Dimension Reduction in Statistical Modeling

by

Linquan Ma

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)

at the
University of Wisconsin-Madison
2022

Date of Final Oral Exam: 11/30/2022

The dissertation is approved by the following members of the Final Committee:

Hyunseung Kang, Assistant Professor, Statistics
Lan Liu, Associate Professor, Statistics
Anru Zhang, Associate Professor, Statistics
Lu Mao, Associate Professor, Biostatistics
Guanhua Chen, Associate Professor, Biostatistics

Abstract

When the data object is described by a large number of features, it is often beneficial to reduce the dimension of the data, so that the statistical analysis can have better efficiencies. Recently, a new dimension reduction method called the envelope method by Cook, Li, and Chiaromonte (2010) has been proposed in the multivariate regressions. It has the potential to gain substantial efficiency over the standard least squares estimator.

Chapter 2 proposes an approach to use envelope method when the predictors and/or the responses are missing at random. When there exists missing data, the envelope method using the complete case observations may lead to biased and inefficient results. We incorporate the envelope structure in the expectation-maximization (EM) algorithm. Our method is guaranteed to be more efficient, or at least as efficient as, the standard EM algorithm. We give asymptotic properties of our method under both normal and non-normal cases.

Chapter 3 extends the envelope model to the mixed effects model for longitudinal data with possibly unbalanced design and time-varying predictors. We show that our model provides more efficient estimators than the standard estimators in mixed effects models.

Chapter 4 proposes a semiparametric variant of the inner envelope model (Su and Cook, 2012) that does not rely on the linear model nor the normality assumption. We show that our proposal leads to globally and locally efficient estimators of the inner envelope spaces. We also present a computationally tractable algorithm to estimate the inner envelope.

The instrumental variables (IV) are frequently used in observational studies to recover the effect of exposure in the presence of unmeasured confounding. A key fact is that the strength of IV matters: an IV with a stronger association with the exposure results in a more accurate estimation of a causal effect. While it is hard to find a stronger IV, we generalize a sufficient dimension method to remove immaterial IVs. Chapter 5 investigates two different ways of incorporating the envelope method into IV regression. We show that the first stage envelope method does not yield any efficiency gain on the standard IV estimator, however, it may reduce the finite sample bias. The second stage envelope can achieve substantial efficiency gain under certain condition.

Contents

List of Figures

List of Tables

Acknowledgements

I'm grateful to have Hyunseung and Lan as my two advisors.

I started doing research in Spring 2017. At that time, I was still a junior student at USTC. I emailed Lan asking if I could spend a summer at UMN doing research with her. She gracefully accepted. After five and half years of hard work, we are working on our sixth paper. At the very beginning, I barely have any exposure in statistics, not to mention the skills for doing research. Lan taught me hand by hand about causal inference, the envelope model, and principles for doing excellent research. It is always a great pleasure working with Lan.

If I search "Hyunseung" in my mailbox, there are over three hundred emails. After each meeting, Hyunseung will send out an email to summarize the meeting, which made my life a lot easier. Hyunseung is always passionate and optimistic about work and life. Some of his energy and attitude toward life passed on me, which makes me a better researcher, and a better person. I'm grateful to Hyunseung for giving me freedom to choose research topics and encouraging and supporting me to go ahead on whichever road I pick.

I'm thankful to my thesis committee members: Prof. Anru Zhang, Prof. Lu Mao, and Prof. Guanhua Chen for the helpful advice and suggestions. I'm thankful to the faculty members for teaching me courses that are important for my Ph.D. research, including Prof. Jun Shao, Prof. Wei-Yin Loh, and Prof. Zhengjun Zhang.

Thanks to my internship managers and mentors, Yu Zhang and Kun Zhou from Amazon; Francesco Mina, Canyao Liu and Zhe Wang from Hudson River Trading. Also thanks to my fellow HRT intern friends. I had a wonderful summer at the 76th floor of 3WTC.

Thanks to Rui Chen, Peng Yu and Peigen Zhou. We discuss how to invest in the stock market every day and lose money. Thanks to my friends who helped me along my journey: Shuai Shao, Muxuan Liang, Tai Qin, Bin Guo, Siyu Wang, Zhenxuan Chen, Dongxue Du, Yuhui Li and Qijun Zhang. Life would be boring without my

friends.

Last but not least, I want to thank my wife Jiatong and my parents, for their endless love and support.

To my wife and parents.

Chapter 1

Introduction

With the advancement of sciences and technologies, industry and scientific data has the tendency to grow in both size and complexity. One characteristic of the complexity lies in the sheer amount of available covariates, which makes it difficult to detect the relationship between covariates and responses.

For example, consider the following multivariate regression model

$$Y_i = \beta^T X_i + \varepsilon_i, \tag{1.1}$$

where ε_i identically and independently (i.i.d) follows $N(0, \Sigma)$, $Y_i \in \mathbb{R}^r$ and $X_i \in \mathbb{R}^p$. The parameter of interest is the regression coefficient $\beta \in \mathbb{R}^{p \times r}$. In the absence of any additional assumptions, classical results show that the OLS is asymptotically efficient (Casella and Berger, 2021), defined as

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y,$$

where $X = (X_1, \ldots, X_n)^T$ and $Y = (Y_1, \ldots, Y_n)^T$.

When there are multiple predictors and responses, usually there exists a more parsimonious model than the full model. There are mainly two different approaches in the statistical literature to find a more parsimonious model. One is variable selection, where the researcher believes that only a few covariates are truly related to the responses, and hence the other covariates can be omitted. In particular, AIC (Akaike, 1974) and BIC (Schwarz et al., 1978) are two widely used variable selection methods. The other approach is dimension reduction. In contrast to variable selection, dimension reduction assumes the response variables only relates to a linear combination of the covariates. Therefore, it is possible that all the

covariates have explanatory power, but the effect is only represented in several linear combinations. The goal of dimension reduction is to successfully identify those linear combinations.

A well-known dimension reduction approach under Model 1.1 is the partial least squares (PLS) regression, which can be traced back to Wold (1966). For many decades, PLS algorithms such as SIMPLS (De Jong, 1993) have been widely used in many applied sciences such as chemometrics, bioinformatics, econometrics and genetics.

The population SIMPLS algorithm produces a sequence of p dimensional vectors $w_1, \ldots, w_k, \ldots \in \mathbb{R}^p$ with initial vector $w_0 = 0$. Given the first k vectors $w_1, \ldots, w_k$, the next direction is obtained by solving the constrained optimization problem

$$\begin{aligned} w_{k+1} = \underset{w}{\operatorname{argmax}}\, & w^T \Sigma_{XY} \Sigma_{YX} w, \text{ subject to} \\ & w^T \Sigma_X W_k = 0 \text{ and } w^T w = 1, \end{aligned} \tag{1.2}$$

where $W_k = (w_1, \ldots, w_k) \in \mathbb{R}^{p \times k}$. By definition, the objective function $w_k^T \Sigma_{XY} \Sigma_{YX} w_k$ monotonically decreases as the algorithm proceeds. Because for any non-zero vector w we have $w^T \Sigma_{XY} \Sigma_{YX} w \geq 0$, this algorithm stops after d steps when $w^T \Sigma_{XY} \Sigma_{YX} w = 0$ for any $w \neq 0$ that satisfies the constraints in (1.2).

In finite samples, we use sample version of the covariance matrix $S_{XY} S_{YX}$ instead of $\Sigma_{XY} \Sigma_{YX}$ and the algorithm stops when the objective function is close to 0. We use $\hat{W}_d$ to denote the PLS estimator. Then, the relationship between X and Y can be obtained by regressing Y on $\hat{W}_d^T X$. That is,

$$\hat{\beta}_{PLS} = \hat{W}_d (\hat{W}_d^T S_X \hat{W}_d)^{-1} \hat{W}_d^T S_{XY}.$$

Cook et al. (2013) built a connection between PLS and the envelope, and showed that the fundamental goal of PLS is to estimate the predictor envelope in a linear model. The predictor envelope considers the following conditions under Model (1.1).

Condition 1.0.1 *$Cov(Y, \Gamma_0^T X \mid \Gamma^T X) = 0$,*

Condition 1.0.2 *$Cov(\Gamma^T X, \Gamma_0^T X) = 0$,*

where $\Gamma \in \mathbb{R}^{p \times d}$ is a semi-orthogonal matrix and $\Gamma_0 \in \mathbb{R}^{p \times (p-d)}$ is its orthogonal complement. Condition 1.0.1 implies that Y relates to X only through $\Gamma^T X$. Condition 1.0.2 implies that $\Gamma_0^T X$ is uncorrelated with $\Gamma^T X$. Together, the two conditions

indicate that $\Gamma_0^T X$ has no correlation with $(\Gamma^T X, Y)$. One can always find an orthogonal matrix (Γ, Γ_0) that satisfies the above two conditions. A trivial choice is $\Gamma = I_p$, where I_p is the identity matrix of size p. The smallest space of Γ satisfying Conditions 1.0.1–1.0.2 is called the predictor envelope. The number of columns of Γ, d, is called the envelope dimension.

Under Conditions 1.0.1–1.0.2, Model (1.1) can be reparametrized as

$$Y_i = \eta^T \Gamma^T X_i + \varepsilon_i, \;\; \Sigma_X = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T, \tag{1.3}$$

where $\eta \in \mathbb{R}^{d \times r}$, $\Omega = \Gamma^T \Sigma_X \Gamma$ and $\Omega_0 = \Gamma_0^T \Sigma_X \Gamma_0$. Once $\hat{\Gamma}$ is estimated (details in the following chapters), the envelope estimator can be obtained by projecting $\hat{\beta}_{OLS}$ onto span$(\hat{\Gamma})$:

$$\hat{\beta}_{env} = \mathrm{P}_{\hat{\Gamma}} \hat{\beta}_{OLS}.$$

Cook et al. (2013) showed that the asymptotic variance of $\hat{\beta}_{env}$ is always no larger than $\hat{\beta}_{OLS}$. They also showed that at the population level, the spanned PLS directions are the same as the predictor envelope under the linear model, and the envelope method outperforms the PLS in estimating β.

We consider a toy example to empirically compare the performance of OLS, PLS and the predictor envelope under Model 1.1. We simulate data with different sample sizes $n = 100, 250, 500, 750$ and 1000 for 100 times. We choose $p = 9$, $\Gamma = (1/3, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3)^T$, $\eta = 2$, $\Omega = 0.5$, $\Omega_0 = 0.2 I_8$. The error term ε_i is generated from $N(0, 9)$, the predictor X_i is from $N(0, \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T)$ and $Y_i = \eta^T \Gamma^T X_i + \varepsilon_i$. The L2-norm $\|\hat{\beta} - \beta\|_2$ for OLS, PLS and the predictor envelope are shown in Figure 1.1.

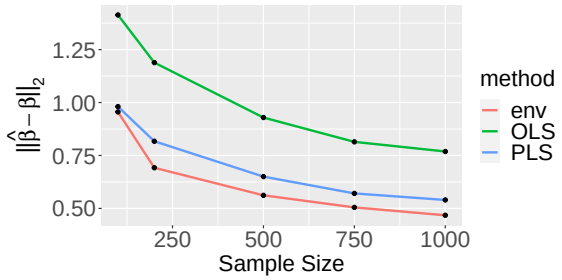


Figure 1.1: A toy example comparing the estimation of β in Model 1.1 using OLS, PLS and the predictor envelope.

The results in Figure 1.1 indicate both PLS and the predictor envelope have better efficiency in estimating β under Conditions 1.0.1–1.0.2, and the predictor

envelope outperforms PLS, which corroborates with Proposition 4.4 and 4.5 in Cook et al. (2013).

As of now, the envelope model gets an increasing amount of attention due to its effectiveness to increase the asymptotic efficiency. Different variants of envelope methods have been proposed in various settings, including response envelope (Cook et al., 2010), partial envelope (Su and Cook, 2011), inner envelope (Su and Cook, 2012), scaled envelope (Cook and Su, 2013), predictor envelope (Cook et al., 2013), reduced rank envelope (Cook et al., 2015), simultaneous envelope (Cook and Zhang, 2015b), model-free envelope (Cook and Zhang, 2015a), sparse envelope (Su et al., 2016) and tensor envelope (Li and Zhang, 2017).

However, there are still many open problems remaining unsolved for the envelope models. In my series of work, I focus on adapting the envelope model to the following problems: (1) when the data contains missing at random predictors and responses; (2) when the data are longitudinal; (3) when the regression functional is not linear; (4) when we are running two stage least squares (2SLS) and have many instrumental variables. Each of the chapter will focus on one specific aforementioned problem. Each chapter, although highly related to each other, is self-contained. Readers can start from any chapters of interests. Chapter 2 is adapted from Ma, Liu, and Yang (2021). Chapter 3 is adapted from Shi, Ma, and Liu (2020). Chapter 4 is adapted from Ma, Liu, and Yang (2021). Chapter 5 is adapted from Ma, Kang, and Liu (2022).

Chapter 2

Envelope method with ignorable missing data

2.1 Introduction

Recently, a new dimension reduction method called the envelope method has been proposed in the multivariate regressions (Cook et al., 2010). Unlike the standard dimension reduction methods, the envelope method assumes the redundancy among responses rather than among predictors. Specifically, it is assumed that there exist some linear combinations of the response variables that do not contribute to the regression. Under such an assumption, the envelope method is shown to have efficiency gain over the ordinary least squares which regresses one response at a time ignoring other responses. Similar redundancy structures have also been extended to hold among the predictors or among both predictors and responses.

It is known that the estimation of the central space may suffer from bias when the correlations between variables are high (Cook, 2018b). The envelope assumptions circumvent the challenge of identifying the central space in the standard dimension reduction problem when the correlation between variables is high, at the cost of obtaining a bigger space containing the parameters of interest, and thus makes the envelope estimates more reliable.

Various envelope methods have been proposed in different settings, including response envelope (Cook et al., 2010), inner envelope (Su and Cook, 2012), scaled envelope (Cook and Su, 2013), reduced rank envelope (Cook et al., 2015), predictor envelope (Cook et al., 2013), simultaneous envelope (Cook and Zhang, 2015b), sparse envelope (Su et al., 2016), tensor envelope (Li and Zhang, 2017), and model-free

envelope (Cook and Zhang, 2015a). Algorithms such as 1-D algorithm (Cook and Zhang, 2016) and envelope coordinate descent (Cook and Zhang, 2018) have also been proposed to effectively and efficiently estimate the envelope models.

A prominent problem when a large number of responses and predictors are collected is the missingness of responses or predictors. Missing data may arise when a subject refuses to respond to certain questions or when the data is not collected. The missing data mechanism is said to be missing at random (MAR) or ignorable if it only depends on the observed data and it is said to be missing not at random (MNAR) or nonignorable if otherwise. As Little and Rubin (2014) suggested, in most MAR scenarios, a complete case analysis would lead to an inefficient or possibly biased results. We assume the missingness mechanism is MAR throughout this paper.

In this chapter, we generalize the envelope method for data with missing predictors and responses. As the parameters under the envelope method are not pointwise identifiable, such a generalization requires special decomposition. The importance of the research lies in several aspects. First, with rapidly advancing technology, it is common that high-dimensional responses are collected to characterize multiple aspects of individuals. Biased and inefficient results will be obtained if the analysis deletes all the observations with missing values. Second, while the standard missing data methods typically suffer from an efficiency loss, as compared to the full data analysis, the method that incorporates dimension reduction can potentially recover substantial efficiency. Third, our proposed method to recover the missing information can also be generalized to the predictor envelope model where the redundancy is assumed among the predictors rather than the responses, as well as to the case where the redundancy is present among both the responses and the predictors. And lastly, to the best of our knowledge, our approach is among the first few in the dimension reduction literature to discuss the case where both responses and predictors are subject to missingness.

2.2 Preliminary

Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ir})^T$ and $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ denote the multivariate responses and predictors for individual i, where T denotes the transpose of a matrix and $i = 1, \ldots, n$. Also, let $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n) \in \mathbb{R}^{r \times n}$ and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathbb{R}^{p \times n}$, where $\mathbf{Y} \in \mathbb{R}^{p \times n}$ denotes that $\mathbf{Y}$ is an element in the set of all real matrices with

dimension $r \times n$. Consider the multivariate linear regression model

$$\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i, \tag{2.1}$$

where $\boldsymbol{\varepsilon}_i$ are identically and independently (i.i.d) distributed following a normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$. Here, the normality of the predictor is assumed for simplicity. We extend later (in proposition 2) that the efficiency of envelope estimator can still be achieved even the distriubtion of the error or predictors are misspecified. Let $R_{X_{ij}} = 1$ if X_{ij} is observed and $R_{X_{ij}} = 0$ if otherwise, where $j = 1, \ldots, p$. Similarly, let $R_{Y_{ik}}$ denote the missing indicator for Y_{ik}, where $k = 1, \ldots, r$. Let $\mathbf{R}_i = (R_{X_{i1}}, \ldots, R_{X_{ip}}, R_{Y_{i1}}, \ldots, R_{Y_{ir}})^T$ denote the vector of missingness indicators of all variables for individual i. Let $\mathbf{Y}_{i,mis}$ and $\mathbf{X}_{i,mis}$ denote the vectors of the missing responses and the predictors for individuals i. Let $\mathbf{Y}_{i,obs}$ and $\mathbf{X}_{i,obs}$ denote the vectors of the observed responses and predictors for individual i. Under such notations, different individuals may have different missing responses and predictors, i.e., the lengths and the components of $\mathbf{Y}_{i,obs}$ and $\mathbf{X}_{i,obs}$ differ from one to another. Let $\mathbf{D}_{i,obs} = (\mathbf{X}_{i,obs}, \mathbf{Y}_{i,obs})^T$ and $\mathbf{D}_{i,mis} = (\mathbf{X}_{i,mis}, \mathbf{Y}_{i,mis})^T$ denote the observed data and the missing data for individual i, respectively. Let y_{ik} and x_{ij} denote the possible value of Y_{ik} and X_{ij}. Then $\mathbf{y}_i = (y_{i1}, \ldots, y_{ir})^T$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ are the possible value of $\mathbf{Y}_i$ and $\mathbf{X}_i$. Let $\mathbf{x}_{i,obs}$ and $\mathbf{x}_{i,mis}$ denote the value of the observed and missing predictors. Define $\mathbf{y}_{i,obs}$ and $\mathbf{y}_{i,mis}$ similarly. We assume the missingness is ignorable:

Assumption 2.2.1 (ignorability) $\mathbf{R}_i \perp\!\!\!\perp \mathbf{D}_{i,mis} | \mathbf{D}_{i,obs}$.

Assumption 2.2.1 implies that given the observed data, the failure to observe a variable does not depend on the unobserved data. This particular type of missingness is called missing at random (MAR) or ignorable missingness. A complete case analysis is inefficient and can be seriously biased (Little, 1992). Throughout the paper, we assume both covariates and responses are missing at random, which has also been assumed in Chen et al. (2008) and Hristache and Patilea (2017).

