	94.4	0.002

Table A.14: Simulation results for the *column* parameters and covariate parameters with $p_x = 20$, E_{kr} is generated from the matrix normal distribution, and $n = 2000$

Parameter	σ	m	Naïve Estimator					Method 1 Estimator					Method 2 Estimator				
			Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE	Bias%	ESE	ASE	CR%	MSE
β_1	0.25	2	-10.595	0.042	0.043	74.2	0.005	3.374	0.052	0.053	95.2	0.003	4.240	0.053	0.051	91.8	0.003
		5	-1.839	0.048	0.046	93.4	0.002	5.265	0.054	0.051	92.8	0.004	5.466	0.054	0.050	91.4	0.004
		10	1.417	0.047	0.047	95.2	0.002	5.272	0.050	0.050	93.0	0.003	5.335	0.050	0.049	92.6	0.003
	0.5	2	-35.970	0.034	0.034	0.2	0.034	-11.690	0.052	0.053	77.0	0.006	-8.514	0.056	0.049	80.0	0.005
		5	-17.336	0.042	0.041	43.6	0.009	0.944	0.057	0.054	94.6	0.003	2.555	0.059	0.051	91.6	0.004
		10	-7.845	0.044	0.044	81.2	0.003	4.151	0.052	0.053	94.4	0.003	4.822	0.053	0.050	93.0	0.003
	0.75	2	-54.839	0.029	0.028	0.0	0.076	-30.599	0.048	0.046	14.2	0.026	-27.578	0.052	0.043	16.8	0.022
		5	-33.423	0.036	0.035	1.6	0.029	-9.366	0.055	0.054	80.8	0.005	-6.271	0.059	0.050	83.2	0.004
		10	-19.378	0.039	0.040	32.4	0.011	-0.190	0.053	0.054	96.0	0.003	1.683	0.056	0.051	93.8	0.003
β_7	0.25	2	-9.698	0.046	0.043	74.8	0.004	4.378	0.056	0.054	92.4	0.004	5.233	0.057	0.051	90.0	0.004
		5	-1.327	0.049	0.046	93.4	0.002	5.800	0.054	0.051	92.8	0.004	6.006	0.055	0.050	91.2	0.004
		10	2.022	0.050	0.047	94.6	0.003	5.899	0.053	0.050	91.8	0.004	5.961	0.053	0.049	91.2	0.004
	0.5	2	-35.004	0.036	0.035	1.4	0.032	-10.349	0.055	0.054	78.4	0.006	-7.291	0.060	0.050	80.2	0.005
		5	-16.922	0.044	0.041	43.6	0.009	1.417	0.058	0.055	94.2	0.003	3.022	0.059	0.051	91.6	0.004
		10	-7.287	0.047	0.044	83.8	0.003	4.777	0.056	0.053	94.0	0.004	5.452	0.057	0.051	91.0	0.004
	0.75	2	-53.947	0.029	0.028	0.0	0.074	-29.202	0.049	0.047	14.6	0.024	-26.254	0.053	0.043	18.8	0.020
		5	-33.083	0.038	0.035	0.8	0.029	-8.926	0.056	0.054	81.8	0.005	-5.870	0.060	0.050	82.8	0.004
		10	-18.885	0.042	0.040	36.2	0.011	0.398	0.057	0.054	94.4	0.003	2.294	0.059	0.051	92.2	0.004
β_{18}	0.25	2	-10.403	0.042	0.043	72.4	0.004	3.574	0.053	0.053	95.6	0.003	4.373	0.054	0.051	93.0	0.003
		5	-1.981	0.044	0.046	94.8	0.002	5.063	0.049	0.051	95.6	0.003	5.244	0.049	0.050	93.6	0.003
		10	1.421	0.046	0.047	95.6	0.002	5.260	0.049	0.050	95.0	0.003	5.313	0.049	0.049	94.0	0.003
	0.5	2	-35.660	0.033	0.035	0.0	0.033	-11.242	0.052	0.053	77.0	0.006	-8.275	0.056	0.050	79.6	0.005
		5	-17.452	0.038	0.041	43.0	0.009	0.647	0.051	0.054	96.4	0.003	2.161	0.053	0.051	95.2	0.003
		10	-7.761	0.043	0.044	82.6	0.003	4.190	0.052	0.053	95.8	0.003	4.840	0.053	0.051	92.2	0.003
	0.75	2	-54.537	0.027	0.028	0.0	0.075	-30.127	0.047	0.047	15.6	0.025	-27.181	0.051	0.043	19.8	0.021
		5	-33.454	0.033	0.035	0.6	0.029	-9.587	0.050	0.054	82.2	0.005	-6.682	0.053	0.050	84.8	0.004
		10	-19.223	0.040	0.040	33.0	0.011	-0.105	0.054	0.054	94.4	0.003	1.740	0.057	0.051	91.6	0.003
γ	0.25	2	-6.621	0.090	0.086	92.4	0.009	4.834	0.103	0.100	92.6	0.011	5.540	0.105	0.099	92.0	0.012
		5	0.133	0.097	0.089	92.4	0.009	5.936	0.104	0.096	91.6	0.012	6.122	0.104	0.095	91.0	0.012
		10	3.142	0.098	0.091	92.6	0.010	6.347	0.101	0.094	92.4	0.011	6.404	0.102	0.094	91.8	0.011
	0.5	2	-26.726	0.076	0.075	56.6	0.024	-7.680	0.102	0.103	94.0	0.012	-4.815	0.109	0.103	93.2	0.012
		5	-12.416	0.090	0.083	85.6	0.012	2.209	0.109	0.101	92.4	0.012	3.783	0.113	0.101	90.4	0.013
		10	-4.370	0.094	0.087	92.2	0.009	5.557	0.106	0.099	92.4	0.012	6.217	0.107	0.098	91.2	0.012
	0.75	2	-40.955	0.066	0.067	14.2	0.046	-22.922	0.093	0.098	77.6	0.022	-19.894	0.101	0.098	79.2	0.020
		5	-25.012	0.082	0.076	61.6	0.022	-6.290	0.109	0.103	92.4	0.013	-3.041	0.117	0.103	90.6	0.014
		10	-13.691	0.088	0.082	84.0	0.013	1.926	0.108	0.102	93.8	0.012	3.868	0.113	0.101	92.4	0.013

Appendix B

Technical Components for Chapter 3

B.1 Regularity Conditions

The following standard regularity conditions are required for the establishment of the asymptotic results for the estimators described in Chapter 3.