In multivariate regression with fully observed data, the envelope method (Cook et al., 2010) is motivated by the observation that some characteristics of the responses are unaffected by the changes of the predictors. For example, in a randomized trial, the difference between the repeated measures of the blood pressure of a patient in the treatment group (or the control group) may only reflect the aging over time rather than the treatment effect. Define a matrix $\mathbf{O} \in \mathbb{R}^{r\times r}$ as orthonor-

mal if and only if it satisfies $\mathbf{O}^T\mathbf{O} = \mathbf{I}_r$, where $\mathbf{I}_r$ denotes the identity matrix with dimension r. Assume there exists an orthonormal matrix $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{r\times r}$ such that

$$\mathbf{\Gamma}_0^T\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \quad \text{(i)}$$

$$\mathbf{\Gamma}^T\mathbf{Y} \perp\!\!\!\perp \mathbf{\Gamma}_0^T\mathbf{Y}|\mathbf{X} \quad \text{(ii)}$$

where $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$, $\mathbf{\Gamma}_0 \in \mathbb{R}^{r\times(r-u)}$, and $0 \leq u \leq r$. These assumptions imply that the distribution of $\mathbf{\Gamma}_0^T\mathbf{Y}$ is not affected by the predictor $\mathbf{X}$ marginally or through an association of $\mathbf{\Gamma}^T\mathbf{Y}$. For example, assume $\mathbf{Y} = (Y_1, Y_2)$. Suppose $Y_1 = \boldsymbol{\beta}\mathbf{X} + \varepsilon_1$ and $Y_2 = -\boldsymbol{\beta}\mathbf{X} + \varepsilon_2$, where ε_1 and ε_2 follow two normal distributions and they are independent of each other. The predictors $\mathbf{X}$ do not affect the summation of responses $Y_1 + Y_2$. Additionally, it can be verified that $Y_1 - Y_2$ is independent of $Y_1 + Y_2$; thus, $Y_1 + Y_2$ can be completely discarded in the regression. That is, the regression of $\mathbf{Y}$ on $\mathbf{X}$ can be replaced with the regression of $Y_1 - Y_2$ on $\mathbf{X}$. In this example, $\mathbf{\Gamma} = (1, -1)^T/\sqrt{2}$, and $\mathbf{\Gamma}_0 = (1, 1)^T/\sqrt{2}$. The combinations of responses that are involved in the regression, $\mathbf{\Gamma}^T\mathbf{Y}$, is called the material part of $\mathbf{Y}$, and the part that is uninvolved, $\mathbf{\Gamma}_0^T\mathbf{Y}$, is called the immaterial part of $\mathbf{Y}$. Hence, the main focus of the envelope method is to find the column space of $\mathbf{\Gamma}$, i.e., span($\mathbf{\Gamma}$), that fully contains the information of $\boldsymbol{\beta}$, i.e., find an envelope of $\boldsymbol{\beta}$.

It has been shown that under the normality of the error term, the independence assumptions (i) and (ii) are equivalent to the reparameterization (i)* span($\boldsymbol{\beta}$) $\subseteq$ span($\mathbf{\Gamma}$) and (ii)* $\mathbf{\Sigma} = \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2$, where $\mathbf{\Sigma}_1 = \mathbf{P}_{\mathbf{\Gamma}}\mathbf{\Sigma}\mathbf{P}_{\mathbf{\Gamma}}$, $\mathbf{\Sigma}_2 = \mathbf{Q}_{\mathbf{\Gamma}}\mathbf{\Sigma}\mathbf{Q}_{\mathbf{\Gamma}}$, $\mathbf{P}_{\mathbf{\Gamma}} = \mathbf{\Gamma}\mathbf{\Gamma}^T$ is the projection matrix onto span($\mathbf{\Gamma}$), and $\mathbf{Q}_{\mathbf{\Gamma}} = \mathbf{I}_r - \mathbf{P}_{\mathbf{\Gamma}}$ (Cook et al., 2010). Under (i)*, we have $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}$, where $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$. Also, by setting $\mathbf{\Omega} = \mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma}$, and $\mathbf{\Omega}_0 = \mathbf{\Gamma}_0^T\mathbf{\Sigma}\mathbf{\Gamma}_0$, we can alternatively reparameterize the covariance matrix as $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$. The subspace satisfying (i)* and (ii)* is not unique, but the envelope is uniquely defined as the smallest subspace satisfying these assumptions. Note $\mathbf{\Gamma}^T\mathbf{Y} \in \mathbb{R}^{u\times n}$; thus, the dimension u is known as the envelope dimension. Once an estimate of the basis $\mathbf{\Gamma}$, $\hat{\mathbf{\Gamma}}$, is obtained, $\hat{\boldsymbol{\beta}}_{env}$ is obtained by projecting the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ onto the estimated envelope space, $\hat{\boldsymbol{\beta}}_{env} = \mathbf{P}_{\hat{\mathbf{\Gamma}}}\hat{\boldsymbol{\beta}}$.

Figure 2.1 demonstrates the intuition of efficiency gain of the envelope method when there is no missing data, or equivalently, with the full data. Consider two groups of individuals (the group with $X = 1$ is denoted by triangles and the other with $X = 0$ is by circle dots), where each point (triangle or circle dot) denotes one individual. Two responses Y_1 and Y_2 are collected for each individual. Suppose that we are interested in estimating the group difference on Y_1, the standard maximum

likelihood estimation (MLE) projects all the data onto the Y_1 axis, ignoring information on Y_2 completely. The density curves of the two group distributions of Y_1 are given at the bottom in Figure 2.2(a). The two curves are hard to distinguish as they almost overlapped. The full data MLE for the group difference is 0.11 with the bootstrap standard error being 0.12 and the p-value being 0.37. Thus, it is hard to distinguish between the two groups. While the true difference between the two group mean of Y_1, 0.32, is contained in the 95% confidence interval of the full data MLE, the large variability of the estimator makes the point estimate deviate from the true parameter value.

The idea of the envelope method is to reduce the noise in the original data by projecting each observation onto the direction that contains all the information related to the regression. The two groups are best distinguished along the direction of the black solid line. In contrast, the two groups have almost identical distribution along the direction that is orthogonal to the black solid line. That is, the information that is orthogonal to the black solid line does not contribute to the distinction between the two groups. Thus, eliminating that part of variation does not sacrifice any relevant information for the regression, but instead makes the regression more efficient. An estimate of the black solid line is shown as the purple dashed line in Figure 2.2(b). All the points are thus first projected onto the estimated direction $\hat{\mathbf{\Gamma}}^T\mathbf{Y}$, then projected onto the Y_1 axis. For example, a data point A was first projected onto the estimated envelope direction with an intersection B, and then projected onto the Y_1 axis. Cook et al. (2010) showed that the envelope method can achieve substantial efficiency gain when the envelope direction is aligned with the eigenspaces of $\mathbf{\Sigma}$ that correspond to relatively small eigenvalues. In that way, linear combinations of $\mathbf{Y}$ with larger variances can be eliminated by the projection. In Figure 2.2(b), the direction that can better distinguish the two groups is aligned with the direction that the data has less variability, so the envelope method is expected to provide substantial efficiency gain. The density curves of the two groups under the envelope estimation are shown at the bottom of Figure 2.2(b) and they have much smaller spreads. The envelope estimator for the group difference is 0.32 with the standard error being 0.03 and the p-value < 0.001. Thus, it is much easier to distinguish between the two groups.

Now, consider the case where the predictors $\mathbf{X}$ are fully observed but some values of the responses are missing (see Figure 2.2). The missingness mechanism is as follows. For an individual i for $i = 1, \ldots, 150$, if $X_i = 1$ and if Y_{i1} is among the largest 30 $Y_{i'1}$ for $i' = 1, \ldots, 150$, then Y_{i2} is missing. If $X_i = 0$ and if Y_{i2} is

among the largest 45 $Y_{i'2}$ for $i' = 1, \ldots, 150$, then Y_{i1} is missing. Such missingness mechanism is MAR, and the missing rate is 30% for Y_1, and 20% for Y_2. The hollow triangle represents Y_1 missing, and the hollow circle dot represents Y_2 missing. The standard EM method is shown in Figure 2.3(a). Although being an asymptotically unbiased method, the standard EM estimates of the group difference is 0.11. Similar as the full data MLE, the point estimate of the standard EM also deviates from the true parameter value due to the large variability. The bootstrap standard error is 0.12 with the p-value being 0.37. The spreads of the two group densities are again relatively large, resulting in a relatively inefficient estimate.

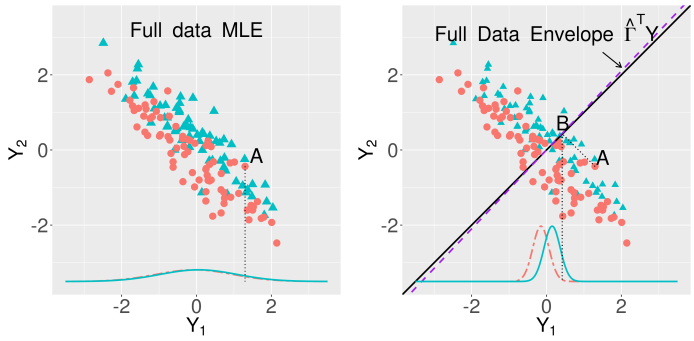
The existing envelope methods for solving $\mathbf{\Gamma}$ all require the data to be fully observed (Cook et al., 2010; Cook and Zhang, 2016). Figure 2.3(b) shows the complete case envelope where all the observations with missing data are deleted from the analysis. The estimated complete case envelope direction is shown as the blue dashed line in Figure 2.3(b), which is far from the true envelope direction (black solid line). This leads to a severe bias: even the sign of the estimated parameter is incorrect. The complete case envelope estimate is -1.63 with the bootstrap standard error being 0.15 and the p-value < 0.001.

Our method is shown in Figure 2.3(c). Different from the complete case analysis, we use both the complete cases and the partially missing information. Our proposed method is asymptotically unbiased when the missing pattern is MAR. The estimated envelope direction is shown as the red dashed line. Our method recovers the envelope direction and achieves significant efficiency gain over the standard EM as the density curves have much smaller spreads. The EM envelope estimator is 0.31 with the bootstrap standard error 0.04 and the p-value < 0.001. It is interesting to see that our method may even outperform the full data MLE as the efficiency gain by the envelope method outweighs the information loss due to missing data in this illustrative example.

2.3 The Observed Data Likelihood

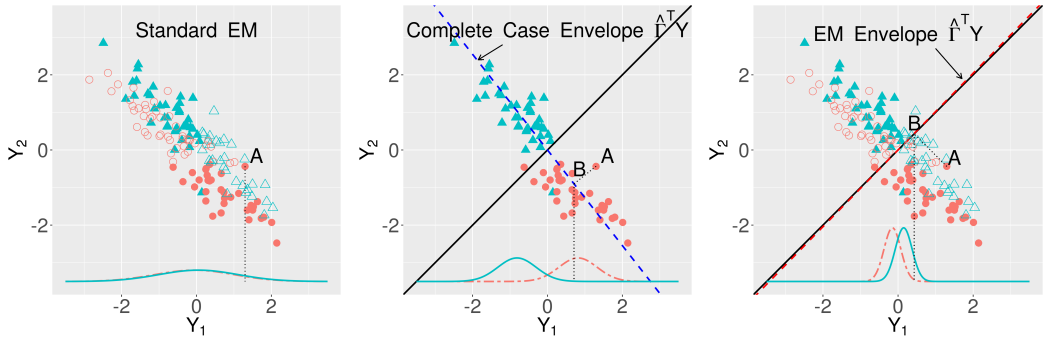
The envelope method proposed by Cook et al. (2010) utilizes the full data likelihood function $L_{full} = \prod_{i=1}^{n} f(\mathbf{y}_i|\mathbf{x}_i; \boldsymbol{\eta}, \mathbf{\Gamma}, \mathbf{\Omega}_0, \mathbf{\Omega})$ to obtain the MLE of the parameters. In the presence of missing data, we replace the full data likelihood with the observed data likelihood

Figure 2.1: Intuitive illustration of the envelope method without missing data. Two groups are shown using circle dots ($X = 0$) and triangles ($X = 1$). The solid line is the true envelope direction, the dashed lines are the estimated envelope. The density curves of the two groups using envelope method are shown at the bottom of each subfigure.



(a) Full data MLE (b) Full data envelope

Figure 2.2: Intuitive illustration of the envelope method in the presence of missing data. Two groups are shown using circle dots ($X = 0$) and triangles ($X = 1$). Hollow circle dots or triangles indicate one of the components of $\mathbf{Y}$ is missing: the hollow triangle has Y_1 missing, and the hollow circle dot has Y_2 missing. The solid line is the true envelope direction, the dashed lines are the estimated envelope using different methods. The density curves of the two groups using different methods are shown at the bottom of each subfigure.



(a) Standard EM (b) CC Envelope (c) EM Envelope

$$
\begin{aligned}
L_{obs} &= \prod_{i=1}^{n} f(\mathbf{y}_{i,obs}|\mathbf{x}_{i,obs};\boldsymbol{\eta},\boldsymbol{\Gamma},\boldsymbol{\Omega}_0,\boldsymbol{\Omega}) \\
&\propto \prod_{i=1}^{n} \int\int f(\mathbf{y}_{i,obs},\mathbf{y}_{i,mis}|\mathbf{x}_i;\boldsymbol{\eta},\boldsymbol{\Gamma},\boldsymbol{\Omega}_0,\boldsymbol{\Omega}) f(\mathbf{x}_{i,obs},\mathbf{x}_{i,mis};\boldsymbol{\rho}) d\mathbf{x}_{i,mis} d\mathbf{y}_{i,mis},
\end{aligned}
$$

where $\boldsymbol{\rho}$ is the parameter for the predictors' distribution and $\propto$ denotes proportional to, i.e., a multiplicative constant is omitted. Let $\chi_{i,mis}$ denote the set of predictors $\mathbf{X}_i$ that is missing for individual i. For example, if $\mathbf{X}_{i,mis} = X_{i1}$, then $\chi_{i,mis} = \{X_{i1}\}$. Write $\chi_{i,mis} = \emptyset$ when all the p predictors are observed for this individual. Since $\int f(\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis}|\mathbf{x}_i; \boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}_0, \boldsymbol{\Omega}) d\mathbf{y}_{i,mis} = f(\mathbf{y}_{i,obs}|\mathbf{x}_i; \boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}_0, \boldsymbol{\Omega})$, we can simplify the observed data likelihood as

$$L_{obs} \propto \prod_{i \in \{\chi_{i,mis} = \emptyset\}} f(\mathbf{y}_{i,obs}|\mathbf{x}_i; \boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}_0, \boldsymbol{\Omega}) \prod_{i \in \{\chi_{i,mis} \neq \emptyset\}} \int f(\mathbf{y}_{i,obs}|\mathbf{x}_i; \boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}_0, \boldsymbol{\Omega}) f(\mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}; \boldsymbol{\rho}) d\mathbf{x}_{i,mis}.$$

The first part of the observed data likelihood corresponds to the likelihood of the individuals with fully observed predictors. The second part corresponds to the likelihood of individuals with missing predictors. Hence, it is easy to see that observed data likelihood utilizes more information than the complete data likelihood.

The observed data likelihood is in general hard to calculate as it involves the multivariate integral. Closed form observed data likelihood exists under certain distributions. Example A.2.1 in the Appendix derives the closed form of the observed data likelihood when predictors and responses follow a joint normal distribution. However, in general, the integral in the observed data likelihood may result in a complicated form. Cook and Zhang (2015a) pointed out that the envelope method performs poorly when the objective function on its first order derivatives do not have a closed form. Even when the observed data likelihood is available in a closed form, the parameter is typically complicatedly intertwined in the likelihood. Together with the fact that the parameter is not pointwise identifiable, it is very challenging to calculate the maximum likelihood estimates. Such a challenge was also identified in Cook and Zhang (2015a) in the context of generalized linear models. In this paper, we propose an EM envelope algorithm that can identify and estimate the envelope space with missing data.

2.4 The EM Envelope

2.4.1 The EM envelope algorithm

Let $l_{full}(\boldsymbol{\theta}|L) = \log L_{full}(\boldsymbol{\theta}|L)$ denote the log of full data likelihood. Then, the logarithm of full data likelihood of $(\mathbf{X}, \mathbf{Y})$ is

$$\begin{aligned} l_{full}(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) &= \log\{f_{y|x}(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\} + \log\{f_x(\mathbf{x}|\boldsymbol{\theta})\} \\ &= \sum_{i=1}^{n}[-\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y}_i - \boldsymbol{\beta}\mathbf{x}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\beta}\mathbf{x}_i) + \log\{f_x(\mathbf{x}_i|\boldsymbol{\rho})\}] + C \\ &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{\Delta}_i + C, \end{aligned}$$

where $\boldsymbol{\Delta}_i = (\mathbf{y}_i - \boldsymbol{\beta}\mathbf{x}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\beta}\mathbf{x}_i) + 2\log\{f_x(\mathbf{x}_i|\boldsymbol{\rho})\}$ and $C = -(nr\log 2\pi)/2$. We firstly do the E-step, where

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) = \mathbb{E}\{l_{full}(\boldsymbol{\theta}|L)|\mathbf{D}_{obs}; \boldsymbol{\theta}_t\} = \int l_{full}(\boldsymbol{\theta}|L) f(\mathbf{D}_{mis}|\mathbf{D}_{obs}; \boldsymbol{\theta}_t) d\mathbf{D}_{mis}.$$

Under model (2.1), $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}_0, \boldsymbol{\Omega}, \boldsymbol{\rho})$. Note here, we abuse the notations a little. More rigorously, we should vectorize the parameters (with duplicated terms in covariance matrix deleted) on the right as $\boldsymbol{\theta}$ is a vector. We omit the vectorization notation for the simplicity of illustration. Recall that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, we can also use $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\rho})$ as the new parameters for the reparameterization. In the E-step, we have

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) = \mathbb{E}\{l_{full}(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y})|\mathbf{D}_{obs}; \boldsymbol{\theta}_t\} = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}(\boldsymbol{\Delta}_i|\mathbf{D}_{i,obs}; \boldsymbol{\theta}_t) + C.$$

Since $\mathbb{E}(\mathbf{Y}_i^T\boldsymbol{\Sigma}\mathbf{Y}_i) = \mathbb{E}\{\mathrm{tr}(\boldsymbol{\Sigma}\mathbf{Y}_i\mathbf{Y}_i^T)\} = \mathrm{tr}\{\boldsymbol{\Sigma}\mathbb{E}(\mathbf{Y}_i\mathbf{Y}_i^T)\}$, we have

$$\begin{aligned} \mathbb{E}(\boldsymbol{\Delta}_i|\mathbf{D}_{i,obs}; \boldsymbol{\theta}_t) &= \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}\mathbb{E}(\mathbf{Y}_i\mathbf{Y}_i^T|\mathbf{D}_{i,obs}; \boldsymbol{\theta}_t) + \boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\mathbb{E}(\mathbf{X}_i\mathbf{X}_i^T|\mathbf{D}_{i,obs}; \boldsymbol{\theta}_t) \\ &\quad -2\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\mathbb{E}(\mathbf{Y}_i\mathbf{X}_i^T|\mathbf{D}_{i,obs}; \boldsymbol{\theta}_t)\} - \mathbb{E}[2\log\{f_x(\mathbf{X}_i|\boldsymbol{\rho})\}|\mathbf{D}_{i,obs}; \boldsymbol{\theta}_t]. \end{aligned}$$

Let $\mathbf{A}_{i1,t} = \mathbb{E}(\mathbf{Y}_i\mathbf{Y}_i^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}_t)$, $\mathbf{A}_{i2,t} = \mathbb{E}(\mathbf{Y}_i\mathbf{X}_i^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}_t)$, $\mathbf{A}_{i3,t} = \mathbb{E}(\mathbf{X}_i\mathbf{X}_i^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}_t)$, $\mathbf{A}_{j,t} = \sum_{i=1}^n \mathbf{A}_{ij,t}$, $j = 1,\ldots,3$. Thus,

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| + \sum_{i=1}^n \mathbb{E}(\boldsymbol{\Delta}|\mathbf{D}_{i,obs};\boldsymbol{\theta}_t) + C \\ &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}^{-1}\big(\sum_{i=1}^n \mathbf{A}_{i1,t} - 2\sum_{i=1}^n \mathbf{A}_{i2,t}\boldsymbol{\beta}^T + \boldsymbol{\beta}\sum_{i=1}^n \mathbf{A}_{i3,t}\boldsymbol{\beta}^T\big)\} \\ &\quad + \mathbb{E}[\log\{f_x(\mathbf{x}_i|\boldsymbol{\rho})\}|\mathbf{D}_{i,obs};\boldsymbol{\theta}_t] + C \\ &\propto -n\log|\boldsymbol{\Sigma}| - \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}\big(\mathbf{A}_{1,t} - 2\mathbf{A}_{2,t}\boldsymbol{\beta}^T + \boldsymbol{\beta}\mathbf{A}_{3,t}\boldsymbol{\beta}^T\big)\} \\ &\quad + \mathbb{E}[2\log\{f_x(\mathbf{x}_i|\boldsymbol{\rho})\}|\mathbf{D}_{i,obs};\boldsymbol{\theta}_t] + 2C. \end{aligned}$$

After the E-step, we do the M-step. However, the parameters under the envelope method are not pointwise identifiable (Cook et al., 2010), the EM algorithm for the envelope method is not straightforward and requires a special decomposition in the M-step. We imitate that of the full data likelihood in Cook et al. (2010) to isolate the parameter to be optimized from the other parameters. We decompose $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$ as $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) = Q_1(\boldsymbol{\rho}|\boldsymbol{\theta}_t) + Q_2(\boldsymbol{\beta},\boldsymbol{\Sigma}|\boldsymbol{\theta}_t)$, where $Q_1(\boldsymbol{\rho}|\boldsymbol{\theta}_t) = \mathbb{E}[2\log\{f_x(\mathbf{X}_i|\boldsymbol{\rho})\}|\mathbf{D}_{obs};\boldsymbol{\theta}_t] + 2C$, and $Q_2(\boldsymbol{\beta},\boldsymbol{\Sigma}|\boldsymbol{\theta}_t) = -n\log|\boldsymbol{\Sigma}| - \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}\big(\mathbf{A}_{1,t} - 2\mathbf{A}_{2,t}\boldsymbol{\beta}^T + \boldsymbol{\beta}\mathbf{A}_{3,t}\boldsymbol{\beta}^T\big)\}$. As $Q_1(\boldsymbol{\rho}|\boldsymbol{\theta}_t)$ only involves $\boldsymbol{\rho}$, the maximizer of $Q_1(\boldsymbol{\rho}|\boldsymbol{\theta}_t)$ is $\boldsymbol{\rho}_{t+1} = \arg\max_{\boldsymbol{\rho}\in\boldsymbol{\Pi}} \mathbb{E}[2\log\{f_x(\mathbf{x}_i|\boldsymbol{\rho})\}|\mathbf{D}_{obs};\boldsymbol{\theta}_t]$, where $\boldsymbol{\Pi}$ is the parameter space of $\boldsymbol{\rho}$.

To find the maximizer of $Q_2(\boldsymbol{\beta},\boldsymbol{\Sigma}|\boldsymbol{\theta}_t)$, note under the envelope assumptions (i) and (ii), we have $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$, where $\boldsymbol{\Sigma}_1 = \mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{P}_{\boldsymbol{\Gamma}}$, $\boldsymbol{\Sigma}_2 = \mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{Q}_{\boldsymbol{\Gamma}}$ with $\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2 = \mathbf{0}$, and $\mathrm{Span}(\boldsymbol{\beta}) \subseteq \mathrm{Span}(\boldsymbol{\Sigma}_1)$. This implies $\boldsymbol{\Sigma}_2\boldsymbol{\beta} = 0$. Additionally, as $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_1^\dagger + \boldsymbol{\Sigma}_2^\dagger$, where $\dagger$ indicates the Moore-Penrose inverse, we can write Q_2 as:

$$\begin{aligned} Q_2(\boldsymbol{\beta},\boldsymbol{\Sigma}|\boldsymbol{\theta}_t) = &-n\log\det{}_0\boldsymbol{\Sigma}_1 - \mathrm{tr}\{\boldsymbol{\Sigma}_1^\dagger(\mathbf{A}_{1,t} - 2\mathbf{A}_{2,t}\boldsymbol{\beta}^T + \boldsymbol{\beta}\mathbf{A}_{3,t}\boldsymbol{\beta}^T)\} \\ &- n\log\det{}_0\boldsymbol{\Sigma}_2 - \mathrm{tr}\big(\boldsymbol{\Sigma}_2^\dagger\mathbf{A}_{1,t}\big), \end{aligned}$$

where $\det_0(\mathbf{A})$ denotes the product of its non-zero eigenvalues. Further, we have $Q_2(\boldsymbol{\beta},\boldsymbol{\Sigma}|\boldsymbol{\theta}_t) = Q_{2,1}(\boldsymbol{\beta},\boldsymbol{\Sigma}_1|\boldsymbol{\theta}_t) + Q_{2,2}(\boldsymbol{\Sigma}_2|\boldsymbol{\theta}_t)$, where $Q_{2,1}(\boldsymbol{\beta},\boldsymbol{\Sigma}_1|\boldsymbol{\theta}_t) = -n\log\det{}_0\boldsymbol{\Sigma}_1 - \mathrm{tr}\{\boldsymbol{\Sigma}_1^\dagger(\mathbf{A}_{1,t} - 2\mathbf{A}_{2,t}\boldsymbol{\beta}^T + \boldsymbol{\beta}\mathbf{A}_{3,t}\boldsymbol{\beta}^T)\}$, and $Q_{2,2}(\boldsymbol{\Sigma}_2|\boldsymbol{\theta}_t) = -n\log\det{}_0\boldsymbol{\Sigma}_2 - \mathrm{tr}\big(\boldsymbol{\Sigma}_2^\dagger\mathbf{A}_{1,t}\big)$. Suppose for the moment, $\boldsymbol{\Sigma}_1$ is fixed. Then, from

$$\begin{aligned} &\mathrm{tr}\{\boldsymbol{\Sigma}_1^\dagger(\mathbf{A}_{1,t} - 2\mathbf{A}_{2,t}\boldsymbol{\beta}^T + \boldsymbol{\beta}\mathbf{A}_{3,t}\boldsymbol{\beta}^T)\} \\ = &\mathrm{tr}\{\boldsymbol{\Sigma}_1^\dagger(\mathbf{A}_{1,t} - \mathbf{A}_{2,t}\mathbf{A}_{3,t}^{-1}\mathbf{A}_{2,t}^T)\} + \mathrm{tr}\{(\mathbf{A}_{3,t}^{\frac{1}{2}}\boldsymbol{\beta}^T - \mathbf{A}_{3,t}^{-\frac{1}{2}}\mathbf{A}_{2,t}^T)\boldsymbol{\Sigma}_1^\dagger(\mathbf{A}_{3,t}^{\frac{1}{2}}\boldsymbol{\beta}^T - \mathbf{A}_{3,t}^{-\frac{1}{2}}\mathbf{A}_{2,t}^T)^T\}, \end{aligned}$$

the maximizer of $Q_{2,1}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_1|\boldsymbol{\theta}_t)$ subjects to $\text{Span}(\boldsymbol{\beta}) \subseteq \text{Span}(\boldsymbol{\Sigma}_1)$ with $\boldsymbol{\Sigma}_1$ fixed is $\boldsymbol{\beta}_{t+1} = \mathbf{P}_{\boldsymbol{\Sigma}_1}\hat{\boldsymbol{\beta}}_{std,t} = \mathbf{P}_{\boldsymbol{\Sigma}_1}\mathbf{A}_{2,t}\mathbf{A}_{3,t}^{-1}$, where $\hat{\boldsymbol{\beta}}_{std,t} = \mathbf{A}_{2,t}\mathbf{A}_{3,t}^{-1}$. Since $\mathbf{Q}_{\boldsymbol{\Sigma}_1}\boldsymbol{\Sigma}_1^{\dagger} = \mathbf{0}$, we have $Q_{2,1}(\boldsymbol{\beta}_{t+1}, \boldsymbol{\Sigma}_1|\boldsymbol{\theta}_t) = -n\log\det_0\boldsymbol{\Sigma}_1 - \text{tr}\{\boldsymbol{\Sigma}_1^{\dagger}(\mathbf{A}_{1,t} - \mathbf{A}_{2,t}\mathbf{A}_{3,t}^{-1}\mathbf{A}_{2,t}^T)\}$.

In order to maximize $Q_{2,1}(\boldsymbol{\beta}_{t+1}, \boldsymbol{\Sigma}_1|\boldsymbol{\theta}_t)$, $Q_{2,2}(\boldsymbol{\Sigma}_2|\boldsymbol{\theta}_t)$ over $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, we use the Lemma 4.3 in Cook et al. (2010), which is reviewed as Lemma A.3.1 in the Appendix. Suppose matrix $\boldsymbol{\Gamma}$ is given, then by Lemma A.3.1, we have $\boldsymbol{\Sigma}_{1,t+1} = \mathbf{P}_{\boldsymbol{\Gamma}}(\mathbf{A}_{1,t} - \mathbf{A}_{2,t}\mathbf{A}_{3,t}^{-1}\mathbf{A}_{2,t}^T)\mathbf{P}_{\boldsymbol{\Gamma}}/n$ and $\boldsymbol{\Sigma}_{2,t+1} = \mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{A}_{1,t}\mathbf{Q}_{\boldsymbol{\Gamma}}/n$. Hence, $Q_{2,1}(\boldsymbol{\beta}_{t+1}, \boldsymbol{\Sigma}_{1,t+1}|\boldsymbol{\theta}_t) = C_1 - n\log\det_0\{\mathbf{P}_{\boldsymbol{\Gamma}}(\mathbf{A}_{1,t} - \mathbf{A}_{2,t}\mathbf{A}_{3,t}^{-1}\mathbf{A}_{2,t}^T)\mathbf{P}_{\boldsymbol{\Gamma}}\}$, $Q_{2,2}(\boldsymbol{\Sigma}_{2,t+1}|\boldsymbol{\theta}_t) = C_2 - n\log\det_0\big(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{A}_{1,t}\mathbf{Q}_{\boldsymbol{\Gamma}}\big)$, where $C_1 = nu\log n - nu$ and $C_2 = n(r-u)(\log n - 1)$. Finally, we find the matrix $\boldsymbol{\Gamma}$ to minimize the function $\log\det\{\mathbf{P}_{\boldsymbol{\Gamma}}(\mathbf{A}_{1,t} - \mathbf{A}_{2,t}\mathbf{A}_{3,t}^{-1}\mathbf{A}_{2,t}^T)\mathbf{P}_{\boldsymbol{\Gamma}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{A}_{1,t}\mathbf{Q}_{\boldsymbol{\Gamma}}\}$. The elements in $\boldsymbol{\Gamma}$ are not pointwise identifiable; however, as the objective function above is a function of $\text{Span}(\boldsymbol{\Gamma})$, we only need to estimate the span of the column space of $\boldsymbol{\Gamma}$, which is identifiable. In order to find the MLE of $\text{Span}(\boldsymbol{\Gamma})$, full Grassmannian optimization is needed (Cook et al., 2010, 2016). However, to simplify the calculation and reduce the computation burden, we apply the 1-D algorithm proposed by Cook and Zhang (2016) to solve $\boldsymbol{\Gamma}$. The 1-D algorithm only provide a $\sqrt{n}$-consistent estimate of $\boldsymbol{\Gamma}$ rather than the most efficient estimate. However, we still find good performance of EM envelope method with 1-D algorithm in simulation studies. Details about the algorithm are in the Appendix.

2.4.2 Selection of the envelope dimension

The selection of the envelope dimension can be viewed as a diagnostic or model selection for the envelope model. Model selection criteria for missing data problem such as the likelihood ratio test and the information criteria including AIC, BIC, typically involve the observed data likelihood. As mentioned, the observed data likelihood may be complicated and not in a closed form. Hence, it is ideal if the calculation of the model selection criteria could be obtained directly from the EM output. Ibrahim et al. (2008) proposed the information criteria for missing data problems. They used the fact that $\mathbb{E}\{\log f(\mathbf{D}_{obs}|\boldsymbol{\theta})|\mathbf{D}_{obs}; \boldsymbol{\theta}_t\} = Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) - H(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$, where $H(\boldsymbol{\theta}|\boldsymbol{\theta}_t) = \mathbb{E}\{\log f(\mathbf{D}_{mis}|\mathbf{D}_{obs}; \boldsymbol{\theta})|\mathbf{D}_{obs}; \boldsymbol{\theta}_t\}$ and $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$ was defined in Section 2.4.1. The Q function can be computed from the EM output and the H function can be analytically approximated as part of the EM output.

Eck and Cook (2017) recommended using the BIC to select the envelope dimension, because the AIC tends to over select the true dimension and the likelihood ratio testing is inconsistent. Thus, we generalize the BIC for the missing data problem

following Ibrahim et al. (2008) as $\text{BIC}_{H,Q} = -2Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) + 2H(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) + pu \log n$. The penalty term is $pu \log n$ because under the envelope model, there are $pu + r(r+1)/2$ unknown parameters in total, and only pu varies with dimension u. The asymptotic properties of $\text{BIC}_{H,Q}$ are given in Ibrahim et al. (2008).

The computation of the H function is not straightforward since it may not have a closed form. Ibrahim et al. (2008) proposed a method for approximating the H function through the truncated Hermite expansion with MCMC sampling. Alternatively, an approximation of BIC_Q could be obtained by omitting $H(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})$, where $\text{BIC}_Q = -2Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) + pu \log n$. When the proportion of missing information is small, the use of BIC_Q is adequate.

Alternatively, if the distribution of the error term $\boldsymbol{\varepsilon}_i$ is not normal, the BIC_Q method we proposed above will not be accurate. Under this scenario, we adopt the bootstrap method proposed in Ye and Weiss (2003); Dong and Li (2010) for choosing the envelope dimension u.

Our target is the u-dimensional subspace spanned by the basis matrix $\boldsymbol{\Gamma}$. By bootstrapping data $(\mathbf{X}, \mathbf{Y})$ for a total of b times, we get the corresponding estimate of the envelope space $\hat{\boldsymbol{\Gamma}}^1, \ldots, \hat{\boldsymbol{\Gamma}}^b$. If the proposed dimension is $u^* > u$, then $\text{span}(\hat{\boldsymbol{\Gamma}})$ can be any space that contains $\text{span}(\boldsymbol{\Gamma})$, and thus, should be variational. As for evaluating the variability of $\hat{\boldsymbol{\Gamma}}^1, \ldots, \hat{\boldsymbol{\Gamma}}^b$, we use the *vector correlation coefficient* q proposed by Hotelling (1936a). Suppose $\mathbf{A}$ and $\mathbf{B} \in \mathbb{R}^{r \times u}$ are semi-orthonormal matrices, then

$$q^2(\mathbf{A}, \mathbf{B}) = |\mathbf{B}^T \mathbf{A} \mathbf{A}^T \mathbf{B}|.$$

We see that $q^2(\mathbf{A}, \mathbf{B}) \in [0, 1]$ and higher value of q^2 indicates higher correlation between the two subspaces. When $q^2(\mathbf{A}, \mathbf{B}) = 1$, $\text{span}(\mathbf{A}) = \text{span}(\mathbf{B})$. Hence, we choose the largest dimension u^* such that

$$\frac{1}{b} \sum_{j=1}^{b} q^2(\hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\Gamma}}^j) > 0.95.$$

We use the BIC_Q method in Section 2.5.1 when we assume normal errors, and the bootstrap method in Section 2.5.2 when the error is not normal. Empirically, these two methods show good performances.

Also, Eck and Cook (2017) suggested dimension selection can be entirely avoided by using a weighted average of envelope estimators, one for each possible dimension. They also showed that the weighted envelope estimator is $\sqrt{n}$-consistent, where the standard error can be well approximated by the residual bootstrap.

2.4.3 Asymptotics

The following propositions guarantees the efficiency gain of the EM envelope estimator $\hat{\boldsymbol{\beta}}_{em\_env}$ over the standard EM estimator $\hat{\boldsymbol{\beta}}_{em\_std}$. Specifically, proposition 2.4.1 establishes the result when the density for $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are both correctly specified. Proposition 2.4.2 extends the result to the case where both $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are misspecified. In this case, under some regularity conditions, our EM envelope estimator is consistent and has efficiency gain over the standard estimator using the same misspecified likelihood. The proofs are in the Appendix.

Proposition 2.4.1 *Assume (i) and (ii) holds, and the regularity condition A.1.1 in the Appendix holds for* $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\rho})$. *Then,* $\sqrt{n}\{\mathrm{vec}(\hat{\boldsymbol{\beta}}_{em\_env}) - \mathrm{vec}(\boldsymbol{\beta})\} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{env})$, $\sqrt{n}\{\mathrm{vec}(\hat{\boldsymbol{\beta}}_{em\_std}) - \mathrm{vec}(\boldsymbol{\beta})\} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{std})$, *and* $\mathbf{V}_{env} \leq \mathbf{V}_{std}$.

When the envelope dimension $u = r$, the envelope reduces to the standard maximum likelihood estimate. Thus, if the envelope assumptions (i) and (ii) do not hold, the envelope dimension u is expected to be the same as r. That is, even when the envelope assumptions do not hold, the EM envelope estimator performs as well as the standard EM estimator. However, following a similar argument as in Cook et al. (2010), if the assumptions (i) and (ii) hold and if the variability of the immaterial part is relatively large, then the efficiency gain would be substantial. In other words, if the noise unrelated to the regression is large, then the envelope method, which subtracts this part of the noise from the estimation, has substantial efficiency gain. Consequently, fewer sample sizes are needed to detect the same effect size for our method as compared with the standard EM. It is also worth mentioning that the closed form of $\mathbf{V}_{env}$ is in general difficult to calculate with no closed form in most cases. We suggest to use bootstrap method to examine the standard error of $\hat{\boldsymbol{\beta}}_{em\_env}$.

Proposition 2.4.1 requires the correct specification of the distribution of $\mathbf{X}_i$ and $\boldsymbol{\varepsilon}_i$. However, we extend in Proposition 2.4.2 the asymptotic normality of missing data envelope estimator when both $\mathbf{X}_i$ and $\boldsymbol{\varepsilon}_i$ are misspecified. Note that the definition of envelope model is changed to defined directly on the parameters following Cook and Zhang (2015a). Here, the working distribution of error term is still joint Gaussian distribution as it is the most commonly used distribution for multivariate regression.

Proposition 2.4.2 *Assume (i) and (ii) holds, the error term* $\boldsymbol{\varepsilon}_i$ *and covariates* $\mathbf{X}_i$ *have finite* $(4+\delta)$*-th moment for some* $\delta > 0$, *and the regularity condition*

A.1.1 in the Appendix holds for $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\rho})$. *If we obtain our EM envelope estimator* $\hat{\boldsymbol{\beta}}_{em\_env}$ *by treating* $\mathbf{X}_i$ *and* $\boldsymbol{\varepsilon}_i$ *as normally distributed, then we have* $\sqrt{n}\{\text{vec}(\hat{\boldsymbol{\beta}}_{em\_env}) - \text{vec}(\boldsymbol{\beta})\} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{env})$, $\sqrt{n}\{\text{vec}(\hat{\boldsymbol{\beta}}_{em\_std}) - \text{vec}(\boldsymbol{\beta})\} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{std})$.