(C.1) The expectations of the first derivative of $U^c(\theta; Y_k^c)$, $U^o(\theta; Y_k^*)$ and $U_v^o(\eta)$ with respect to θ exist and are not singular at θ_0.

(C.2) The second derivatives of $U^c(\theta; Y_k^c)$, $U^o(\theta; Y_k^*)$ and $U_v^o(\eta)$ with respect to θ exist and are continuous and bounded in a neighborhood of θ_0.

(C.3) The expectation of the first derivative of $\delta_k S_k(\phi)$ with respect to ϕ exists and is not singular at ϕ_0.

(C.4) The second derivatives of $U^o(\theta; Y_k^*)$ and $U_v^o(\eta)$ with respect to ϕ exist and are continuous and bounded in a neighborhood of ϕ_0.

B.2 Proof of Theorem 3.1

Noting that $\hat{\theta}_c$ is the solution of (3.6), i.e., $\sum_{k=1}^{n} U^c(\hat{\theta}_c; Y_k^c) = 0$, we apply the first-order Taylor series expansion to the equation around θ_0:

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}} \sum_{k=1}^{n} U^c(\hat{\theta}_c; Y_k^c) \\ &= \frac{1}{\sqrt{n}} \sum_{k=1}^{n} U^c(\theta_0; Y_k^c) + \Big\{ \frac{1}{n} \sum_{k=1}^{n} \frac{\partial U^c(\theta_0; Y_k^c)}{\partial \theta^\intercal} \Big\} \Big\{ \sqrt{n}(\hat{\theta}_c - \theta_0) \Big\} \\ &\quad + o_p(\sqrt{n} \|\hat{\theta}_c - \theta_0\|). \end{aligned} \tag{B.1}$$

By Condition (C.2) and the Central Limit Theorem, we have that

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} U^c(\theta_0; Y_k^c) \xrightarrow{d} N(0, \Sigma_c) \text{ as } n \to \infty, \tag{B.2}$$

thus, $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} U^c(\theta_0; Y_k^c) = O_p(1)$, where $\Sigma_c = \mathrm{E}\{U^c(\theta_0; Y_k^c) U^c(\theta_0; Y_k^c)^\intercal\}$. By Condition (C.1) and the Law of Large Numbers, we obtain that

$$\frac{1}{n} \sum_{k=1}^{n} \frac{\partial U^c(\theta_0; Y_k^c)}{\partial \theta^\intercal} \xrightarrow{p} \Gamma_c \text{ as } n \to \infty, \tag{B.3}$$

thus, $\frac{1}{n} \sum_{k=1}^{n} \frac{\partial U^c(\theta_0; Y_k^c)}{\partial \theta^\intercal} = O_p(1)$, where $\Gamma_c = \mathrm{E}\{\partial U^c(\theta_0; Y_k^c)/\partial \theta^\intercal\}$. Then (B.1) shows that

$$O_p(1) + O_p(1) \times \sqrt{n} \|\hat{\theta}_c - \theta_0\| + o_p(\sqrt{n} \|\hat{\theta}_c - \theta_0\|) = O_p(1), \tag{B.4}$$

implying that $\|\hat{\theta}_c - \theta_0\|$ is of order $O_p(\frac{1}{\sqrt{n}})$.

Combining (B.1), (B.2) and (B.3), then by the Slutsky's Theorem, we obtain that

$$\sqrt{n}(\hat{\theta}_c - \theta_0) \xrightarrow{d} N(0, \Gamma_c \Sigma_c (\Gamma_c^{-1})^\intercal) \text{ as } n \to \infty.$$

B.3 Proof of Theorem 3.2.

Noting that $\hat{\eta}_v = (\hat{\phi}_v^\intercal, \hat{\theta}_v^\intercal)^\intercal$ is the solution of (3.8), we apply the first-order Taylor series expansion around η_0 to (3.8) with η replaced by $\hat{\eta}_v$:

$$\sqrt{n}\begin{pmatrix}\hat{\theta}_v - \theta_0\\ \hat{\phi}_v - \phi_0\end{pmatrix} = -\begin{pmatrix}\frac{1}{n}\sum\limits_{k=1}^{n}\partial U^c(\theta_0,\phi_0;Y_k^c)/\partial\theta^\intercal & \frac{1}{n}\sum\limits_{k=1}^{n}\partial U^c(\theta_0,\phi_0;Y_k^c)/\partial\phi^\intercal\\ 0 & \frac{1}{n}\sum\limits_{k=1}^{n}\delta_k\times\partial S_k(\phi_0)/\partial\phi^\intercal\end{pmatrix}^{-1} \times\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\begin{pmatrix}U^c(\theta_0,\phi_0;Y_k^c)\\ \delta_k S_k(\phi_0)\end{pmatrix} + o_p(1). \tag{B.5}$$

By Conditions (C.1)-(C.4), applying the Central Limit Theorem to the right-hand-side of (B.5) leads to the asymptotic distribution of $\sqrt{n}(\hat{\eta}_v - \eta_0)$ as

$$\sqrt{n}(\hat{\eta}_v - \eta_0) \xrightarrow{d} N(0, \Gamma_U^{-1}(\eta_0)\Sigma_U(\eta_0)[\Gamma_U^{-1}(\eta_0)]^\intercal) \text{ as } n\to\infty,$$

where

$$\Gamma_U(\eta_0) = \begin{pmatrix}E\{\partial U^c(\theta_0,\phi_0;Y_k^c)/\partial\theta^\intercal\} & E\{\partial U^c(\theta_0,\phi_0;Y_k^c)/\partial\phi^\intercal\}\\ 0 & E\{\delta_k\times\partial S_k(\phi_0)/\partial\phi^\intercal\}\end{pmatrix}$$

and

$$\Sigma_U(\eta_0) = E\left\{\begin{pmatrix}U^c(\theta_0,\phi_0;Y_k^c)\\ \delta_k S_k(\phi_0)\end{pmatrix}\begin{pmatrix}U^c(\theta_0,\phi_0;Y_k^c)\\ \delta_k S_k(\phi_0)\end{pmatrix}^\intercal\right\}.$$