2.5 Simulations

2.5.1 Normal errors

In this subsection, we compare six different estimators: the EM envelope estimator $\hat{\boldsymbol{\beta}}_{em\_env}$, the complete case (CC) envelope estimator $\hat{\boldsymbol{\beta}}_{cc\_env}$, the full data envelope $\hat{\boldsymbol{\beta}}_{full\_env}$, the standard EM estimator $\hat{\boldsymbol{\beta}}_{em\_std}$, the standard complete case (CC) estimator $\hat{\boldsymbol{\beta}}_{cc\_std}$, and the full data MLE $\hat{\boldsymbol{\beta}}_{full\_std}$. The complete case estimators only utilize the observations that do not have any predictors or responses missing, whereas the full data estimators use the full data without any missingness. In practice, the full data estimators cannot be calculated with the missing data. The full data envelope sets a theoretical maximal efficiency possibly gained from incorporating the envelope structures. Jia et al. (2010) compared the envelope method with ridge regression and Curds and Whey introduced by Breiman and Friedman (1997). They concluded that the envelope model has the best performance when $u < p < r < n$ in the classical domain. Therefore, we only compare our method with ordinary least squares estimators. We carry out the simulations in the following steps.

Step 1. Set the population size $n = 500$. Generate parameters $\tilde{\boldsymbol{\Gamma}} \in \mathbb{R}^{r\times u}$, $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{r\times p}$, where $r = 20$, $p = 5$ and $u = 3$, and the elements are independently generated from $U(0, 1)$ and $U(-10, 10)$. By Schmidt orthogonalization, we get $\boldsymbol{\Gamma}$ from $\tilde{\boldsymbol{\Gamma}}$, where $\boldsymbol{\Gamma}$ satisfies $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_{u\times u}$. Set the true regression coefficients as $\boldsymbol{\beta} = \mathbf{P}_{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\beta}}$. Generate a matrix $\mathbf{N} \in \mathbb{R}^{p\times p}$ and set $\boldsymbol{\Sigma}_x = \mathbf{N}\mathbf{N}^T$, $\boldsymbol{\Sigma}_\varepsilon = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, where $\boldsymbol{\Omega} = 0.1\mathbf{I}_r$, $\boldsymbol{\Omega}_0 = 1000\mathbf{I}_r$.

Step 2. Generate the full data $(\mathbf{X}_i, \mathbf{Y}_i)$ for each individual i, where $\mathbf{X}_i \overset{i.i.d}{\sim} N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and $\mathbf{Y}_i|\mathbf{X}_i \overset{i.i.d}{\sim} N(\boldsymbol{\beta}\mathbf{X}, \boldsymbol{\Sigma}_\varepsilon)$ and each element of $\boldsymbol{\mu}_x$ is generated from $U(-10, 10)$.

Step 3. Generate the missingness as follows. Set three missingness mechanisms for the predictors as $\text{logit}\mathbb{P}(R_{X_{i,4}} = 1|x_{i,1}, x_{i,2}, x_{i,3}) = 1 - x_{i,1} - 2x_{i,2} - 3x_{i,3}$, $\text{logit}\mathbb{P}(R_{X_{i,3}} = 1|x_{i,1}, x_{i,4}) = 1 - x_{i,1} - 2x_{i,4}$, and $\text{logit}\mathbb{P}(R_{X_{i,5}} = 1|x_{i,1}) = 1 - x_{i,1}$. Also, set five missingness mechanisms for the responses as $\text{logit}\mathbb{P}(R_{Y_{i,2}} =$

$1, R_{Y_{i,4}} = 1|x_{i,1}, y_{i,8}, y_{i,9}) = 2 - x_{i,1} - y_{i,8} - 3y_{i,9}$, $\text{logit}\mathbb{P}(R_{Y_{i,3}} = 1|x_{i,2}, y_{i,4}, y_{i,6}) = 1 - x_{i,2} - 3y_{i,4} - y_{i,6}$, $\text{logit}\mathbb{P}(R_{Y_{i,7}} = 1, R_{Y_{i,8}} = 1, R_{Y_{i,9}} = 1|y_{i,1}, y_{i,2}, y_{i,3}) = 2 - 2y_{i,1} - y_{i,2} - 3y_{i,3}$, $\text{logit}\mathbb{P}(R_{Y_{i,1}} = 1, R_{Y_{i,10}} = 1|x_{i,1}, x_{i,2}) = 1 - x_{i,1} - x_{i,2}$ and $\text{logit}\mathbb{P}(R_{Y_{i,5}} = 1, R_{Y_{i,6}} = 1|x_{i,1}, x_{i,2}, y_{i,1}, y_{i,10}) = 1 - x_{i,1} - x_{i,2} - y_{i,1} - y_{i,10}$. For each individual, we randomly choose one missingness mechanism for the predictors and one missingness mechanism for the responses. Then, we generate the missingness indicators $(R_{X_{i,1}}, \ldots, R_{X_{i,p}}, R_{Y_{i,1}}, \ldots, R_{Y_{i,r}})$, for $i = 1, \ldots n$. We obtain the observed data for predictors and responses.

Step 4. Calculate $\hat{\boldsymbol{\beta}}_{em\_env}$, $\hat{\boldsymbol{\beta}}_{cc\_env}$, $\hat{\boldsymbol{\beta}}_{full\_env}$, $\hat{\boldsymbol{\beta}}_{em\_std}$, $\hat{\boldsymbol{\beta}}_{cc\_std}$, and $\hat{\boldsymbol{\beta}}_{full\_std}$, where $\hat{\boldsymbol{\beta}}_{em\_env}$ is calculated from the EM envelope algorithm using BIC_Q to select the envelope dimension.

Step 5. Repeat the Steps 2–5 for 1000 times.

Under the missingness mechanisms above, each predictor suffers from about 10%–15% missingness and each response about 5%–10%. The median MSEs are 4.44×10^{-5}, 2.00×10^{-4}, 1.02×10^{-5}, 5.34×10^{-2}, 0.69 and 5.23×10^{-2} for the EM envelope, the complete case envelope, the full data envelope, the standard EM, the standard complete case analysis and the full data MLE, respectively. Detailed comparisons of the six estimators are given in Figure A.1 and Table A.1 in the Appendix. For the EM envelope estimator, by using BIC_Q to choose the envelope dimension, out of 1000 times of simulations, we correctly estimated the envelope dimension $u = 3$ at an accuracy of 98.6%. The envelope dimension $u = 2$ is selected 12 times and $u = 4$ is selected 2 time. The overselection $u = 4$ still provides a correct model, although the point estimate may not be as efficient as compared with that using the correct u. The underestimation of $u = 2$ could introduce some bias. As expected, the standard complete case analysis suffers from both large variance and large bias. In contrast, the EM envelope is asymptotically unbiased and the most efficient among the four estimators using the observed data, despite the occasional underestimation of u. In this simulation setting, the variance of the immaterial part of the responses is relatively large. Thus, by eliminating the variability of the immaterial part, the EM envelope estimate outperforms the standard EM. This confirms the efficiency gain in Proposition 2.4.1. Similar to the illustrative example in Section **??**, the EM envelope also outperforms the full data MLE in this simulation, emphasizing the advantage of incorporating a dimension reduction method to recover the efficiency loss due to missing data. The performance of the EM envelope is close to the full

data envelope in this case.

In this specific setting, the complete case envelope outperforms the standard EM. This is an interesting case as the complete case envelope is biased but the standard EM is not. However, the ordering of the two is not certain in general. Intuitively, if the proportion of missingness is low, the complete case envelope estimate resembles the EM envelope estimate, and thus outperforms the standard EM. If the proportion of missingness is high, the complete case envelope is both biased and inefficient while the standard EM is still unbiased although inefficient. When the bias of the complete case envelope dominates the MSE, the standard EM outperforms the complete case envelope. When the proportion of missingness is not at extremes (too high or too low), the complete case envelope is not necessarily better or worse than the standard EM. The standard EM estimate may have a smaller bias but a relatively larger variance while the complete case envelope may have a larger bias and a smaller variance.

We carried out another simulation study, where the steps were the same as above, except we replaced $\boldsymbol{\Omega}_0 = 1000\mathbf{I}_q$ with $\boldsymbol{\Omega}_0 = 10\mathbf{I}_q$ in Step 2. This is a case where the variance of the immaterial part is not as large. The median MSEs of the EM envelope, the complete case envelope, the full data envelope, the standard EM, the standard complete case analysis and the full data MLE are: 1.06×10^{-4}, 6.16×10^{-4}, 8.58×10^{-5}, 5.42×10^{-4}, 6.81×10^{-3} and 5.24×10^{-4}. Detailed comparisons of the six methods are given in Figure A.2 and Table A.2 in the Appendix. Out of 1000 simulations, the envelope dimension is correctly estimated as $u = 3$ with an accuracy of 89.8%, while the rest 10.2% yields an estimated envelope dimension $u > 3$. As mentioned, overselection can still provide us with the correct model, but may lead to inefficient estimation. The EM envelope and the standard complete case analysis remain the best and the worst estimators using the observed data in terms of the MSEs, the standard EM now outperforms the complete case envelope. Again, the EM envelope outperforms the full data MLE.

2.5.2 Non-normal errors

In order to confirm the efficiency gain in Proposition 2.4.2, we did several sets of simulations to compare $\hat{\boldsymbol{\beta}}_{em\_env}$ with $\hat{\boldsymbol{\beta}}_{em\_std}$ when the error term $\boldsymbol{\varepsilon}_i$ is not normal. We set the sample size $n = 500$, $r = 10$, $p = 5$, and $u = 2$. Again, we generate $\tilde{\boldsymbol{\Gamma}} \in \mathbb{R}^{r\times u}$ and $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{r\times p}$ from $U(0,1)$ and $U(-10,10)$, and get the final form of $\boldsymbol{\Gamma}$ and $\boldsymbol{\beta}$ by Schmidt orthogonalization and projection of $\tilde{\boldsymbol{\beta}}$ onto span($\boldsymbol{\Gamma}$). After generating the

positive definite matrix $\mathbf{\Sigma}_x$, we generate $\mathbf{X}_i \overset{i.i.d}{\sim} t_5(\mathbf{0}, \mathbf{\Sigma}_x)$, where $t_\nu(\boldsymbol{\mu}, \mathbf{\Sigma})$ represent the multivariate t distribution with location parameter $\boldsymbol{\mu}$, scale parameter $\mathbf{\Sigma}$ and degrees of freedom ν. In order to satisfy the independence conditions $\mathbf{\Gamma}_0^T\mathbf{Y}_i \perp\!\!\!\perp \mathbf{X}_i$ and $\mathbf{\Gamma}^T\mathbf{Y}_i \perp\!\!\!\perp \mathbf{\Gamma}_0^T\mathbf{Y}_i|\mathbf{X}_i$, we generate $\boldsymbol{\varepsilon}_i$ through the following procedure. Firstly draw elements of $\boldsymbol{\varepsilon}_{i1} \in \mathbb{R}^u$ and $\boldsymbol{\varepsilon}_{i2} \in \mathbb{R}^{r-u}$ independently of two distributions f_1 and f_2. Then we set $\boldsymbol{\varepsilon}_i = \mathbf{\Gamma}\boldsymbol{\varepsilon}_{i1} + \mathbf{\Gamma}_0\boldsymbol{\varepsilon}_{i2}$. We simulated three sets of f_1 and f_2:

1. $f_1 \sim t_5(0, 1)$, $f_2 \sim t_5(0, 100)$.
2. $f_1 \sim U(-1, 1)$ and $f_2 \sim U(-5, 5)$.
3. $f_1 \sim \text{Laplace}(0, 1)$ and $f_2 \sim \text{Laplace}(0, 10)$.

Then, we get the response variable $\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i$. After that, we generate missingness using the same way as in the previous subsection.

The median MSE of $\hat{\boldsymbol{\beta}}_{em\_env}$ for the t, uniform and Laplace distributions are 1.23×10^{-3}, 8.42×10^{-5}, 1.45×10^{-3}. For $\hat{\boldsymbol{\beta}}_{em\_env}$, the median MSEs are 7.67×10^{-3}, 4.16×10^{-4}, 9.48×10^{-3} respectively. Also, all the envelope dimensions are correctly specified through the bootstrap method. Detailed comparisons are given in Table A.3. We see that even the error terms are not normal, as long as the envelope independence conditions hold, our EM envelope estimator outperforms the standard estimator.

2.6 Data Analysis

In this section, we apply our proposed method to the Chronic Renal Insufficiency Cohort (CRIC) study. The CRIC study recruited 3939 participants from April 8, 2003 through September 3, 2008 and continued through March 31, 2013 (Feldman et al., 2003). The study cohort was a racially and ethnically diverse group aged from 21 to 74 years with mild to moderate chronic kidney disease (CKD). Each study subject was given extensive clinical evaluation, and the information collected included quality of life, dietary assessment, physical activity, health behaviors, depression, cognitive function, and blood and urine specimens.

To prevent the development of severe clinical events, it is important to identify the CKD patients with high risk of end-stage renal diseases (ESRD) in their early stages. A variety of risk factors for ESRD have been identified in the literature (Budoff et al., 2011; He et al., 2012; Madjid and Fatemi, 2013; Bansal et al., 2013;

Ferguson et al., 2013; Anderson et al., 2015). It is of interest to investigate the difference in the distributions of baseline biomarkers among the patients who develop ESRD versus who do not. Correlation among risk factors have often been observed in the literature (Capuano et al., 2003); however, it has not been fully utilized in the statistical analyses for predicting ESRD and CVD. Our method leveraged the correlation among the risk factors and biomarkers to improve the efficiency of the analysis. Additionally, it is of interest to explore modifiable biomarkers, which are the biomarkers that are significantly differently distributed for patients who develop ESRD adjusting for the established biomarkers.

The study participants were distinguished by the ESRD status (binary, 1 for ESRD and 0 for no ESRD) within five years of enrollment. We assumed death before the progression of ESRD and withdraw from study were independent of the ESRD disease status. Thus, we focused our analysis on the remaining 3205 patients. In our analysis, we also adjusted for gender, age, race, systolic, and diastolic blood pressures, and hemoglobin. The biomarkers and risk factors are urine albumin, urine creatinine, high sensitivity C-reactive protein (HS_CRP), brain natriuretic peptide (BNP), chemokine ligand 12 (CXCL12), fetuin A, fractalkine, myeloperoxidase (MPO), neutrophil gelatinase associated lipocalin (NGAL), fibrinogen, troponini, urine calcium, urine sodium, urine potassium, urine phosphate, high sensitive troponin T (TNTHS), aldosterone, C-peptide, insulin value, total parathyroid hormone (Total PTH), CO_2, 24-hour urine protein, and estimated glomerular filtration rate (EGFR). We performed a log transformation on the highly skewed biomarkers and risk factors. In addition, we divided fetuin A by 10^4 as its scale was quite different from other biomarkers.

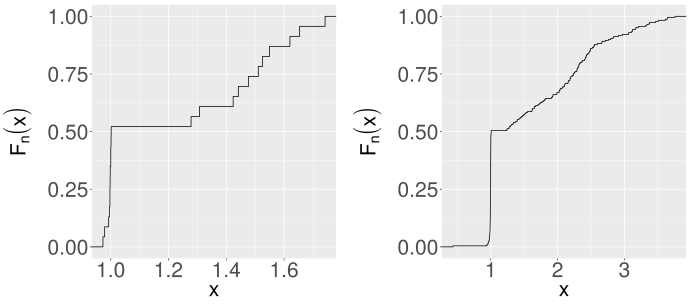
We first assessed the difference in the distributions of baseline biomarkers versus the ESRD status, unadjusted for the established biomarkers. All the biomarkers except the EGFR had some missingness ranging from <1% to 6%. Also, as for the predictors, hemoglobin and BMI had relatively low missing rate (there are 15 observations with hemoglobin missing and 5 observations with BMI missing). As the proportion of missingness was relatively low, we used the BIC_Q given in Section 2.4.2 to select the envelope dimension. The EM envelope method reduced the dimension of the biomarkers from $r = 23$ to $u = 15$. The point estimates, bootstrap standard errors, confidence intervals and $p-$values for the mean difference of biomarkers among ESRD patients versus no ESRD patients are given in the Appendix. The magnitude of the point estimates of our method is in general slightly smaller than those of the standard EM. For example, the coefficient for urine albumin is 0.56

using our method and 2.54 using the standard EM. This is because in each EM iteration, the envelope estimate is the projection of the standard estimates onto the envelope direction. The reduction in the magnitude is interpreted as the noise subtracted from the original estimates. As we mentioned in Section 2.4.3, the closed form of the standard errors of our method are difficult to obtain. Hence, we carried out the nonparametric bootstrap for 1000 times, that is, we resample individuals with replacement. The standard errors of our method is also generally smaller than those of the standard method. For example, Figure 3.3 further shows the empirical cumulative density distributions of the estimated standard errors of the standard EM versus our method. Again, the estimated standard errors are in general smaller (on the right hand side of 1 in Figure 3.3) using our method than using the standard EM indicating the efficiency gain using our method. The mean of the ratio is 1.24 for coefficients corresponding to ESRD and 1.62 for all coefficients. That is, on average, our method is about 24% more efficient than the standard method for the coefficients corresponding to ESRD and 62% more efficient for all coefficients. The same set of biomarkers (all the aforemenioned biomarkers except HS CRP, fetuin A and insulin value) were found by our method and the standard EM, to be significantly different among patients with and without ESRD. Table A.4 and Table A.5 in the Appendix present details of the results.

It is found in the literature that although many novel biomarkers are found to be marginally significantly associated with the ESRD status, such an association often disappears after adjusting for the established biomarkers (Foster et al., 2015; Park et al., 2017; Inker et al., 2017). That is, they are not as useful as modifiable biomarkers. We next assess the mean difference of baseline biomarkers among patients with and without the ESRD status, adjusted for the established biomarkers. The EGFR and the amount of urine protein excreted are two established biomarkers for predicting the ESRD. Thus, in the subsequent analysis, we use the two variables as predictors rather than responses. The estimated envelope dimension is $u = 17$. The point estimates, bootstrap standard errors, confidence intervals and $p-$values for the mean difference of biomarkers for different ESRD status adjusting for the EGFR and the urine protein are given in Table A.4. The point estimates and the standard errors are again in general smaller using our method as compared with using the standard EM. Figure 2.4 shows the empirical distribution of the ratio between the estimated standard errors of the two methods. The mean of the ratio is 1.92 for coefficients corresponding to the ESRD and 1.86 for all coefficients. Comparing Figure 3.3 and Figure 2.4, we see that the EM envelope method achieves even

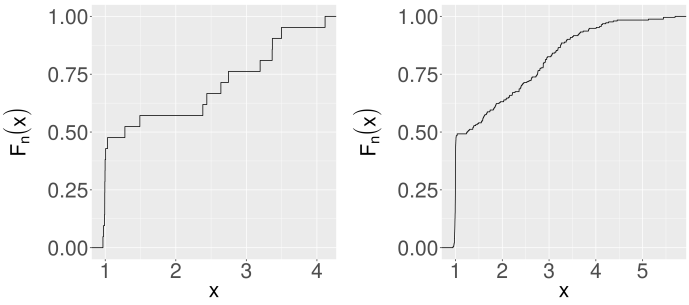
higher efficiency gain when we adjust for the established biomarkers versus not. As found in the literature, after adjusting for the established biomarkers, the majority of biomarkers that have been investigated are no longer significant. We observe the same phenomenon using both our method and the standard EM. However, among the few biomarkers that remain significant, there is some discrepancy between the standard EM and our method: our method found HS CRP, aldosterone, and C-peptide significant which were not shown in standard EM; whereas standard EM found NGAL, which was not found in our method. As our method is more efficient for finite sample, the results of which are more precise than those of the standard EM.

Figure 2.3: The empirical cumulative distribution of the ratio between the standard errors of the standard EM and our method without adjusting for the established biomarkers.



(a) Coefficients for ESRD (b) All coefficients

Figure 2.4: The empirical cumulative distribution of the ratio between the standard errors of the standard EM and our method adjusted for the established biomarkers.



(a) Coefficients for ESRD (b) All coefficients

2.7 Discussion

In this paper, we proposed the EM envelope method to achieve more efficient estimation for coefficients in the multivariate regression with missing data. Specifically, we assumed the redundancy exists in the response variables and thus could be omitted in the regression to reduce noise. A similar redundancy structure may also occur among the predictors or among both predictors and responses. Our method can be similarly derived under those scenarios. For example, if we assume there exist a linear combination of predictors that does not contribute to the regression and assume the missingness mechanism of predictors and responses are MAR, then our method could be adapted to gain efficiency by discarding the immaterial part of the variance among the predictors. A similar derivation can be made by changing the covariance matrix $\mathbf{\Sigma}$ in this paper to $\mathbf{\Sigma}_x$, the covariance matrix of predictors.

An alternative approach to calculate an envelope estimate with missing data is to use the model free approach proposed by Cook and Zhang (2015a). Specifically, we can calculate the standard EM estimator together with its asymptotic variance using the Louis formula. However, the calculation of the asymptotic variance of the EM estimator requires calculating the conditional expectation of the outer product of the complete data score vector, an inherently problem-specific task that usually requires much computational effort as discussed in Meng and Rubin (1991). Also, this method requires estimating an envelope in $\mathbb{R}^{pq}$ space instead of $\mathbb{R}^q$, which makes the problem more challenging. A detailed comparison of the empirical performances of such model free envelope based on the standard EM estimator versus the EM envelope method is left for future work.

Envelope method has been generalized to GLM (Cook and Zhang, 2015a) with univariate response. How to adapt GLM envelope method with multiple responses even without missing data is still an open problem. Hence, our paper only focused on linear model envelope method, which is the most widely used case.

Throughout this paper, our method is proposed assuming the missing data mechanism is ignorable. When the data is nonignorably missing, a selection model is needed to be specified. We also leave it as a future research topic.

2.8 Software

The corresponding R package is available at `https://github.com/mlqmlq/missing_env`.

Chapter 3

Mixed effects envelope models

3.1 Introduction

3.1.1 Literature review

Over the past three decades, an increasing amount of literature has emerged on the topic of sufficient dimension reduction (SDR). Li (1991) proposed the sliced inverse regression to reduce the dimension of the predictors. That is, assuming the response only depends on a linear combination of the predictors, one regresses the predictors $\mathbf{X}$ against the response $\mathbf{Y}$ to circumvent any model-fitting process. Cook (1998) defined the central subspace as the subspace with the minimal dimension such that the response is independent of the predictors given the projection of the predictors onto the space. Other SDR methods include but not limited to sliced average variance estimation (Cook and Weisberg, 1991a), principal Hessian direction (Li, 1992), contour regression (Li et al., 2005), inverse regression estimation (Cook and Ni, 2005), directional regression (Li and Wang, 2007), likelihood-acquired directions (Cook and Forzani, 2009), discretization-expectation estimation (Zhu et al., 2010), non-elliptically distributed predictors (Li and Dong, 2009; Dong and Li, 2010), dimension reduction based on canonical correlation (Fung et al., 2002; Zhou and He, 2008), and average partial mean estimation (Zhu et al., 2010). However, the aforementioned methods all focus on the dimension reduction of predictors with univariate response.

Recently, Cook et al. (2010) proposed a new sufficient dimension reduction method called the envelope method to reduce the dimension of *responses* in multivariate regression. Specifically, Cook et al. (2010) considered the following multi-

variate linear regression model

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i, \tag{3.1}$$

where i indicates the i^{th} individual, $\mathbf{Y}_i,\ \boldsymbol{\varepsilon}_i \in \mathbb{R}^r$, $\mathbf{X}_i \in \mathbb{R}^p$ and the parameter of interest is $\boldsymbol{\beta}$.

The key idea of the envelope method is to assume the existence of redundancy in responses that do not contribute to the estimation of $\boldsymbol{\beta}$ in model (3.1), so that the estimation of $\boldsymbol{\beta}$ is more efficient by leveraging this condition. Cook et al. (2010) assumes that there exists an orthogonal matrix $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{r\times r}$, where $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ and $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{r\times(r-u)}$, with $0 \le u \le r$ satisfying the following conditions:

Condition 3.1.1 $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i \perp\!\!\!\perp \mathbf{X}_i$;

Condition 3.1.2 $\boldsymbol{\Gamma}^T\mathbf{Y}_i \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{Y}_i \mid \mathbf{X}_i$,

where $\perp\!\!\!\perp$ indicates independence. Under Conditions 1 and 2, $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i$ is redundant for a fixed effect regression (Cook et al., 2010). The $\boldsymbol{\Sigma}_\varepsilon$-envelope is uniquely defined to be the smallest subspace satisfying these conditions. Once the basis $\widehat{\boldsymbol{\Gamma}}$ is obtained, the envelope estimator is obtained by projecting the ordinary least square estimator onto the estimated envelope space. Cook et al. (2010) showed that the envelope estimator can achieve efficiency gain over the OLS estimator. Following the definition in their paper, we define the variance of $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i \mid \mathbf{X}_i$ as the material part variance, and $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i$ as the immaterial part variance. The efficiency gain will be substantial if the variation of the immaterial part is relatively large as compared with that of the material part.

The envelope methods have been developed in different settings, including response envelope (Cook et al., 2010), partial envelope (Su and Cook, 2011), inner envelope (Su and Cook, 2012), scaled envelope (Cook and Su, 2013), predictor envelope (Cook et al., 2013), reduced rank envelope (Cook et al., 2015), simultaneous envelope (Cook and Zhang, 2015b), model-free envelope (Cook and Zhang, 2015a), and tensor envelope (Li and Zhang, 2017).

Longitudinal data, also known as panel data, collects repeated measurements of the same subjects over time. As a distinctive feature of longitudinal data, measures that are collected repeatedly over time are typically correlated and redundancy typically exists among responses. Hence, reducing data to lower dimensions can improve efficiency while still preserving all relevant information on regression. Additionally,

data may be unbalanced in the sense that subjects are not measured at the same time points. Moreover, the predictors may depend on time and may have different trajectories over time across individuals. For example, in a study of the effect of smoking on body weight, the number of cigarettes smoked per day may stay the same for some individuals but not for others. These features distinguish the longitudinal data from the cross-sectional data in terms of appropriate analyses. However, the study of sufficient dimension reduction for longitudinal data is quite limited. Pfeiffer et al. (2012) developed first-moment sufficient dimension reduction techniques to replace the original predictors with longitudinal nature. Bi and Qu (2015) applied the quadratic inference function to longitudinal data sufficient dimension reduction. The literature is even more scarce on dimension reduction in mixed effects models with high-dimensional response. Some early work has been done by Zhou et al. (2010) using a reduced rank model for spatially correlated hierarchical functional data and by Hughes and Haran (2013) using sparse reparameterization to reduce spatial confounding.

In this chapter, we propose a mixed effects envelope model. Similar to the standard mixed effects model, the variability of each observation are composed of the within-individual variability and the between-individual variability. The mixed effects envelope model recovers the distribution of the unobserved between-individual random coefficient and reduces noises in the within-individual variation. The mixed effects envelope model inherits both the efficiency gain of the standard envelope model and the flexibility of the standard mixed effects model. Specifically, our methods result in more efficient estimators than those from the standard mixed effects models and can be used for unbalanced data as well as with time-varying predictors.

3.1.2 Notation

Consider a study with n individuals with each individual being measured at a total of J_i time points, where $i = 1, \ldots, n$ and $j = 1, \ldots, J_i$. Let $J_\bullet = \sum_{i=1}^{n} J_i$ denote the total number of the observations across time. For an individual i at time j, let $\mathbf{Y}_{ij} \in \mathbb{R}^r$ denote the responses of length r, let $\mathbf{X}_{ij} \in \mathbb{R}^p$ denote the vector of predictors of length p, and let $\mathbf{Z}_{ij} \in \mathbb{R}^q$ denote the vector of predictors of length q. Predictors $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ can either be stochastic or nonstochastic. Let $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \ldots \mathbf{Y}_{iJ_i})$ denote the responses, and $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots \mathbf{X}_{iJ_i})$, $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \ldots \mathbf{Z}_{iJ_i})$ denote the predictors for one individual at all time points. Let $\mathbb{R}^{r \times p}$ denote the class of all matrices with size

$r \times p$. Let $\mathbb{S}^{r\times r}$ denote the class of all symmetric positive definite matrices of size r. Let $\mathbf{I}_r$ denote the identity matrix of size r. Let $\text{vec}(\cdot)$ denote the vectorization of a matrix by stacking the columns of the matrix on top of one another, and let $\text{vech}(\cdot)$ denote the vectorization of the unique part of each column that lies on or below the diagonal. Let $\dagger$ denote the Moore-Penrose inverse. Also, let $\mathbf{E}_r \in \mathbb{R}^{r^2\times r(r+1)/2}$ to be the expansion matrix such that $\text{vec}(\cdot) = \mathbf{E}_r\text{vech}(\cdot)$ and $\mathbf{C}_r \in \mathbb{R}^{r(r+1)/2\times r^2}$ to be the contraction matrix such that $\text{vech}(\cdot) = \mathbf{C}_r\text{vec}(\cdot)$. Let $\widetilde{\mathbf{Y}}_i = \text{vec}(\mathbf{Y}_i)$ denote the vectorized responses. Let $\mathbf{A} \otimes \mathbf{B}$ denote the kronecker product between $\mathbf{A}$ and $\mathbf{B}$. All the population covariance matrices in this paper are positive definite. We use $\text{span}(\mathbf{A})$ to denote the span of the column vectors of $\mathbf{A}$.

We organize this chapter as follows. In Section 3.2, we give the standard mixed effects model and discuss a special case where classic envelope can be directly applied. In Section 3.3, we propose the mixed effects envelope model as well as provide a graphical illustration of our method. We further illustrate our proposed method in the simulations in Section 3.4 and data analysis in Section 3.5. We conclude with a brief discussion in Section 3.6.

3.2 Preliminary

3.2.1 Mixed effects model

Consider the mixed effects model

$$\mathbf{Y}_{ij} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_{ij} + \mathbf{b}_i\mathbf{Z}_{ij} + \boldsymbol{\varepsilon}_{ij},$$

where $\boldsymbol{\alpha} \in \mathbb{R}^r$ is the intercept, $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ denotes the coefficient for the fixed effects and $\mathbf{b}_i \in \mathbb{R}^{r\times q}$ denotes the random coefficients. Assume $\text{vec}(\mathbf{b}_i)$ identically and independently follows $N(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{b})$, where $\boldsymbol{\Sigma}_\mathbf{b} \in \mathbb{S}^{qr\times qr}$. The residual error $\boldsymbol{\varepsilon}_{ij}$ identically and independently follows $N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, J_i$, where $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \in \mathbb{S}^{r\times r}$. The normality of random effect and error are assumed here for simplicity. We extend our result when they have finite $(4+\delta)$-th moment in Section 3.3. The random effect $\mathbf{b}_i$ is assumed to be independent from the residual error $\boldsymbol{\varepsilon}_{ij}$, i.e., $\mathbf{b}_i \perp\!\!\!\perp \boldsymbol{\varepsilon}_{ij}$. The variance due to random coefficient $\boldsymbol{\Sigma}_\mathbf{b}$ is the between-subject variability and the variance due to the error $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is the within-subject variability. Let $\mathbf{A}_{ij} = \mathbf{Z}_{ij}^T \otimes \mathbf{I}_r$, then $\mathbf{b}_i\mathbf{Z}_{ij} = (\mathbf{Z}_{ij}^T \otimes \mathbf{I}_r)\text{vec}(\mathbf{b}_i) = \mathbf{A}_{ij}\text{vec}(\mathbf{b}_\text{i})$. The covariance of responses across time for the same individual is correlated $\text{Cov}(\mathbf{Y}_{ij}, \mathbf{Y}_{ij'} \mid \mathbf{X}_{ij}, \mathbf{X}_{ij'}, \mathbf{Z}_{ij}, \mathbf{Z}_{ij'}) = \mathbf{A}_{ij}\boldsymbol{\Sigma}_\mathbf{b}\mathbf{A}_{ij'}^T$

if $j \neq j'$, and $\text{Var}(\mathbf{Y}_{ij} \mid \mathbf{X}_{ij}, \mathbf{Z}_{ij}) = \mathbf{A}_{ij}\boldsymbol{\Sigma}_{\mathbf{b}}\mathbf{A}_{ij}^T + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. Let $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_{i1}, \ldots \boldsymbol{\varepsilon}_{iJ_i})$, we can rewrite the model above in a matrix form as

$$\mathbf{Y}_i = \boldsymbol{\alpha} \otimes \mathbf{1}_{J_i}^T + \boldsymbol{\beta}\mathbf{X}_i + \mathbf{b}_i\mathbf{Z}_i + \boldsymbol{\varepsilon}_i. \tag{3.2}$$

Model (3.1) is a special case of model (3.2) when $\boldsymbol{\Sigma}_{\mathbf{b}} = \mathbf{0}$, and $J_i = 1$ for $i = 1, \ldots, n$.

3.2.2 Classic envelope model for a special case of longitudinal data

The classic envelope method can be applied to longitudinal data in a special case: when the data is balanced, the predictors do not vary with time, and random slopes are not included in the model (random intercepts are included). We will show that under this setting, the mixed effects model naturally contains an envelope structure over the observations across time. Under this setting, $J_i = J$ for any i. Also, if we assume $\mathbf{X}_i$ does not vary with time, then (3.2) can be written as

$$\widetilde{\mathbf{Y}}_i = \mathbf{1}_J \otimes \boldsymbol{\alpha} + (\mathbf{1}_J \otimes \boldsymbol{\beta})\mathbf{X}_i + \widetilde{\boldsymbol{\varepsilon}}_i, \tag{3.3}$$

where $\widetilde{\boldsymbol{\varepsilon}}_i \in \mathbb{R}^{rJ}$ i.i.d follows $N(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_{\varepsilon})$ and $\widetilde{\boldsymbol{\Sigma}}_{\varepsilon} = \mathbf{I}_J \otimes \boldsymbol{\Sigma}_{\mathbf{b}} + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. This model is a standard multivariate model, hence we can impose an envelope model on it. Let $\mathcal{E}_{\widetilde{\boldsymbol{\Sigma}}_{\varepsilon}}(\widetilde{\mathcal{B}})$ denote the $\widetilde{\boldsymbol{\Sigma}}_{\varepsilon}$-envelope for $\widetilde{\mathcal{B}}$, where $\widetilde{\mathcal{B}} = \text{span}(\mathbf{1}_J \otimes \boldsymbol{\beta})$. The structure of $\mathcal{E}_{\widetilde{\boldsymbol{\Sigma}}_{\varepsilon}}(\widetilde{\mathcal{B}})$ is given in the following proposition and corollary.

Proposition 3.2.1 *Under model* (3.3)*, the basis for* $\mathcal{E}_{\widetilde{\boldsymbol{\Sigma}}_{\varepsilon}}(\widetilde{\mathcal{B}})$ *is* $\mathbf{1}_J \otimes \boldsymbol{\Phi}$*, where* $\boldsymbol{\Phi}$ *is the basis for* $\mathcal{E}_{(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}/J + \boldsymbol{\Sigma}_{\mathbf{b}})}(\mathcal{B})$*.*

Corollary 3.2.1 *Under model* (3.3)*, the dimension of* $\mathcal{E}_{\widetilde{\boldsymbol{\Sigma}}_{\varepsilon}}(\widetilde{\mathcal{B}}) \subseteq \mathbb{R}^{rJ}$ *cannot exceed* r*.*

Intuitively, although the repeated measures from the same individual are correlated, because neither the fixed effects nor the random effects change over time, we can reduce the dimension of responses by averaging each individual over different time points. That is, model (3.2) naturally results in combinations of the responses of dimension $r(J-1)$ that do not contribute to the regression. This results in a $\widetilde{\boldsymbol{\Sigma}}_{\varepsilon}$-envelope $\mathcal{E}_{\widetilde{\boldsymbol{\Sigma}}_{\varepsilon}}(\widetilde{\mathcal{B}})$ with envelope dimension no greater than r rather than rJ.

Proposition 3.2.1 presents a simple but important observation: if the true model is a mixed effects model but instead we fit a standard multivariate linear regression,

even we have a reduced dimension from rJ to r by the envelope method, we do not gain additional efficiency. This is because the failure to leverage the mixed effects model structure creates redundancy. Such an observation naturally leads us to explore an envelope model that can incorporate the mixed effects model structure to gain further efficiency.

3.3 The mixed effects envelope model

3.3.1 Conditions

Now, we propose the mixed effects envelope model. The key requirement of the classic envelope method is the existence of some linear combination of the responses that do not contribute to the regression. With longitudinal data, because both $\mathbf{X}_i$ and $\mathbf{Z}_i$ are observed predictors, it may seem natural to extend Conditions 1 and 2 by replacing $\mathbf{X}_i$ with $(\mathbf{X}_i, \mathbf{Z}_i)$ as

Condition 1°. $\boldsymbol{\Gamma}_0^T \mathbf{Y}_i \perp\!\!\!\perp (\mathbf{X}_i, \mathbf{Z}_i)$;

Condition 2°. $\boldsymbol{\Gamma}^T \mathbf{Y}_i \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T \mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{Z}_i$.

It has been shown that the standard envelope Conditions 1 and 2 are equivalent to the reparameterization $\text{span}(\boldsymbol{\beta}) \subseteq \text{span}(\boldsymbol{\Gamma})$ and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, where $\boldsymbol{\Omega} = \boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Gamma}$, and $\boldsymbol{\Omega}_0 = \boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Gamma}_0$. Unlike Conditions 1 and 2 which impose conditions on population parameters, Conditions 1° and 2° are equivalent to requiring certain relationships between $\mathbf{Z}_i$ and parameters as shown in the proposition below. In general, Conditions 1° and 2° are hard to satisfy because their validity is contingent on the observed value of $\mathbf{Z}_i$ in the sample.

Proposition 3.3.1 *Conditions 1° and 2° hold under model* (3.2) *if and only if* $\boldsymbol{\Gamma}_0^T\boldsymbol{\beta} = \mathbf{0}$, $\mathbf{Z}_i \otimes \boldsymbol{\Gamma}_0 = \mathbf{0}$, *and* $\mathbf{I}_{J_i} \otimes (\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Gamma}_0) + (\mathbf{Z}_i^T \otimes \boldsymbol{\Gamma}^T)\boldsymbol{\Sigma}_{\mathbf{b}}(\mathbf{Z}_i \otimes \boldsymbol{\Gamma}_0) = \mathbf{0}$.