Since θ is of primary interest, we explicitly express the asymptotic distribution of the estimator $\hat{\theta}_v$ by calculating the product of the corresponding block matrices:

$$\sqrt{n}(\hat{\theta}_v - \theta_0) \xrightarrow{d} N(0, \Gamma_c^{-1}\Sigma_\tau[\Gamma_c^{-1}]^\intercal) \text{ as } n\to\infty,$$

where $\Gamma_c = \mathrm{E}\left\{\frac{\partial U^c(\theta_0,\phi_0;Y_k^c)}{\partial\theta^\intercal}\right\}$, $\Sigma_\tau = \mathrm{E}\{\Omega_k(\theta_0,\phi_0)\Omega_k(\theta_0,\phi_0)^\intercal\}$ and

$$\begin{aligned}\Omega_k(\theta_0,\phi_0) = U^c(\theta_0,\phi_0;Y_k^c) &- \mathrm{E}\Big\{\partial U^c(\theta_0,\phi_0;Y_k^c)/\partial\phi\Big\}\\ &\times\Big[\mathrm{E}\Big\{\partial\delta_k S_k(\phi_0)/\partial\phi\Big\}\Big]^{-1}\times\{\delta_k S_k(\phi_0)\}.\end{aligned}$$

Appendix C

Technical Components for Chapter 4

C.1 Regularity Conditions

(C.1) Assume that $\lambda_n \to 0$, $a_n = O(\frac{1}{\sqrt{n}})$ and $b_n \to 0$ as $n \to \infty$.

(C.2) The expectations of $\partial U_k^c(\theta; Y_k^c)/\partial\theta$, $\partial U_k^o(\theta; Y_k^*)/\partial\theta$ and $\partial U_k^{ov}(\eta)/\partial\eta$ exist and are not singular at θ_0.

(C.3) $\partial^2 U_k^c(\theta; Y_k^c)/\partial\theta\partial\theta^\intercal$, $\partial^2 U_k^o(\theta; Y_k^*)/\partial\theta\partial\theta^\intercal$ and $\partial^2 U_k^{ov}(\eta)/\partial\eta\partial\eta^\intercal$ exist and are continuous and bounded in a neighborhood of θ_0.

(C.4) The variance-covariance matrices of $U_{k,\mathrm{I}}^c(\theta; Y_k^c)$, $U_{k,\mathrm{I}}^o(\theta; Y_k^*)$ and $U_{k,\mathrm{I}}^{ov}(\eta)$ are positive definite at θ_0.

(C.5) The expectations of $\partial\{\delta_k S_k(\phi)\}/\partial\phi$, $\partial U_k^o(\theta; Y_k^*)/\partial\phi$ and $\partial U_k^{ov}(\eta)/\partial\phi$ exist and are not singular at ϕ_0.

(C.6) $\partial^2\{\delta_k S_k(\phi)\}/\partial\phi\partial\phi^\intercal$, $\partial^2 U_k^o(\theta; Y_k^*)/\partial\phi\partial\phi^\intercal$ and $\partial^2 U_k^{ov}(\eta)/\partial\phi\partial\phi^\intercal$ exist and are continuous and bounded in a neighborhood of ϕ_0.

C.2 Proof of Theorem 4.1

Here, we show that given Conditions (C.1)-(C.5), there exists a solution to (4.1), $\hat{\theta}_c$, such that $\|\hat{\theta}_c - \theta_0\| = O_p(\frac{1}{\sqrt{n}} + a_n)$, if a_n and b_n tend to 0 as $n \to \infty$. We adapt the techniques of Ma and Li (2010) to do this.

Define

$$J_1 = \left[E\left\{ \frac{\partial U_k^c(\theta; Y_k^c)}{\partial \theta^\intercal} \right\}\Big|_{\theta_0} \right]^{-1}, \quad \phi_k^*(\theta) = J_1 U_k^c(\theta; Y_k^c), \text{ and } q'_{\lambda_n}(\theta) = J_1 p'_{\lambda_n}(\theta).$$

Write $\alpha_n = \frac{1}{\sqrt{n}} + a_n$ and $U_k^c(\theta) = U_k^c(\theta; Y_k^c)$. Then we consider

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \phi_k^*(\theta) - \sqrt{n} q'_{\lambda_n}(\theta) = 0. \tag{C.1}$$

To show Theorem 4.1, it suffices to show that $\hat{\theta}_c$ is a solution for (C.1) that satisfies $\|\hat{\theta}_c - \theta_0\| = O_p(\alpha_n)$ by the Brouwer fixed-point theorem.

Given Conditions (C.1) and (C.5), for any θ with $\|\theta - \theta_0\| = C\alpha_n$ for some positive constant C, we apply the first-order Taylor-series expansion to left-hand-side of (C.1) around θ_0 and obtain that

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \phi_k^*(\theta) - \sqrt{n} q'_{\lambda_n}(\theta) \\
&= \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \phi_k^*(\theta_0) - \sqrt{n} q'_{\lambda_n}(\theta_0) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{\partial \phi_k^*(\theta_0)}{\partial \theta^\intercal} (\theta - \theta_0)\{1 + o_p(1)\} - \sqrt{n} \frac{\partial q'_{\lambda_n}(\theta_0)}{\partial \theta^\intercal} (\theta - \theta_0)\{1 + o_p(1)\},
\end{aligned} \tag{C.2}$$

By Condition (C.5), we have that

$$\begin{aligned}
&\sqrt{n}(\theta - \theta_0)^\intercal \left\{ \frac{1}{n} \sum_{k=1}^{n} \frac{\partial \phi_k^*(\theta_0)}{\partial \theta^\intercal} \right\} (\theta - \theta_0)\{1 + o_p(1)\} \\
&= \sqrt{n}(\theta - \theta_0)^\intercal J_1 \left\{ \frac{1}{n} \sum_{k=1}^{n} \frac{\partial U_k^c(\theta_0)}{\partial \theta^\intercal} \right\} (\theta - \theta_0)\{1 + o_p(1)\} \\
&= \sqrt{n}\|\theta - \theta_0\|^2 \{1 + o_p(1)\},
\end{aligned} \tag{C.3}$$

where the second step is due to the definition of $\phi_k^*(\cdot)$ and the last step is due to the definition of J_1 and the law of large numbers.

Furthermore, we have that

$$\begin{aligned}
&(\theta-\theta_0)^\intercal\sqrt{n}\left\{\frac{\partial q'_{\lambda_n}(\theta_0)}{\partial\theta^\intercal}\right\}(\theta-\theta_0)\{1+o_p(1)\}\\
&=(\theta-\theta_0)^\intercal\sqrt{n}J_1p''_{\lambda_n}(\theta_0)(\theta-\theta_0)\{1+o_p(1)\}\\
&=(\theta-\theta_0)^\intercal\sqrt{n}O(b_n)(\theta-\theta_0)\{1+o_p(1)\}\\
&=o_p(\sqrt{n}\|\theta-\theta_0\|^2),
\end{aligned}\tag{C.4}$$

where the first step is due to the definition of $q'_{\lambda_n}(\theta)$, the second step is due to Condition (C.5) and the definition of b_n, and the last step is from Condition (C.1).

Then combining (C.2), (C.3) and (C.4), we obtain that for θ with $\|\theta-\theta_0\|=C\alpha_n$,

$$\begin{aligned}
&(\theta-\theta_0)^\intercal\Big\{\frac{1}{\sqrt{n}}\sum_{k=1}^n\phi_k^*(\theta)-\sqrt{n}q'_{\lambda_n}(\theta)\Big\}\\
&=(\theta-\theta_0)^\intercal\Big\{\frac{1}{\sqrt{n}}\sum_{k=1}^n\phi_k^*(\theta_0)-\sqrt{n}q'_{\lambda_n}(\theta_0)\Big\}+\sqrt{n}\|\theta-\theta_0\|^2\{1+o_p(1)\}+o_p(\sqrt{n}\|\theta-\theta_0\|^2)\\
&=(\theta-\theta_0)^\intercal\times\text{Constant}+\sqrt{n}C^2\alpha_n^2+o_p(\sqrt{n}\|\theta-\theta_0\|^2)\\
&=O_p(C\alpha_n)+\sqrt{n}C^2\alpha_n^2+o_p(\sqrt{n}C^2\alpha_n^2).
\end{aligned}\tag{C.5}$$

As long as C is large enough, the second term in (C.5) dominates the first and third terms in (C.5). Thus, for any $\epsilon>0$, as long as C is large enough, we have

$$P\Big[(\theta-\theta_0)^\intercal\Big\{\frac{1}{\sqrt{n}}\sum_{k=1}^n\phi_k^*(\theta)-\sqrt{n}q'_{\lambda_n}(\theta)\Big\}>0\Big]\geq 1-\epsilon,$$

where $\|\theta-\theta_0\|=C\alpha_n$. By the Brouwer fixed-point theorm, with probability at least $1-\epsilon$, there exists at least one solution, $\hat{\theta}_c$ for (C.1) that satisfies $\|\hat{\theta}_c-\theta_0\|=O_p(\alpha_n)$.

C.3 Lemma 4.1 and the Proof.

Lemma 4.1. *Let* $U^c_{\mathrm{II}}(\theta) = \sum\limits_{k=1}^{n} U^c_{k,\mathrm{II}}(\theta; Y^c_k)$*, where* $U^c_{k,\mathrm{II}}(\theta; Y^c_k)$ *is defined in Section 4.1. If the conditions in Theorem 4.1 hold, then with the probability tending to one, for any* θ *satisfying* $\|\theta - \theta_0\| = O(\frac{1}{\sqrt{n}})$*, we have that* $\theta_{\mathrm{II}} = 0$ *are the solutions to* $U^c_{\mathrm{II}}(\theta) = 0$*, where* $\theta_{\mathrm{II}} = (\tilde{\alpha}^{\intercal}_{\mathrm{II}}, \beta^{\intercal}_{\mathrm{II}}, \gamma^{\intercal}_{\mathrm{II}})^{\intercal}$ *is subvector of* θ *defined in Section 4.1.*

Let $d_\theta = (d_\alpha + d_\beta + d_\gamma)$ and $d_{2\theta} = (d_{2\alpha} + d_{2\beta} + d_{2\gamma})$. For $j = 1, ..., d_{2\theta}$, $U^c_{\mathrm{II}j}(\theta)$ denote the jth equation in $U^c_{\mathrm{II}}(\theta)$ and let $\theta_{\mathrm{II}j}$ denote the jth component of θ_{II}. Then applying the first-order Taylor series expansion to $U^c_{\mathrm{II}j}(\theta) - np'_{\lambda_n}(\theta_{\mathrm{II}j})$ around $\theta_{\mathrm{II}0}$, we obtain that

$$\begin{aligned} U^c_{\mathrm{II}j}(\theta) - np'_{\lambda_n}(\theta_{\mathrm{II}j}) = U^c_{\mathrm{II}j}(\theta_0) &+ \sum_{k=1}^{d_\theta} \frac{\partial U^c_{\mathrm{II}j}(\theta)}{\partial \theta_k}(\theta_k - \theta_{k0}) \\ &+ \frac{1}{2}\sum_{k=1}^{d_\theta}\sum_{l=1}^{d_\theta} \frac{\partial^2 U^c_{\mathrm{II}j}(\theta^*)}{\partial\theta_k \partial\theta_l}(\theta_k - \theta_{k0})(\theta_l - \theta_{l0}) - np'_{\lambda_n}(|\theta_{\mathrm{II}j}|)\mathrm{sign}(\theta_{\mathrm{II}j}), \end{aligned} \tag{C.6}$$

where θ^* lies between θ and θ_0.