To modify Condition 1°, we want to find a condition that reduces to Condition 1 when there is no random effect. Recall under (3.1), $\text{vec}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)$ only depends on predictors through $\boldsymbol{\beta}\mathbf{X}_i$ in the mean and to have its distribution free of $\boldsymbol{\beta}$ is the same as to have $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i \perp\!\!\!\perp \mathbf{X}_i$. In other words, Condition 1 can be equivalently expressed as $\mathbb{E}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i \mid \mathbf{X}_i) = \mathbb{E}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)$ under linear model. However, under model (3.2), $\text{vec}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)$ depends on the predictors $(\mathbf{X}_i, \mathbf{Z}_i)$ through both mean and variance. Notice that the parameter of interest $\boldsymbol{\beta}$ only involves in the mean of $\text{vec}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)$. Thus, this motivates us to relax the distributional independence between $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i$ and

$(\mathbf{X}_i, \mathbf{Z}_i)$ to be just mean independence, i.e., $\mathbb{E}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{Z}_i) = \mathbb{E}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)$ so that this condition reduces to Condition 1 when there is no random effect.

To modify Condition 2°, we also want to find a condition that reduces to Condition 2 in the absence of the random effect. Note $\text{vec}(\boldsymbol{\Gamma}^T\mathbf{Y}_i) \mid \text{vec}(\boldsymbol{\Gamma}_0^T\mathbf{Y}), \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i \sim N(\boldsymbol{\mu}^{***}, \boldsymbol{\Sigma}^{***})$, where

$$\boldsymbol{\mu}^{***} = \text{vec}(\boldsymbol{\Gamma}^T\boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\Gamma}^T\mathbf{b}_i\mathbf{Z}_i) + \mathbf{I}_{J_i} \otimes \left\{\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Gamma}_0(\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Gamma}_0)^{-1}\right\} \cdot \left\{\text{vec}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i) - \text{vec}(\boldsymbol{\Gamma}_0^T\boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\Gamma}_0^T\mathbf{b}_i\mathbf{Z}_i)\right\},$$

and $\boldsymbol{\Sigma}^{***} = \mathbf{I}_{J_i} \otimes (\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{\Gamma})^{-1}$. That is, if we conditional on both predictors and random effects, the conditional independence between $\boldsymbol{\Gamma}^T\mathbf{Y}_i$ and $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i$ is equivalent to $\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Gamma}_0 = 0$, i.e., $\boldsymbol{\Gamma}$ reduces $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. This condition reduces to Condition 2 when there is no random effect.

Thus, to develop the mixed effects envelope model, we assume

Condition 1*. $\mathbb{E}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{Z}_i) = \mathbb{E}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)$,

Condition 2*. $\boldsymbol{\Gamma}^T\mathbf{Y}_i \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i$. As mentioned, in the absence of random effects, Condition 1* and 2* will reduce to Condition 1 and 2 under the linear model. Conditions 1* and 2* can be viewed as extensions of Conditions 1 and 2 for longitudinal data. However, unlike the classic envelope condition (Cook et al., 2010), Condition 1* only requires the expectation of $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{Z}_i$ and $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i$ to be the same. The motivation of Condition 1* is that instead of imposing the redundancy of $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i$ on its entire distribution, we just assume the redundancy on its mean, which is easier to satisfy.

Condition 2* assumes the independence between $\boldsymbol{\Gamma}^T\mathbf{Y}_i$ and $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i$ conditional on predictors $(\mathbf{X}_i, \mathbf{Z}_i)$, as well as on the unobservable $\mathbf{b}_i$. Equivalently, the redundancy of responses is within individuals rather than across. In other words, Condition 2* excludes the possibility of $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i$ contributing to the regression through a correlation with $\boldsymbol{\Gamma}^T\mathbf{Y}_i$ given any individual, although such correlation may be present in the population. Here, different from the original envelope model, under model (3.2), $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is the variance of outcomes given predictors and random effects, which is only the within-subject variation not including the between-subject variation. Thus, when the study is balanced, even when the classic envelope model does not have much efficiency gain, the mixed effects envelope may achieve substantial efficiency gain: decomposing part of the variability into material and immaterial variability may be possible even when decomposing the total variability may not be possible. The idea

of using part of parameters to form an envelope was also adopted in the development of partial envelope (Su and Cook, 2011), where part of parameters in the mean model are used.

Other than having clear interpretations, Conditions 1* and 2* also facilitate the reparameterization of the original parameters in (3.2). Under Condition 1*, $\mathbb{E}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i \mid \mathbf{X}_i, \mathbf{Z}_i)$ is free of $\boldsymbol{\beta}$, which indicates $\text{span}(\boldsymbol{\beta}) \subseteq \text{span}(\boldsymbol{\Gamma})$. Additionally, since Condition 2* conditions on the random effects, we have $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$. We define the smallest reducing subspace that satisfies Conditions 1* and 2* as the mixed effects envelope, or $\boldsymbol{\Sigma}_\varepsilon$-mean envelope and write it as $\bar{\mathcal{E}}_{\boldsymbol{\Sigma}_\varepsilon}(\mathcal{B})$. Under Conditions 1* and 2*, model (3.2) can be written as:

$$\mathbf{Y}_{ij} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X}_{ij} + \mathbf{b}_i\mathbf{Z}_{ij} + \boldsymbol{\varepsilon}_{ij}, \tag{3.4}$$

where $\boldsymbol{\varepsilon}_{ij}$ identically and independently follows $N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$, and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$.

Under the mixed effects envelope model (3.4), the number of variational independent parameters changes from $r + rp + r(r+1)/2 + qr(qr+1)/2$ to $r + up + r(r+1)/2 + qr(qr+1)/2$. Since $u \leq r$, the number of parameters of mixed effects envelope model is no more than the standard mixed effects model. When J is large, the number of parameters in the mixed effects envelope model (3.4) can be substantially fewer than those in the standard envelope model, but there is no general relationship between the number of parameters in these two envelope models.

Under Conditions 1* and 2* and model (3.2), the covariance $\text{Var}(\widetilde{\mathbf{Y}}_i \mid \mathbf{X}_i, \mathbf{Z}_i)$ has a specific heteroscedastic error structure $\text{Var}(\widetilde{\mathbf{Y}}_i \mid \mathbf{X}_i, \mathbf{Z}_i) = \mathbf{A}_i\boldsymbol{\Sigma}_{\mathbf{b}}\mathbf{A}_i^T + \mathbf{I}_{J_i} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, where $\mathbf{A}_i = \mathbf{Z}_i^T \otimes \mathbf{I}_r$. Another heteroscedastic error model was considered in Su and Cook (2013), where the predictors are only indicators for the subpopulation and individuals in the same subpopulation have the same distribution. Following Su and Cook (2013), Park et al. (2017) generalized the multivariate envelope mean model to groupwise envelope regression models with heteroscedastic error. In their setting, the envelope is assumed to be the intersection of subspaces that contains columns of all the coefficients across populations. Under model (3.2), it is possible that $\mathbf{Z}_i$ is different for all individuals, then we have n single individual subpopulations. This situation cannot be directly handled in their framework.

3.3.2 Graphical illustration

Before diving into the estimation details of the mixed effects envelope model, we first provide a graphical illustration of the classic envelope, the standard mixed effects estimator and our mixed effects envelope estimator, under the mixed effects model. We generate the outcomes $\mathbf{Y}_i$ from (1) with $\mathbf{Z}_{ij} = 1$. To compare the classic envelope with the mixed effects envelope, we only consider the setting where data is balanced and the predictors are time-invariant.

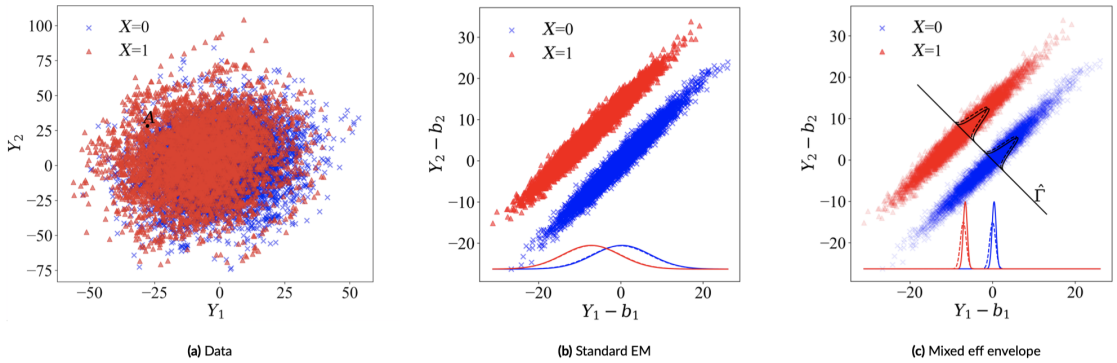
Consider two groups of individuals with $X_i = 0$ and 1 respectively. We generate $n = 2000$ individuals with $J_i = 5$ observation for each group. For individual i at time point j, we generate a bivariate response $\mathbf{Y}_{ij} = (Y_{ij1}, Y_{ij2})^T$ from the mixed effects model (1) with $\boldsymbol{\alpha} = \mathbf{0}$ and $\boldsymbol{\beta} = (-7.07, 7.07)^T$. We are interested in examining the mean group difference.

Figure 3.1a presents the raw data, where we directly implement the OLS method and the classic envelope model. The OLS estimator is $(-7.54, 6.38, -7.17, 6.83, -7.23, 6.66, -7.81, 6.28, -7.68, 6.27)^T$ with the MSE 1.10. Model (3.3) ignores the fact that some responses are repeated measures over time and does not distinguish them from different measures collected at the same time point. As a result, the OLS estimator of model (3.3) is relatively inefficient.

By applying the classic envelope method, the estimated envelope dimension is $\widehat{u} = 2$. The envelope estimate for the group difference is $(-7.46, 6.52, -7.49, 6.49, -7.41, 6.53, -7.50, 6.49, -7.57, 6.39)^T$ with the MSE 0.93. The classic envelope removes some redundancy in the responses as compared with the OLS estimator from model (3.3). However, as we show below, such redundancy can easily be removed by incorporating the mixed effects model structure.

Figure 3.1b shows the performance of the estimator under model (1) using the expectation-maximization (EM) algorithm, a standard mixed effects model estimator. For comparison, we keep the OLS estimate of $\mathbf{Y}_{ij} - \mathbf{b}_i$ in two groups when $\mathbf{b}$ is assumed observed (dashed curve in Figure 3.1b) as a benchmark. The solid curves denote the estimated distribution of $\mathbf{Y}_{ij} - \mathbf{b}_i$ using the standard EM. The density curves related to the EM algorithm are not simply projections of scattered points obtained by subtracting $\mathbf{b}$, and the scatter points just provide intuitions. The EM algorithm separates the between-subject variability from the within-subject variability when calculating the point estimates, hence, the solid and dashed density curves almost overlap completely. The group difference using the standard EM is $(-7.48, 6.48)^T$ with the MSE 0.95. The MSE is similar to that of the classic envelope

Figure 3.1: Graphical illustration of the standard mixed effects estimator and the mixed effects envelope estimator, with a random intercept. Individuals of the two groups are represented by cross dots ($X = 0$) and triangles ($X = 1$). The scatter points in Figure 3.1a demonstrate the original data, whereas Figure 3.1b and 3.1c demonstrate the data of $\mathbf{Y} - \mathbf{b}$ as if $\mathbf{b}$ is observed. Solid curves are from the EM-type estimates, and the dashed curves are from the estimates when $\mathbf{b}$ is given.



in (5), which confirms that incorporating the mixed effects model already eliminates some noise in the repeated measures.

Figure 3.1c illustrates the performance of our mixed effects envelope method. The solid curves at the bottom are obtained by applying our method on the data set in Figure 3.1a. Unlike the standard EM algorithm which only recovers the between-subject variability, our method additionally reduces the within-subject variability. The estimated envelope dimension $\widehat{u} = 1$. The group difference estimated from our method is $(-6.97, 6.99)^T$ with the MSE 0.40. The solid and dashed curves are similar, indicating that our method provides a similar point estimate as the standard envelope estimator with random effects subtracted. While the classic envelope estimator has almost no efficiency gain over the standard EM estimator (similar MSEs), our mixed effects envelope method achieves substantial efficiency gain over the standard EM estimator (the MSE ratio is only about 0.4) and even more efficiency gain over the OLS estimator from model (4) (the MSE ratio is about 0.36). This shows that by leveraging both the mixed effects model and the envelope structure, our method may achieve a much greater amount of efficiency gain as compared with either method.

3.3.3 Maximum likelihood estimation

Under the reparameterization implied by Conditions 1* and 2*, we first investigate the likelihood function given the observed data. Recall that $\mathbf{A}_i = \mathbf{Z}_i^T \otimes \mathbf{I}_r$. We have

$$L(\boldsymbol{\theta}, \mathbf{Y}; \mathbf{X}, \mathbf{Z}) \propto \prod_{i=1}^{n} |\mathbf{I}_{J_i} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} + \mathbf{A}_i \boldsymbol{\Sigma}_{\mathbf{b}} \mathbf{A}_i^T|^{-1/2} \exp\Big\{ -\frac{1}{2}\{\mathrm{vec}(\mathbf{Y}_i) - \boldsymbol{\alpha} \otimes \mathbf{1}_{J_i} - \mathrm{vec}(\boldsymbol{\beta}\mathbf{X}_i)\}^T (\mathbf{I}_{J_i} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} + \mathbf{A}_i \boldsymbol{\Sigma}_{\mathbf{b}} \mathbf{A}_i^T)^{-1} \{\mathrm{vec}(\mathbf{Y}_i) - \boldsymbol{\alpha} \otimes \mathbf{1}_{J_i} - \mathrm{vec}(\boldsymbol{\beta}\mathbf{X}_i)\} \Big\}, \tag{3.5}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Sigma}_{\mathbf{b}})$, $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$ and $\propto$ denotes proportional to. The formula above is obtained by square completion and the Woodbury matrix identity. As the likelihood function $L(\boldsymbol{\theta}, \mathbf{Y}; \mathbf{X}, \mathbf{Z})$ have a complicated form in $\boldsymbol{\theta}$, the MLE of model (3.4) does not have a closed form in general.

In order to obtain the maximum likelihood estimate, we combine the EM-algorithm and the envelope structure. The resulting algorithm is not trivial since the parameters are not element-wise identifiable and random effects are not observable. Due to space constraints, we only briefly describe the steps here and relegate the technical details of the algorithm in the Supplementary Materials. For any predetermined envelope dimension u, we start with an initial value for all parameters. We calculate the E-step and then, during the M-step, we decompose the expectations from the E-step such that all the other parameters can be optimized individually given $\boldsymbol{\Gamma}$, and then we optimize over $\boldsymbol{\Gamma}$. We iterate such EM process till convergence.

We adapt the BIC in the classic envelope models (Eck and Cook, 2017) to estimate u in the mixed effects envelope model. Under model (3.4), BIC is $-2l(\widehat{\boldsymbol{\theta}}; \mathbf{Y} \mid \mathbf{X}, \mathbf{Z}) + \log(J_\cdot)pu$, where $l(\widehat{\boldsymbol{\theta}}; \mathbf{Y} \mid \mathbf{X}, \mathbf{Z})$ is the log of the likelihood $L(\boldsymbol{\theta}, \mathbf{Y}; \mathbf{X}, \mathbf{Z})$ given in (3.5). The penalty coefficient in BIC is $\log(J_\cdot)$ rather than $\log(n)$ to take the longitudinal feature of data into consideration (Jones, 2011). Also, the likelihood in BIC is the observed data likelihood rather than the full data one. We summarize the mixed effects envelope algorithm in the appendix.

3.3.4 Efficiency Gain

We discuss the asymptotic variance of the mixed effects envelope estimator. The parameters of the envelope model is vector $\boldsymbol{\phi} = (\boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Sigma}_{\mathbf{b}})$. A more rigorous notation is $\boldsymbol{\phi} = (\mathrm{vec}(\boldsymbol{\eta}), \mathrm{vec}(\boldsymbol{\Gamma}), \mathrm{vech}(\boldsymbol{\Omega}), \mathrm{vech}(\boldsymbol{\Omega}_0), \mathrm{vech}(\boldsymbol{\Sigma}_{\mathbf{b}}))$. We omit the vectorization notations here. We are interested in the property of the parame-

ter $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ and $\boldsymbol{\Sigma}_{\mathbf{b}}$, which can be viewed as functions of $\boldsymbol{\phi}$. Generally, we have $\mathbf{h}(\boldsymbol{\phi}) = (\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}, \boldsymbol{\Sigma}_{\mathbf{b}}) = (\boldsymbol{\Gamma}\boldsymbol{\eta}, \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \boldsymbol{\Sigma}_{\mathbf{b}}) = (h_1(\boldsymbol{\phi}), h_2(\boldsymbol{\phi}), h_3(\boldsymbol{\phi}))$. Let $\boldsymbol{\theta} = \mathbf{h}(\boldsymbol{\phi})$, $\widehat{\boldsymbol{\theta}}_{mix\cdot env}$ and $\widehat{\boldsymbol{\theta}}_{mix\cdot em}$ denote the mixed effects envelope and standard EM estimates under (3.2). The asymptotic variance of our estimator can be calculated using Shapiro (1986).