Now we examine the terms on the right hand side of (C.6) using the assumption that $\|\theta - \theta_0\| = O(\frac{1}{\sqrt{n}})$. The first term has order $O_p(\sqrt{n})$ by Condition (C.4), the second term has order $O_p(\sqrt{n})$ due to Condition (C.6), and the third term has order $O_p(\sqrt{n})$ by Condition (C.5). Hence (C.6) becomes

$$U^c_{\mathrm{II}j}(\theta) - np'_{\lambda_n}(\theta_{\mathrm{II}j}) = -\sqrt{n}\{\sqrt{n}p'_{\lambda_n}(|\theta_{\mathrm{II}j}|)\mathrm{sign}(\theta_{\mathrm{II}j}) + O_p(1)\}.$$

By Condition (C.1), λ_n is sufficiently small, $a_n = O(\frac{1}{\sqrt{n}})$ when n is large enough, then $\sqrt{n}p'_{\lambda_n}(|\theta_{\mathrm{II}j}|) = \infty$ by Condition (4.5). Thus the sign of $U^c_{\mathrm{II}j}(\theta) - np'_{\lambda_n}(\theta_{\mathrm{II}j})$ is decided by the negative of $\mathrm{sign}(\theta_{\mathrm{II}j})$. By the continuity of $U^c_{\mathrm{II}j}(\theta) - np'_{\lambda_n}(\theta_{\mathrm{II}j})$, we obtain that it is zero at $\theta_{\mathrm{II}j} = 0$.

C.4 Proof of Theorem 4.2

Theorem 4.2(a) comes from Lemma 4.1 immediately. To show Theorem 4.2(b), let $U_{\mathrm{I}}^c(\theta_{\mathrm{I}})$ denote the $\sum_{k=1}^{n} U_{k,\mathrm{I}}^c\{\theta; Y_k^c\}$, where $U_{k,\mathrm{I}}^c\{\theta; Y_k^c\}$ is defined in Section 4.1. Then we we apply the Taylor series expansion to

$$0 = U_{\mathrm{I}}^c(\hat{\theta}_{c,\mathrm{I}}) - np'_{\lambda_n,\mathrm{I}}(\hat{\theta}_{c,\mathrm{I}}),$$

and obtain that

$$\begin{aligned}
0 &= U_{\mathrm{I}}^c(\theta_{\mathrm{I}0}) + \left\{\frac{\partial U_{\mathrm{I}}^c(\theta_{\mathrm{I}0})}{\partial \theta_{\mathrm{I}}^{\mathsf{T}}} + o_p(n)\right\}(\hat{\theta}_{c\mathrm{I}} - \theta_{\mathrm{I}0}) - ng_\theta - n\{\Sigma_\theta + o_p(1)\}(\hat{\theta}_{c\mathrm{I}} - \theta_{\mathrm{I}0}) \\
&= U_{\mathrm{I}}^c(\theta_{\mathrm{I}0}) + n\left[E\left\{\frac{\partial U_{k,\mathrm{I}}^c(\theta_{\mathrm{I}0}; Y_k^c)}{\partial \theta^{\mathsf{T}}}\right\} - \Sigma_\theta\right](\hat{\theta}_{c\mathrm{I}} - \theta_{\mathrm{I}0}) \\
&\quad - n\left[E\left\{\frac{\partial U_{k,\mathrm{I}}^c(\theta_{\mathrm{I}0}; Y_k^c)}{\partial \theta^{\mathsf{T}}}\right\} - \Sigma_\theta\right] \times \left[E\left\{\frac{\partial U_{k,\mathrm{I}}^c(\theta_{\mathrm{I}0}; Y_k^c)}{\partial \theta^{\mathsf{T}}}\right\} - \Sigma_\theta\right]^{-1} g_\theta + o_p(\sqrt{n}) \\
&= U_{\mathrm{I}}^c(\theta_{\mathrm{I}0}) + n\left[E\left\{\frac{\partial U_{k,\mathrm{I}}^c(\theta_{\mathrm{I}0}; Y_k^c)}{\partial \theta^{\mathsf{T}}}\right\} - \Sigma_\theta\right]\left[\hat{\theta}_{c\mathrm{I}} - \theta_{\mathrm{I}0} - \left[E\left\{\frac{\partial U_{k,\mathrm{I}}^c(\theta_{\mathrm{I}0}; Y_k^c)}{\partial \theta^{\mathsf{T}}}\right\} - \Sigma_\theta\right]^{-1} g_\theta\right] \\
&\quad + o_p(\sqrt{n})
\end{aligned}$$

where in the first step, the first two terms and the last two terms, respectively, come from the Taylor series expansion of $U_{\mathrm{I}}^c(\hat{\theta}_{c,\mathrm{I}})$ and of $np'_{\lambda_n,\mathrm{I}}(\hat{\theta}_{c,\mathrm{I}})$ with g_θ and Σ_θ defined after Theorem 1.

Consequently,

$$\begin{aligned}
&\sqrt{n}\left[\hat{\theta}_{c,\mathrm{I}} - \theta_{\mathrm{I}0} - \left[E\left\{\frac{\partial U_{k,\mathrm{I}}^c(\theta_{\mathrm{I}0}; Y_k^c)}{\partial \theta^{\mathsf{T}}}\right\} - \Sigma_\theta\right]^{-1} g_\theta\right] \\
&\quad = -n^{-1/2}\left[E\left\{\frac{\partial U_{k,\mathrm{I}}^c(\theta_{\mathrm{I}0}; Y_k^c)}{\partial \theta^{\mathsf{T}}}\right\} - \Sigma_\theta\right]^{-1} U_{\mathrm{I}}^c(\theta_{\mathrm{I}0}) + o_p(1),
\end{aligned}$$

and Theorem 4.2(b) thus follows from the central limit theorem together with Condition (C.6).

C.5 Proof of Theorem 4.3.

We show Theorem 4.3 using similar techniques in Appendix C.2. Let $\eta = (\theta^\intercal, \phi^\intercal)^\intercal$. Combining (4.1) and (3.7), this two-stage estimation procedure can be expressed as a single procedure for ease of establishing the asymptotic results of the resulting estimator, $\hat{\eta}_v$. Solving

$$\sum_{k=1}^{n} \begin{pmatrix} U_{1k}^c(\theta, \phi; Y_k^c) - np'_{\lambda_n}(\tilde{\alpha}) \\ U_{2k}^c(\theta, \phi; Y_k^c) - np'_{\lambda_n}(\beta) \\ U_{3k}^c(\theta, \phi; Y_k^c) - np'_{\lambda_n}(\gamma) \\ \delta_k S_k(\phi) \end{pmatrix} = 0, \tag{C.7}$$

gives a consistent estimator for η, say $\hat{\eta}_v = (\hat{\theta}_v^\intercal, \hat{\phi}_v^\intercal)^\intercal$, provided regularity conditions. Then, we obtain

$$\sqrt{n}(\hat{\theta}_v - \theta_0) = -\sqrt{n}\left[E\Big\{\frac{\partial U_k^c(\theta_0, \phi_0; Y_k^c)}{\partial \theta^\intercal}\Big\} - \Sigma_\theta^*\right]^{-1} E\Big\{U_k^{*c}(\theta_0, \phi_0; Y_k^c) - p'_{\lambda_n}(\theta)\Big\} + o_p(1),$$

where $\Sigma_\theta^* = \text{diag}\{p''_{\lambda_n}(\tilde{\alpha}), p''_{\lambda_n}(\beta), p''_{\lambda_n}(\gamma)\}$,

$$U_k^{*c}(\theta, \phi; Y_k^c) = U_k^c(\theta, \phi; Y_k^c) - \\ \Big\{\frac{1}{n}\sum_{k=1}^{n} \partial U_k^c(\theta, \phi; Y_k^c)/\partial\phi\Big\} \times \Big[\frac{1}{n}\Big\{\sum_{k=1}^{n} \delta_k \times \partial S_k(\phi)/\partial\phi\Big\}\Big]^{-1} \times \{\delta_k S_k(\phi)\}.$$

Now we denote

$$J_2 = \Big[E\Big\{\frac{\partial U_k^{*c}(\theta_0, \phi_0; Y_k^c)}{\partial \theta^\intercal}\Big\} - \Sigma_\theta^*\Big]^{-1} \quad \text{and} \quad q'_{\lambda_n}(\theta) = J_2 p'_{\lambda_n}(\theta),$$

and define

$$\phi_k^{*c}(\theta) = J_2 \times U_k^{*c}(\theta, \phi; Y_k^c),$$

where we treat ϕ as fixed. Denote $\alpha_n = \frac{1}{\sqrt{n}} + a_n$.

Then to prove Theorem 4.3, it suffices to show that

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} \phi_k^{*c}(\theta) - \sqrt{n} q'_{\lambda_n}(\theta) = 0. \tag{C.8}$$

has a solution, $\hat{\theta}_v$ that satisfies $\|\hat{\theta}_v - \theta_0\| = O_p(\alpha_n)$.

For any θ with $\|\theta - \theta_0\| = C\alpha_n$ for some positive constant C, given Conditions (C.1) and (C.5), we apply the first-order Taylor series expansion to (C.7) around θ_0 and obtain that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\phi_k^{*c}(\theta) - \sqrt{n}q'_{\lambda_n}(\theta) \\
=& \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\phi_k^{*c}(\theta_0) - \sqrt{n}q'_{\lambda_n}(\theta_0) \\
&+ \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{\partial\phi_k^{*c}(\theta_0)}{\partial\theta^{\intercal}}(\theta-\theta_0)\{1+o_p(1)\} - \sqrt{n}\frac{\partial q'_{\lambda_n}(\theta_0)}{\partial\theta^{\intercal}}(\theta-\theta_0)\{1+o_p(1)\}.
\end{aligned} \tag{C.9}$$

By Condition (C.5), we have that

$$\begin{aligned}
&\sqrt{n}(\theta-\theta_0)^{\intercal}\Big\{\frac{1}{n}\sum_{k=1}^{n}\frac{\partial\phi_k^{*c}(\theta_0)}{\partial\theta^{\intercal}}\Big\}(\theta-\theta_0)\{1+o_p(1)\} \\
=& \sqrt{n}(\theta-\theta_0)^{\intercal}J_2\Big\{\frac{1}{n}\sum_{k=1}^{n}\frac{\partial U_k^{*c}(\theta_0,\phi_0;Y_k^c)}{\partial\theta^{\intercal}}\Big\}(\theta-\theta_0)\{1+o_p(1)\} \\
=& \sqrt{n}(\theta-\theta_0)^{\intercal}\Big[E\Big\{\frac{\partial U_k^{*c}(\theta_0,\phi_0;Y_k^c)}{\partial\theta^{\intercal}}\Big\}+o_p(1)\Big]^{-1}\Big\{\frac{1}{n}\sum_{k=1}^{n}\frac{\partial U_k^{*c}(\theta_0,\phi_0;Y_k)}{\partial\theta^{\intercal}}\Big\}(\theta-\theta_0)\{1+o_p(1)\} \\
=& \sqrt{n}\|\theta-\theta_0\|^2\{1+o_p(1)\}.
\end{aligned} \tag{C.10}$$

Furthermore, we have that

$$\begin{aligned}
&(\theta-\theta_0)^{\intercal}\sqrt{n}\left\{\frac{\partial q'_{\lambda_n}(\theta_0)}{\partial\theta^{\intercal}}\right\}(\theta-\theta_0)\{1+o_p(1)\} \\
=& (\theta-\theta_0)^{\intercal}\sqrt{n}J_2p''_{\lambda_n}(\theta_0)(\theta-\theta_0)\{1+o_p(1)\} \\
=& (\theta-\theta_0)^{\intercal}\sqrt{n}O(b_n)(\theta-\theta_0)\{1+o_p(1)\} \\
=& o_p(\sqrt{n}\|\theta-\theta_0\|^2)
\end{aligned} \tag{C.11}$$