Proposition 3.3.2 *Under model* (3.2) *and assume envelope conditions (i)* and (ii)* hold, then* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{mix\cdot em} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$, *and* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{mix\cdot env} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{mix\cdot env})$ *where* $\mathbf{V}_{mix\cdot env} = \mathbf{G}(\mathbf{G}^T\mathbf{V}\mathbf{G})^{\dagger}\mathbf{G}$, *the form of* $\mathbf{V}$ *is given in the appendix, and* $\mathbf{G}$ *is given by*

$$\begin{pmatrix} \mathbf{I}_p \otimes \boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \mathbf{I}_r & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{C}_r(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \mathbf{I}_r - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) & \mathbf{C}_r(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_r(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{r-u} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_{qr(qr+1)/2} \end{pmatrix},$$

Moreover, $\mathbf{V}^{-\frac{1}{2}}(\mathbf{V} - \mathbf{V}_0)\mathbf{V}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{V}^{-\frac{1}{2}}\mathbf{G}(\mathbf{G}^T\mathbf{V}^{-1}\mathbf{G})^{\dagger}\mathbf{G}^T\mathbf{V}^{-\frac{1}{2}} \geq 0$, *so the mixed effects envelope always has no larger asymptotic variance.*

In order to provide some insights on occasions where our estimator can be efficient as compared with the standard method, we compare $\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}})]$ using the mixed envelope model with the standard model under a relatively simple setting. Specifically, we set $r = 2$, $J = 2$, $p = 1$, $\mathbf{Z}_{i,j} = 1$ for all i, j, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_0^2 \end{pmatrix}$, $\boldsymbol{\Sigma}_{\mathbf{b}} = \begin{pmatrix} \sigma_b^2 & 0 \\ 0 & \sigma_b^2 \end{pmatrix}$, $\boldsymbol{\Gamma} = (1, 0)^T$, $\boldsymbol{\Gamma}_0 = (0, 1)^T$ and $\eta = 1$. In this specific case, we have the close form formula

$$\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{mix\cdot em})] = \begin{pmatrix} \dfrac{\sigma_1^2(\sigma_1^2 + 2\sigma_b^2)}{\sigma_b^2\sigma_{x_1}^2 + \sigma_1^2\sigma_{x_2}^2} & 0 \\ 0 & \dfrac{\sigma_0^2(\sigma_0^2 + 2\sigma_b^2)}{\sigma_b^2\sigma_{x_1}^2 + \sigma_0^2\sigma_{x_2}^2} \end{pmatrix},$$

and

$$\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}_{mix\cdot env})] = \begin{pmatrix} \dfrac{\sigma_1^2(\sigma_1^2 + 2\sigma_b^2)}{\sigma_b^2\sigma_{x_1}^2 + \sigma_1^2\sigma_{x_2}^2} & 0 \\ 0 & \sigma_{\beta_2}^2 \end{pmatrix},$$

where $\sigma^2_{\beta_2} = \left[\frac{\sigma_b^2\sigma^2_{x_1} + \sigma_0^2\sigma^2_{x_2}}{\sigma_0^2(\sigma_0^2+2\sigma_b^2)} + \frac{4(\sigma_1^2-\sigma_0^2)^2(\sigma_1^2\sigma_0^2+2\sigma_1^2\sigma_b^2+2\sigma_0^2\sigma_b^2+2\sigma_b^4)}{\sigma_1^2\sigma_0^2(\sigma_1^2+2\sigma_b^2)(\sigma_0^2+2\sigma_b^2)}\right]^{-1}$, $\sigma^2_{x_1} = \sum_{i=1}^n (x_{i1}-x_{i2})^2/n$ and $\sigma^2_{x_2} = \sum_{i=1}^n (x_{i1}^2+x_{i2}^2)/n$. As long as $\sigma_1^2 \neq \sigma_0^2$ and $\sigma_0^2 > 0$,

$$\sigma^2_{\beta_2} < \frac{\sigma_0^2(\sigma_0^2+2\sigma_b^2)}{\sigma_b^2\sigma^2_{x_1}+\sigma_0^2\sigma^2_{x_2}}.$$

The ratio

$$\frac{\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}^{(2)}_{mix\cdot em})]}{\text{avar}[\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}^{(2)}_{mix\cdot env})]} = 1 + \frac{4(\sigma_1^2-\sigma_0^2)^2(\sigma_1^2\sigma_0^2+2\sigma_1^2\sigma_b^2+\sigma_0^2\sigma_b^2+2\sigma_b^4)}{\sigma_1^2(\sigma_1^2+2\sigma_b^2)(\sigma_b^2\sigma^2_{x_1}+\sigma_0^2\sigma^2_{x_2})}$$

tends to $+\infty$ as $\sigma_0^2 \to +\infty$. Therefore, the efficiency gain is large when σ_0^2 is large relative to σ_1^2. Consequently, in this case, fewer samples are needed to detect the same effect size for our method as compared with the standard EM.

The consistency and efficiency gain of the mixed effects envelope estimator in Proposition 3.3.2 is derived based on the normality of the error and random effect. In the next proposition, we justify the $\sqrt{n}$-consistency of $\widehat{\boldsymbol{\theta}}_{mix\cdot env}$ without the normality conditions on the error and random effect.

Proposition 3.3.3 *If the error $\boldsymbol{\varepsilon}_{ij}$ and random effect $\mathbf{b}_i$ have finite $(4+\delta)$-th moments for some $\delta > 0$, and the regularity conditions in the appendix hold, then $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{mix\cdot em} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \widetilde{\mathbf{V}})$, and $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{mix\cdot env} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \widetilde{\mathbf{V}}_{mix\cdot env})$ for some covariance matrices $\widetilde{\mathbf{V}}$ and $\widetilde{\mathbf{V}}_{mix\cdot env}$. In addition, we have $\widetilde{\mathbf{V}}_{mix\cdot env} = \mathbf{G}(\mathbf{G}^T\mathbf{J}\mathbf{G})^\dagger\mathbf{G}^T\mathbf{J}\widetilde{\mathbf{V}}\mathbf{J}\mathbf{G}(\mathbf{G}^T\mathbf{J}\mathbf{G})^\dagger\mathbf{G}^T$. The definition of $\mathbf{J}$ is given in the Appendix.*

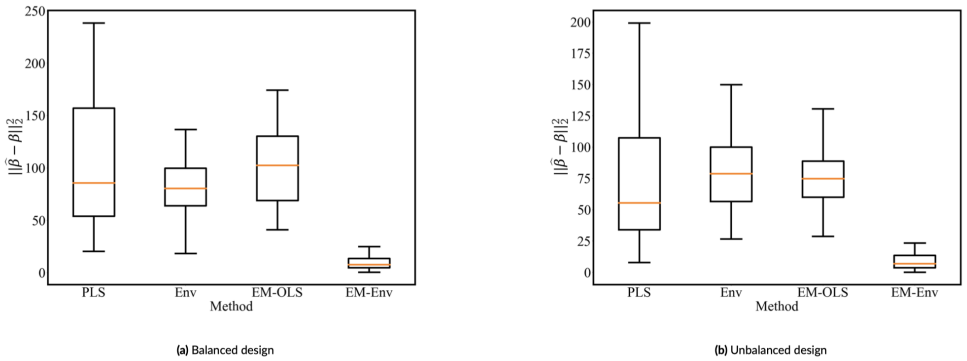
3.4 Simulations

In this section, we carry out simulations to compare the finite sample efficiency of our estimator with the standard EM method, the response envelope method and the response PLS method using the SIMPLS algorithm. The response PLS is a counterpart algorithm of the standard PLS algorithm but to reduce the dimension reduction rather than that for predictors. A detailed algorithm can be found in Cook (2018a). The response envelope and the response PLS methods do not take the time dependency among responses into consideration, but still provide consistent estimators. Although envelopes and PLS are asymptotically equivalent as Cook et al. (2013) suggested, their finite sample properties are different. We first consider a balanced case by generating a population of $n = 50$ individuals, and each has

$r = 10$ responses measured at each of the 5 time points ($J_i = J = 5,\ i = 1, \ldots, 50$). We set $p = 6$ and $q = 2$ for the fixed and random effects.

We generate parameters $\mathbf{\Gamma}$ and $\boldsymbol{\beta}_0$ of size $r \times u$ and $r \times p$, where $u = 1$. The elements of $\mathbf{\Gamma}$ and $\boldsymbol{\beta}_0$ are from $U(0, 1)$ and $U(-10, 10)$. Let $\boldsymbol{\beta} = P_{\mathbf{\Gamma}}\boldsymbol{\beta}_0$. We generate a matrix $\mathbf{B}$ of dimension $qr \times qr$ with each element from $U(-10, 10)$ and let $\mathbf{\Sigma_b} = \mathbf{B}\mathbf{B}^T$. Let $\mathbf{\Omega} = 0.01\mathbf{I}_u$ and $\mathbf{\Omega}_0 = 100\mathbf{I}_{(r-u)}$ and let $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0{\mathbf{\Gamma}_0}^T$. For each individual, $\mathbf{X}_{i1}, \mathbf{X}_{i2}, \mathbf{X}_{i3}$ vary with time, and $\mathbf{X}_{i4}, \mathbf{X}_{i5}, \mathbf{X}_{i6}$ stay fixed for all the time points, where each predictor (time varying or fixed over time) is independently generated from $U(-10, 10)$. Then, generate $\mathbf{Z}_i = (\mathbf{Z}_{i1}^T, \mathbf{Z}_{i2}^T)^T$, where $\mathbf{Z}_{i1} = \mathbf{1}_{1\times J}$, and each element of $\mathbf{Z}_{i2}$ follows $U(-10, 10)$. Generate vector $\boldsymbol{\varepsilon}_{ij} \in \mathbb{R}^r$, where each column follows $N(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}})$. Also generate $\mathbf{b}_i \in \mathbb{R}^{r\times q}$, where $\text{vec}(\mathbf{b}_i)$ is from the normal distribution $N(\mathbf{0}, \mathbf{\Sigma_b})$. Set $\mathbf{Y}_{ij} = \boldsymbol{\beta}\mathbf{X}_{ij} + \mathbf{b}_i\mathbf{Z}_{ij} + \boldsymbol{\varepsilon}_{ij}$. We then calculate $\widehat{\boldsymbol{\beta}}_{mix\cdot em}$, $\widehat{\boldsymbol{\beta}}_{mix\cdot env}$ $\widehat{\boldsymbol{\beta}}_{env}$ and $\widehat{\boldsymbol{\beta}}_{pls}$, and repeat the above procedure for 100 times.

Figure 3.2: Empirical distribution of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$



(a) Balanced design

(b) Unbalanced design

We compute the square of l_2 norm $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$ for each simulation. The boxplot of l_2^2 error across 100 simulations is given in Figure 3.2a, where we suppress the outliers to make the figure clean. The mixed envelope estimates are significantly better in terms of both bias and variance than the standard EM estimates. For example, the mean l_2^2 error of the mixed envelope estimate for $\boldsymbol{\beta}$ is 12.12, while that is 104.77, 79.40 and 105.99 for the standard EM, response envelope and response PLS estimates respectively. Also, 99 out of 100 of our method selected the correct envelope dimension $u = 1$.

Then, we examine the results where J_i is uniformly generated from $\{5, 6, 7, 8, 9\}$. Other steps in the previous simulation remain unchanged. The mean l_2^2 error of $\widehat{\boldsymbol{\beta}}_{mix\cdot env}$ is 9.30, while that is 69.11, 79.05 and 72.16 for $\widehat{\boldsymbol{\beta}}_{mix\cdot em}$, $\widehat{\boldsymbol{\beta}}_{env}$ and $\widehat{\boldsymbol{\beta}}_{pls}$.

Also, the envelope dimension is always correctly estimated as $\widehat{u} = 1$. The empirical distribution of l_2^2 error is shown in Figure 3.2b. Our proposed mixed effects envelope estimator has much smaller MSE than the standard EM estimator even in relatively small samples.

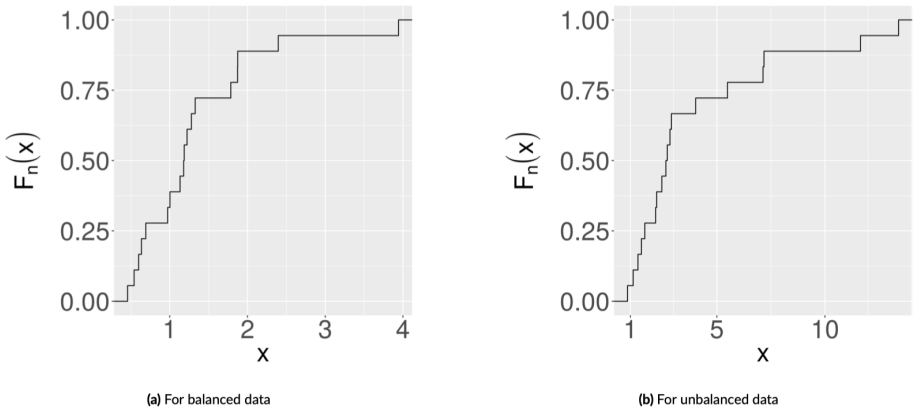
3.5 Data Analysis

In this section, we apply our proposed method to the Action to Control Cardiovascular Risk in Diabetes (ACCORD) study. The ACCORD randomized-control trial aimed at determining whether cardiovascular disease (CVD) event rates can be reduced in people with diabetes. Participants are between the ages of 40 and 82. All participants have Type 2 diabetes and an especially high risk for heart attack and stroke.

We are interested in the treatment effect on the quality of life and changes in health outcomes. The responses were collected at four time points, which are 12, 24, 36, 48 months after the beginning of the trial. We consider 2054 participants who responded to the survey, among whom 1156 individuals responded to all surveys. In our analysis, the response variables were treatment satisfaction, depression scale, aggregate physical activity score, aggregate mental score, aggregate interference score, symptom and distress score, systolic blood pressure (SBP), diastolic blood pressure (DBP), and heart rate. The predictors were age and treatment, where we consider the intensive glycemia treatment ($T = 1$) and the standard glycemia treatment ($T = 0$).

We first assessed the difference in the quality of life versus glycemia level and age for people who attended all four surveys ($J_i = 4$). All responses except systolic blood pressure and diastolic blood pressure had a missing rate less than 1.5%. Since the missing rate is low, we imputed the missing data using its mean value. We assume there is only random intercept in our model. The mixed effects envelope method reduced the dimension of the response variable from $r = 9$ to $\widehat{u} = 4$. The point estimates, bootstrap standard errors, and p-values for the regression parameter is given in Table B.1 in the Supplementary Material. The magnitude of the point estimates of our method is in general slightly smaller than those of the standard EM. For example, the coefficient for treatment satisfaction with respect to treatment is 0.46 using our method and 0.69 using the standard EM. As the envelope estimate is obtained by projecting the standard estimates onto the envelope direction, the reduction in the magnitude can be interpreted as the noise subtracted from the

Figure 3.3: The empirical cumulative distribution of the ratio between the estimated standard errors of the standard EM and that of our method for ACCORD data.



original estimates. As mentioned in Section 3.3.3, the closed form of the standard errors of our method are difficult to obtain. Therefore, we used the nonparametric bootstrap. Figure 3.3a shows the empirical cumulative density distributions of the estimated standard errors of the standard EM versus that of our method. The estimated standard errors are in general smaller (on the right hand side of 1 in Figure 3.3a) using our method than using the standard EM, which indicates the efficiency gain of our method. The mean ratio of the coefficient standard error using our method over the standard EM is 1.33. That is, to achieve the same mean power among all predictors, our methods require 75.2% of the original sample.

Then, we repeated the analysis by including people with less than four surveys, thus J_i varies across each person. In this case, the total number of observations $J_{\cdot}$ increases from 1156 to 2054. The missing rate is about the same, so we imputed them using their mean again. The estimated envelope dimension is $\widehat{u} = 1$. The point estimate, bootstrap standard errors and p-values for the regression coefficients are given in Table B.2 in the Supplementary Material. It is worth noticing that our method found SBP and DBP corresponding to age significant, whereas the standard method found physical score corresponding to age significant. Figure 3.3b shows the empirical distribution of the ratio between the estimated standard errors of the two methods. The mean ratio of the coefficient standard error using our method over the standard EM is 4.07. That is, to achieve the same mean power among all predictors, our methods require 24.6% of the original sample. All the regression coefficients except symptom and distress score corresponding to age have a smaller standard error, which again shows that our method is more efficient.

3.6 Discussion

In this chapter, we proposed the mixed effects envelope method to achieve a more efficient estimation than the traditional EM in longitudinal studies. Although this chapter is motivated by the repeated measures problem, the mixed effects envelope model can also be used in clustered data to achieve efficiency gain. For example, patients are nested in physicians, who are in turn nested in clinics. Such clustered data also features in correlations between observations.

The mixed effects model is closely related to the missing data problem since the random effects can be viewed as missing for all observations. Moreover, the missing data techniques may be combined with the mixed effects model to further relax conditions. Ma et al. (2021) discussed the envelope method under the ignorable missingness of predictors and covariates. In this chapter, we assume that the measures collected at each visit are balanced and repeated measures may be collected at different time points across individuals. Such a condition may be violated when a different number of measures are collected every time. One possible solution is to use the union of responses as the balanced response and frame this into a missing data problem. The extension of the mixed effects models with missing data is left as a future research avenue.

In this chapter, we considered a heteroscedastic error induced by the mixed effects model. Su and Cook (2013) proposed an alternative method for another heteroscedastic error covariance structure under the multivariate regression. Both covariance structures allow us to formulate the regression with heteroscedastic variance as a variation of the original envelope model and thus we can use the original computation to obtain the MLE of the likelihood. How to generalize the model with a general heteroscedastic variance structure is left for future research.

In many contemporary studies and applications, the dimension of data can be much larger than the number of observations. Under such high-dimensional settings, more refined variants of the envelope method are desired. Many studies have adapted the envelope method to high-dimensional settings under sparsity conditions. Su et al. (2016) proposed the envelope model for response variable selection under high-dimensional settings. Zhu and Su (2020) incorporated the envelope model with partial least squares for high-dimensional regression. We leave the extension of our mixed effects model for high-dimensional data as a future research topic.

Additional information for this chapter, including the proof of propositions, graphical illustration of the response envelope, EM-algorithm for the mixed effects

envelope model, derivation of $\boldsymbol{\mu}_{\mathbf{b}_i,t}$ and $\boldsymbol{\Sigma}_{\mathbf{b}_i,t}$, the mixed effects envelope algorithm, and Tables B.1 and B.2 for the data analysis are available in the appendix.

Chapter 4

Semiparametric Efficient Inner Envelope

4.1 Introduction

4.1.1 Background

In 2010, Cook et al. (2010) proposed a new approach to dimension reduction in multivariate linear regression models called the envelope approach, with the aim of more efficiently estimating the underlying regression coefficients. Briefly, for each individual $i = 1, \ldots, n$, consider a multivariate linear regression model where we regress r-dimensional, multivariate responses $\mathbf{Y}_i \in \mathbb{R}^{r\times 1}$ onto p regressors $\mathbf{X}_i \in \mathbb{R}^{p\times 1}$, i.e.

$$\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}). \tag{4.1}$$

Here, the term $\boldsymbol{\varepsilon}_i \in \mathbb{R}^{r\times 1}$ is a random vector that is independent from $\mathbf{X}_i$ and follows a multivariate normal distribution with mean $\mathbf{0}$ and unknown covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$. The parameter $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ is the unknown matrix regression coefficients of interest. A natural estimator of $\boldsymbol{\beta}$ is the ordinary least square (OLS) estimator, which takes the form $\hat{\boldsymbol{\beta}}^T_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ where $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^T \in \mathbb{R}^{n\times r}$. However, when the number of responses r is large, Cook and co-authors proposed to improve on the OLS estimator by considering an "envelope" subspace in the space of the responses $\mathbb{R}^r$. Specifically, if there is a linear combination of the responses that is an ancillary for $\boldsymbol{\beta}$, Cook et al. (2010) proposed a new estimator of $\boldsymbol{\beta}$, called the envelope estimator, and showed that the new envelope estimator is at least as efficient as the OLS estimator; see Section 4.2.2

for details.

Since the introduction of the envelope estimator, there has been an explosion of work on using envelope-based dimension reduction methods for multivariate regression models (Cook et al., 2010; Su and Cook, 2011, 2012; Cook and Su, 2013; Cook et al., 2013, 2015; Cook and Zhang, 2015b,a; Li and Zhang, 2017; Shi et al., 2020; Cook et al., 2021; Rekabdarkolaee et al., 2020; Ma et al., 2021). Our paper focuses on one popular extension and improvement of the original work, the inner envelope estimator of Su and Cook (2012). Specifically, the inner envelope estimator supposes that there is an extra subspace that is dependent on some responses, but becomes independent after conditioning on the regressors; see Section 4.2.2 for the exact definitions. Critically, the inner envelope estimator of $\boldsymbol{\beta}$ can achieve efficiency gains even when the envelope estimator offers no gains.