Then combining (C.9), (C.10) and (C.11), we obtain that

$$
\begin{aligned}
&(\theta-\theta_0)^\intercal\Big\{\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\phi_k^{*c}(\theta)-\sqrt{n}q'_{\lambda_n}(\theta)\Big\}\\
&=(\theta-\theta_0)^\intercal\Big\{\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\phi_k^{*c}(\theta_0)-\sqrt{n}q'_{\lambda_n}(\theta_0)\Big\}+\sqrt{n}\|\theta-\theta_0\|^2\{1+o_p(1)\}+o_p(\sqrt{n}\|\theta-\theta_0\|^2)\\
&=(\theta-\theta_0)^\intercal\times\text{Constant}+\sqrt{n}C^2\alpha_n^2+o_p(\sqrt{n}\|\theta-\theta_0\|^2)\\
&=O_p(C\alpha_n)+\sqrt{n}C^2\alpha_n^2+o_p(\sqrt{n}C^2\alpha_n^2).
\end{aligned}
\tag{C.12}
$$

As long as C is large enough, the second terms in (C.12) dominates the first and third terms in (C.12). Thus, for any $\epsilon > 0$, as long as C is large enough, the probability of (C.12) larger than zero is at least $1-\epsilon$. By the Brouwer fixed-point theorem, with probability at least $1-\epsilon$, there exists at least one solution, $\hat{\theta}_v$ for (C.7) that satisfies $\|\hat{\theta}_v-\theta_0\| = O_p(\alpha_n)$.

C.6 Proof of Theorem 4.4

Lemma 4.2 *Let $U^c_{\mathrm{II}}(\theta,\phi)$ denote $\sum_{k=1}^{n} U^c_{k,\mathrm{II}}\{\theta,\phi;Y^c_k\}$, where $U^c_{k,\mathrm{II}}\{\theta,\phi;Y^c_k\}$ is defined before Theorem 4.4 in Section 4.2. If the conditions in Theorem 4.4 hold, then with the probability tending to one, for any θ satisfying $\|\theta-\theta_0\| = O(\frac{1}{\sqrt{n}})$, we have that $\theta_{\mathrm{II}} = 0$ are the solutions to $U^c_{\mathrm{II}}(\theta,\phi) = 0$.*

The proof is similar to that of Lemma 4.1 in Appendix C.3 with $U^c_{\mathrm{II}j}(\theta)$ replaced by $U^c_{\mathrm{II}j}(\theta,\phi)$.

Theorem 4.4(a) comes from Lemma 4.2 immediately. Then, similar to the proof of Theorem 4.2(b) in Appendix D, based on Condition (C.6), we apply the Taylor series expansion to $0 = \sum_{k=1}^{n} U^c_{k,\mathrm{I}}(\theta_{\mathrm{I}0},\phi_0;Y^c_k) - p'_{\lambda_n,\mathrm{I}}(\theta_{\mathrm{I}})$ to obtain the Theorem 4.4(b).

Appendix D

Technical Components for Chapter 5

D.1 Full Conditional Distribution of Hyperparameters

As we defined in (5.6), the prior distribution for hyperparameters, λ_{α_i}, λ_{β_i}, λ_{γ_i} and a are half-Cauchy distribution. It suffices to show the full conditional distribution of λ_{γ_i} only, and all full conditional distribution of other hyperparameters can be derived using same techniques:

$$\begin{aligned}\pi(\lambda_{\alpha_i}|\alpha,\beta,\gamma,a) &= \pi(\lambda_{\alpha_i}|\alpha_i,a)\\ &\propto \pi(\lambda_{\alpha_i})\times\pi(\alpha_i|\lambda_{\alpha_i},a)\\ &\propto \frac{2}{\pi}\frac{1}{1+\lambda_{\alpha_i}^2}\times\exp(-\frac{\alpha_i^2}{2\lambda_{\alpha_i}^2a^2}).\end{aligned}$$

The density function of $\pi(\lambda_{\alpha_i}|\alpha,\beta,\gamma,a)$ cannot be identified as any known distribution. Thus, Slice-sampling algorithm (Polson et al. 2014) is used to generate λ_{α_i} as we present in Section 5.2.3.

D.2 Full Conditional Distribution of $\alpha^{(r)}$

As we claimed in (5.5), the full conditional distribution of $\alpha^{(r)}$ is

$$\begin{aligned}
\pi(\alpha^{(r)}|w,\beta^{(r)},\mathcal{B}_{-r},\{\mathbb{Y},x\}) &\propto \Big\{\prod_{k=1}^{n} P(Y_k=y_k|\mathcal{B})\Big\} f(w|\mathcal{B})\pi(\alpha^{(r)}|\beta^{(r)},\mathcal{B}_{-r}) \\
&\propto \prod_{k=1}^{n}\Big\{\frac{\exp(<x_k,\mathfrak{B}>)^{y_k}}{1+\exp(<x_k,\mathfrak{B}>)}\Big\}\cosh\Big(\frac{|<x_k,\mathfrak{B}>|}{2}\Big) \\
&\quad \times \exp\Big\{-\frac{(<x_k,\mathfrak{B}>)^2 w_k}{2}\Big\}\pi(\alpha^{(r)}|\lambda_{\alpha^{(r)}},a) \\
&= 2^{-n}\pi(\alpha^{(r)}|\lambda_{\alpha^{(r)}},a)\prod_{k=1}^{n}\exp\Big\{y_k(<x_k,\mathfrak{B}>)-\frac{<x_k,\mathfrak{B}>}{2} \\
&\quad -\frac{(<x_k,\mathfrak{B}>)^2 w_k}{2}\Big\} \\
&\propto \exp\Big\{-\frac{1}{2}\alpha^{(r)\intercal}\Sigma^{-1}_{\alpha^{(r)}}\alpha^{(r)}+\sum_{k=1}^{n}\Big(y_k-\frac{1}{2}\Big)\alpha^{(r)\intercal}x_k\beta^{(r)} \\
&\quad -\frac{(\alpha^{(r)\intercal}x_k\beta^{(r)})^2}{2}w_k-\alpha^{(r)\intercal}x_k\beta^{(r)}\Big(<x_k,\mathfrak{B}_{-r}>\Big)w_k\Big\} \\
&= \exp\Big[-\frac{1}{2}\alpha^{(r)\intercal}\Sigma^{-1}_{\alpha^{(r)}}\alpha^{(r)}-\frac{1}{2}\alpha^{(r)\intercal}x^{\intercal}_{\beta^{(r)}}\Omega(w)x_{\beta^{(r)}}\alpha^{(r)} \\
&\quad +x_{\beta^{(r)}}\Big\{y-\frac{1}{2}\mathbf{1}_n-x_{\mathcal{B}_{-r}}(w)\Big\}\alpha^{(r)}\Big] \\
&= \exp\Big[-\frac{1}{2}\alpha^{(r)\intercal}\Big\{x^{\intercal}_{\beta^{(r)}}\Omega(w)x_{\beta^{(r)}}+\Sigma^{-1}_{\alpha^{(r)}}\Big\}\alpha^{(r)}+x_{\beta^{(r)}}y(w)\alpha^{(r)}\Big]
\end{aligned} \tag{D.1}$$

where the third step is from the fact that $\cosh(\mathrm{u})=\frac{1+\exp(2\mathrm{u})}{2\exp(\mathrm{u})}$, $x_{\beta^{(r)}}=(x_1\beta^{(r)},...,x_n\beta^{(r)})^{\intercal}$, $y=(y_1,...,y_n)^{\intercal}$, $y(w)=y-\frac{1}{2}\mathbf{1}_n-x_{\mathcal{B}_{-r}}(w)$, $x_{\mathcal{B}_{-r}}(w)=\{(<x_1,\mathfrak{B}_{-r}>)w_1,...,(<x_n,\mathfrak{B}_{-r}>)w_n\}^{\intercal}$, $\mathbf{1}_n$ is an $n\times 1$ unit vector, $\Omega(w)=diag(w)$ and $\Sigma_{\alpha^{(r)}}=diag(\lambda^2_{\alpha^{(r)}}a^2)$. We can observe that (D.1) is the kernel of a multivariate normal with mean $m_{\alpha^{(r)}}(w)$ and covariance $\Sigma_{\alpha^{(r)}}(w)$ such that

$$\begin{aligned}
m_{\alpha^{(r)}}(w) &= \Sigma_{\alpha^{(r)}}(w)x_{\beta^{(r)}}y(w), \\
\Sigma_{\alpha^{(r)}}(w) &= \Big\{x^{\intercal}_{\beta^{(r)}}\Omega(w)x_{\beta^{(r)}}+\Sigma^{-1}_{\alpha^{(r)}}\Big\}^{-1}.
\end{aligned}$$

D.3 Full Conditional Distribution of $\beta^{(r)}$

The full conditional distribution of $\beta^{(r)}$ is

$$\begin{aligned}
\pi(\beta^{(r)}|w,\alpha,\{\mathbb{Y},x\}) &\propto \Big\{\prod_{k=1}^{n} P(Y_k = y_k|\mathcal{B})\Big\} f(w|\mathcal{B})\pi(\beta^{(r)}|\alpha^{(r)},\mathcal{B}_{-r}) \\
&\propto \prod_{k=1}^{n} \Big\{\frac{\exp(< x_k, \mathfrak{B} >)^{y_k}}{1+\exp(< x_k, \mathfrak{B} >)}\Big\} \cosh\Big(\frac{| < x_k, \mathfrak{B} > |}{2}\Big) \\
&\quad \times \exp\Big\{-\frac{(< x_k, \mathfrak{B} >)^2 w_k}{2}\Big\}\pi(\beta^{(r)}|\lambda_{\beta^{(r)}}, a) \\
&= 2^{-n}\pi(\beta^{(r)}|\lambda_{\beta^{(r)}}, a)\prod_{k=1}^{n}\exp\Big\{y_k(< x_k, \mathfrak{B} >) - \frac{< x_k, \mathfrak{B} >}{2} \\
&\quad - \frac{(< x_k, \mathfrak{B} >)^2 w_k}{2}\Big\} \\
&\propto \exp\Big\{-\frac{1}{2}\beta^{(r)\intercal}\Sigma^{-1}_{\beta^{(r)}}\beta^{(r)} + \sum_{k=1}^{n}\Big(y_k - \frac{1}{2}\Big)\alpha^{(r)\intercal}x_k\beta^{(r)} \\
&\quad - \frac{(\alpha^{(r)\intercal}x_k\beta^{(r)})^2}{2}w_k - \alpha^{(r)\intercal}x_k\beta^{(r)}\Big(< x_k, \mathfrak{B}_{-r} > \Big)w_k\Big\} \\
&= \exp\Big[-\frac{1}{2}\beta^{(r)\intercal}\Sigma^{-1}_{\beta^{(r)}}\beta^{(r)} - \frac{1}{2}\beta^{(r)\intercal}x^{\intercal}_{\alpha^{(r)}}\Omega(w)x_{\alpha^{(r)}}\beta^{(r)} \\
&\quad + x_{\alpha^{(r)}}\Big\{y - \frac{1}{2}\mathbf{1}_n - x_{\mathcal{B}_{-r}}(w)\Big\}\beta^{(r)}\Big] \\
&= \exp\Big[-\frac{1}{2}\beta^{(r)\intercal}\Big\{x^{\intercal}_{\alpha^{(r)}}\Omega(w)x_{\alpha^{(r)}} + \Sigma^{-1}_{\beta^{(r)}}\Big\}\beta^{(r)} + x_{\alpha^{(r)}}y(w)\beta^{(r)}\Big]
\end{aligned}\tag{D.2}$$

where $x_{\alpha^{(r)}} = (x_1^{\intercal}\alpha^{(r)}, ..., x_n^{\intercal}\alpha^{(r)})^{\intercal}$, and $\Sigma_{\beta^{(r)}} = diag(\lambda^2_{\beta^{(r)}}a^2)$. We can observe that (D.2) is the kernel of a multivariate normal with mean $m_{\beta^{(r)}}(w)$ and covariance $\Sigma_{\beta^{(r)}}(w)$ such that

$$\begin{aligned}
m_{\beta^{(r)}}(w) &= \Sigma_{\beta^{(r)}}(w)x_{\alpha^{(r)}}y(w), \\
\Sigma_{\beta^{(r)}}(w) &= \Big\{x^{\intercal}_{\alpha^{(r)}}\Omega(w)x_{\alpha^{(r)}} + \Sigma^{-1}_{\beta^{(r)}}\Big\}^{-1}.
\end{aligned}$$