Despite its superior performance, the inner envelope estimator relies on parametric assumptions (e.g. model (4.1)) to guarantee its efficiency gains over the envelope estimator or the OLS estimator. In general, most envelope-based approaches rely on a linear relationship between the responses and the regressors (Cook et al., 2010; Su and Cook, 2012; Cook and Su, 2013; Li and Zhang, 2017; Shi et al., 2020; Cook et al., 2021) and/or the joint normality of the error terms (Cook et al., 2010; Cook and Su, 2013; Li and Zhang, 2017; Cook et al., 2021), even though the envelope-based assumptions are not stated with respect to a particular parametric model; see, for example, our high-level descriptions of the envelope or the inner envelope above or the exact definitions in Section 4.2.2. Often, parametric assumptions are imposed out of theoretical convenience, especially to leverage the useful properties of multivariate normal distributions concerning independence from its covariance matrix or the variational independence between the parameters for the mean and the covariance matrix.

More importantly, in practice, these parametric assumptions are often violated or infeasiable for certain data types. For example, if the responses are binary, the existing inner envelope estimator cannot be applied because, unlike the multivariate normal model in (4.1), $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ are generally not variationally independent for binary responses; so far, there is no inner envelope estimator for logistic or multinomial multivariate regression. A similar problem occurs when some responses are continuous while others are binary. Or, if the responses form non-linear relationships with the regressors, say the regressors have higher-order polynomial terms, or the regression errors are heteroskedastic, the existing inner envelope estimator may be inconsistent and/or inefficient. In Section 4.2.3, we show the inconsistency of the

existing inner envelope estimator when model (4.1) is violated.

While some progress has been made to relax these assumptions in envelope-based methods (Cook and Zhang, 2015a; Zhang et al., 2018), to the best of our knowledge, there is no work on relaxing the parametric assumptions underlying the inner envelope estimator. This paper aims to resolve this through a semiparametric generalization of the inner envelope.

4.1.2 Our Contributions

At a high level, our semiparametric generalization relies on efficiently estimating a more fundamental quantity discussed in the original inner envelope paper called the inner envelope space (Su and Cook, 2012). In particular, in the original work, estimation of $\boldsymbol{\beta}$ was done in two steps where the first step involved estimating the inner envelope space and the second step involved a projection of the responses and the regressors onto the estimated inner envelope space; see Section 4.2.2 for details. Our main insight in the paper is to realize that (a) the inner envelope space in the original work can actually be defined without assuming a parametric model (4.1) or even a particular type of responses (i.e. continuous, binary, or even a mixture of them), and (b) the inner envelope space can be uniquely parametrized as finite dimensional, semi-orthogonal basis matrices. These insights serve as the basis to develop locally and globally semiparametric efficient estimators, including a locally robust estimator that remains consistent and asymptotically normal even if some of the underlying models are mis-specified.

In addition to being robust to parametric modelling assumptions, our semiparametric generalization has two additional byproducts. First, we show a computationally simple procedure to evaluate all regular and asymptotically linear (RAL) estimators based on the generalized method of moments (GMMs). Our procedure is dramatically simpler than the original approach in Su and Cook (2012) based on solving a non-convex objective function over a Grassmannian, a non-convex set. Second, since our semiparametric generalization does not rely on a parametric model between the responses and the regressors, we show how to use the efficiently estimated inner envelope space to improve predictions from supervised machine learning models, such as XGBoost (Chen et al., 2015) and random forest (Breiman, 2001). In particular, we show that directly using the original responses $\mathbf{Y}_i$ inside a supervised machine learning method may lead to poor, out-of-sample, predictive performance compared to using the dimension-reduced responses generated from our approach.

Our work is related to some recent works on semiparametric relaxation of sufficient dimension reduction (SDR) in traditional, univariate (i.e. single response) linear regression models (Ma and Zhu, 2012, 2013). Typically, SDR methods such as sliced inverse regression (SIR) (Li, 1991), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991b) and directional regression (DR) (Li and Wang, 2007) require the univariate response to be linear and have homoskedastic variance. Ma and Zhu (2012) relaxed these parametric assumptions by directly estimating the dimension-reduced subspace of the regressors. In contrast, our work deals with the case when both the regressors and the responses are multivariate. In particular, our semiparametric relaxation of the inner envelope requires estimating two subspaces, one of which summarizes the relationship between $\mathbf{X}$ and $\mathbf{Y}$ and another nested subspace that contains purely relevant information; see Section 4.3 for the exact definitions. The former subspace, which we denote as $\mathcal{S}_1 \cup \mathcal{S}_2$ below, can be thought of as the generalization of the dimension-reduced subspace of regressors estimated in Ma and Zhu (2012) when there are multivariate responses. But, the latter subspace, denote as $\mathcal{S}_3$ below, is new and arises because of the multivariate responses with a dimension-reducing envelope structure.

The outline of the paper is as follows. Section 4.2 reviews the original envelope (Cook et al., 2010) and the inner envelope (Su and Cook, 2012). The section also illustrates the problem of the existing inner envelope when the parametric modeling assumption is violated. Section 4.3 formalizes our semiparametric generalization of the inner envelope. Sections 4.4 derive the semiparametric efficiency bound of the inner envelope space and show estimators that can achieve this bound. Section 4.5 provides implementation details for our estimators. Sections 4.6 and 4.7 numerically demonstrate the efficiency gains of the proposed estimators in simulation and a study on the relationship between glycemic control and various risk factors for cardiovascular disease. Section 9 contains the proofs of the key results in the paper.

4.2 Preliminaries

4.2.1 Notation

For any matrix $\mathbf{A} \in \mathbb{R}^{r \times p}$, let $\mathcal{A} = \text{span}(\mathbf{A})$ denote the subspace of $\mathbb{R}^r$ spanned by the columns of $\mathbf{A}$ and let $\mathcal{A}^\perp$ denote the orthogonal complement of this subspace. Let $\mathbf{P}_\mathcal{A}$ and $\mathbf{Q}_\mathcal{A} = \mathbf{I} - \mathbf{P}_\mathcal{A}$ denote the orthogonal projection matrix that projects a vector $\mathbf{v} \in \mathbb{R}^r$ onto $\mathcal{A}$ and $\mathcal{A}^\perp$, respectively. Let $\dim(\mathcal{A})$ denote the dimension of

the subspace $\mathcal{A}$. Let $\text{vec}(\mathbf{A})$ denote the vector formed by stacking columns of the matrix $\mathbf{A}$ and $\|\mathbf{A}\|_F = \sqrt{\text{trace}(\mathbf{A}^T\mathbf{A})}$ denote its Frobenius norm. Finally, we say a square matrix $\mathbf{A} \in \mathbb{R}^{r\times r}$ is an orthogonal matrix if $\mathbf{A}^T\mathbf{A} = \mathbf{I}_r$. We say a non-square matrix $\mathbf{A} \in \mathbb{R}^{r\times p}$ is a semi-orthogonal matrix if $\mathbf{A}^T\mathbf{A} = \mathbf{I}_p$.

For two matrices $\mathbf{A} \in \mathbb{R}^{r\times p}$ and $\mathbf{B} \in \mathbb{R}^{s\times q}$, let $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{rs\times pq}$ denote their Kronecker product. Relatedly, for any vector $\mathbf{v} \in \mathbb{R}^r$, let $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$. For any two subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ of $\mathbb{R}^r$, let $\mathcal{S}_1 \oplus \mathcal{S}_2 = \{\mathbf{x}_1 + \mathbf{x}_2 : \mathbf{x}_1 \in \mathcal{S}_1, \mathbf{x}_2 \in \mathcal{S}_2\}$ denote their direct sum. For a function $\mathbf{f}(\mathbf{x}, \mathbf{y}) : \mathbb{R}^{p_1+p_2} \to \mathbb{R}^q$, let $\partial\mathbf{f}/\partial\mathbf{x}^T \in \mathbb{R}^{p_1\times q}$ denote the derivative of $\mathbf{f} = (f_1, \ldots, f_q)$ with respect of $\mathbf{x}$ where the (i, j)-th entry is denoted as $\partial f_i/\partial x_j$.

4.2.2 Review: Parametric Envelope and Inner Envelope Estimators

We review the parametric envelope estimator of Cook et al. (2010) and the parametric inner envelope estimator of Su and Cook (2012). The review is not aimed to be comprehensive; rather, the review is geared towards introducing some important concepts, notably the subspaces $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$, that we will use in our semiparametric generalization. Also, without loss of generality, we will assume both $\mathbf{X}$ and $\mathbf{Y}$ are mean $\mathbf{0}$ random vectors.

For the parametric envelope estimator, consider a subspace $\mathcal{S} \subseteq \mathbb{R}^r$ that satisfies the following conditions in model (4.1):

Condition 4.2.1 $\mathbf{Q}_\mathcal{S}\mathbf{Y}_i \mid \mathbf{X}_i \sim \mathbf{Q}_\mathcal{S}\mathbf{Y}_i$,

Condition 4.2.2 $\mathbf{Q}_\mathcal{S}\mathbf{Y}_i \perp\!\!\!\perp \mathbf{P}_\mathcal{S}\mathbf{Y}_i \mid \mathbf{X}_i$.

The smallest subspace $\mathcal{S}$ that satisfies the above conditions is called the envelope subspace. Broadly speaking, Conditions 4.2.1 and 4.2.2 state that $\mathbf{Q}_\mathcal{S}\mathbf{Y}_i$ is irrelevant in estimating $\boldsymbol{\beta}$. Consequently, we can regress $\mathbf{P}_\mathcal{S}\mathbf{Y}_i$ on $\mathbf{X}_i$ to obtain a more efficient, envelope estimator of $\boldsymbol{\beta}$, especially compared to the naive OLS estimator that regresses the entire responses $\mathbf{Y}_i$ on $\mathbf{X}_i$; see Cook et al. (2010) for the exact arguments.

However, if the subspace satisfying Conditions 4.2.1 and 4.2.2 happens to be the entire space of the multivariate responses (i.e. $\mathbb{R}^r$), the envelope estimator has no efficiency gains over the OLS estimator. In such settings, Su and Cook (2012) proposed the inner envelope estimator of $\boldsymbol{\beta}$, which is defined as follows. Suppose

the responses can be decomposed as the sum of projections onto three orthogonal subspaces $\mathcal{S}_1,\ \mathcal{S}_2,\ \mathcal{S}_3 \subseteq \mathbb{R}^r, \mathcal{S}_1 \oplus \mathcal{S}_2 \oplus \mathcal{S}_3 = \mathbb{R}^r$, with dimensions $\dim(\mathcal{S}_1) = u$, $\dim(\mathcal{S}_2) = p - u$, and $\dim(\mathcal{S}_3) = r - p$. The three subspaces $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ satisfy the following conditions:

Condition 4.2.3 $\mathbf{P}_{\mathcal{S}_3}\mathbf{Y}_i \mid \mathbf{X}_i \sim \mathbf{P}_{\mathcal{S}_3}\mathbf{Y}_i$,

Condition 4.2.4 $\mathbf{P}_{\mathcal{S}_1}\mathbf{Y}_i \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}_1}\mathbf{Y}_i \mid \mathbf{X}_i$.

The largest subspace $\mathcal{S}_1$ satisfying Conditions 4.2.3 and 4.2.4 is called the inner envelope subspace. Unlike Conditions 4.2.1 and 4.2.2, the inner envelope estimator presents a new subspace $\mathcal{S}_2$ where $\mathbf{P}_{\mathcal{S}_2}\mathbf{Y}_i$ is correlated with both $\mathbf{P}_{\mathcal{S}_1}\mathbf{Y}_i$ and $\mathbf{P}_{\mathcal{S}_3}\mathbf{Y}_i$, but $\mathbf{P}_{\mathcal{S}_1}\mathbf{Y}_i$ is independent with $\mathbf{P}_{\mathcal{S}_2}\mathbf{Y}_i$ and $\mathbf{P}_{\mathcal{S}_3}\mathbf{Y}_i$ given $\mathbf{X}$. If $\mathcal{S}_2$ is the null space, Condition 4.2.4 reduces to Condition 4.2.2 and the inner envelope subspace coincides with the envelope subspace, i.e. $\mathcal{S}_1 = \mathcal{S}$. Critically, Conditions 4.2.3 and 4.2.4 do not necessarily rely on a parametric model between $\mathbf{Y}_i$ and $\mathbf{X}_i$, a fact that we exploit in our semiparametric generalization of the inner envelope below.

Under Conditions 4.2.3, 4.2.4, and model (4.1), Su and Cook (2012) proposed a two-step, inner envelope estimator of $\boldsymbol{\beta}$ where the first step involves estimating the aforementioned subspaces $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ and the second step involves estimating the relationship between the predictors and the projected response. Specifically, in the first step, let $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ and $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{r\times(r-u)}$ denote the semi-orthogonal basis matrices for subspaces $\mathcal{S}_1$ and $\mathcal{S}_1^{\perp}$, respectively (i.e. $\text{span}(\boldsymbol{\Gamma}) = \mathcal{S}_1$ and $\text{span}(\boldsymbol{\Gamma}_0) = \mathcal{S}_2 \oplus \mathcal{S}_3$). Also, let $\mathbf{B} \in \mathbb{R}^{(r-u)\times(p-u)}$ be a semi-orthogonal matrix such that $\boldsymbol{\Gamma}_0\mathbf{B}$ is a semi-orthogonal basis matrix for $\mathcal{S}_2$ (i.e. $\text{span}(\boldsymbol{\Gamma}_\mathbf{0}\mathbf{B}) = \mathcal{S}_2$). Intuitively, the matrix $\mathbf{B}$ divides the subspace spanned by $\boldsymbol{\Gamma}_0$ (i.e. $\mathcal{S}_1^{\perp}$) into the subspace specific to $\mathcal{S}_2$ and if $\mathbf{B}_0$ is the orthogonal complement of $\mathbf{B}$, $\mathcal{S}_3$ can be rewritten as $\mathcal{S}_3 = \text{span}(\boldsymbol{\Gamma}_0\mathbf{B}_0)$. Based on these orthogonal relationships, estimating $\boldsymbol{\Gamma}$ and $\mathbf{B}$ is sufficient to estimate all the subspaces $\mathcal{S}_1, \mathcal{S}_2$, and $\mathcal{S}_3$. To estimate $\boldsymbol{\Gamma}$, Su and Cook (2012) proposed the following estimator:

$$\hat{\boldsymbol{\Gamma}} = \underset{\mathbf{G}}{\text{argmin}}\{\log|\mathbf{G}^T\mathbf{S}_{\text{res}}\mathbf{G}| + \log|\mathbf{G}^T\mathbf{S}_{\text{res}}^{-1}\mathbf{G}| + \sum_{i=p-u+1}^{r-u}\log(1+\lambda_{(i)})\} \tag{4.2}$$

such that $\mathbf{G}^T\mathbf{G} = \mathbf{I}_u$.

Here, $\mathbf{G}_0$ is the orthogonal complement of $\mathbf{G}$ and $\lambda_{(i)}$, $i = 1, \ldots, n$ are the ordered descending eigenvalues of $(\mathbf{G}_0^T\mathbf{S}_{\text{res}}\mathbf{G}_0)^{-1/2}(\mathbf{G}_0^T\mathbf{S}_{\text{fit}}\mathbf{G}_0)(\mathbf{G}_0^T\mathbf{S}_{\text{res}}\mathbf{G}_0)^{-1/2}$ with $\mathbf{S}_{\text{fit}}$ and

$\mathbf{S}_{\text{res}}$ being the sample covariances of the fitted and residual vectors, respectively, from the OLS regression of $\mathbf{Y}$ on $\mathbf{X}$. Using the orthogonal relationship, we can also arrive at the estimator $\hat{\boldsymbol{\Gamma}}_0$, which is the orthogonal complement of the estimated $\hat{\boldsymbol{\Gamma}}$. The estimator of $\mathbf{B}$, which we denote as $\hat{\mathbf{B}}$, is similar and for brevity, the details are relegated to the Appendix.

Once the basis matrices are estimated, Su and Cook (2012) proposed to estimate $\boldsymbol{\beta}$ as follows:

$$\hat{\boldsymbol{\beta}}_{\text{IE}} = \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\zeta}}_1 + \hat{\boldsymbol{\Gamma}}_0\hat{\mathbf{B}}\hat{\boldsymbol{\zeta}}_2.$$

Here, $\hat{\boldsymbol{\zeta}}_1, \hat{\boldsymbol{\zeta}}_2$ are the estimated coordinates of the projections onto the estimated subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$. Under the linear model with normal errors in (4.1), Su and Cook (2012) showed that the estimator $\hat{\boldsymbol{\beta}}_{\text{IE}}$ is not only consistent and asymptotically normal, but also has equal or smaller asymptotic variance than $\hat{\boldsymbol{\beta}}_{\text{OLS}}$.

Finally, we make two additional remarks. First, the original inner envelope implicitly assumed the dimension of the subspace $\mathcal{S}_3$ is fixed at $r - p$ and thus, $\boldsymbol{\beta}$ is required to have full column rank. In our semiparametric generalization below, we relax this assumption and allow $\mathcal{S}_3$ to have a dimension other than $r - p$. This relaxation introduces one minor condition on the maximum size of the dimensions to ensure identifability and is discussed in Section 4.3.1. Second, similar to other works on semiparametric generalizations of dimension-reduction approaches Ma and Zhu (2012, 2013, 2014), we assume u and d are known for our theoretical discussions. But, Section 4.5.2 discusses how to choose this dimension from data.

4.2.3 Problem: Inconsistency of the Parametric Inner Envelope When Model (4.1) is Wrong

In this section, we explore the consequences of using the original inner envelope approach based on model (4.1) when that model no longer holds. This exercise intends to mimic a practitioner who may naively conduct dimension reduction using the original inner envelope, but the underlying data generating model may deviate from model (4.1). Our goal in this section is to not only show that this practice can lead to misleading results, but also to provide a concrete rationale for our semiparametric dimension reduction approach we propose in Section 4.3.

To start, consider the case where model (4.1) is correct. Specifically, let there be two regressors generated from independent standard normals and three response variables. The responses are generated from the model $\mathbf{Y} = (X_1 + \varepsilon_1, X_2 + \varepsilon_2, 2X_2 +$

$\varepsilon_3)^T$ where $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are independent, standard normals. Some simple algebra reveals that the subspaces $\mathcal{S}_1, \mathcal{S}_2$ and $\mathcal{S}_3$ satisfying the inner envelope Conditions 4.2.3 and 4.2.4 are spanned by $(1,0,0)^T$, $(0,1,2)^T$ and $(0,2,-1)^T$, respectively; note that the envelope subspace $\mathcal{S}$ is the entire space $\mathbb{R}^3$ and thus, this is an example where the envelope offers no gain compared to the inner envelope. Also, if $\mathbf{S}_{\text{res}}$ and $\mathbf{S}_{\text{fit}}$ in (4.2) are replaced with the true population covariances, the resulting estimator $\hat{\mathbf{\Gamma}}$ is a consistent estimator of the basis matrix that spans $\mathcal{S}_1$.

Now, suppose model (4.1) is incorrect and the true model has non-linear regressors, say $\mathbf{Y} = (X_1^2 + \varepsilon_1, X_2 + \varepsilon_2, 2X_2 + \varepsilon_3)^T$. The true subspaces $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ that satisfy Conditions 4.2.3 and 4.2.4 are the same as those from the previous paragraph. However, the estimated basis matrix from the population version of (4.2) is not $(1,0,0)^T$; that is, there is another estimate $\hat{\mathbf{\Gamma}}$ that is not equal to $(1,0,0)^T$ and that achieves the global minimum in (4.2). Thus, $\hat{\mathbf{\Gamma}}$ is an inconsistent estimator of the true inner envelope subspace. Similarly, if the true model for $\mathbf{Y}$ is $\mathbf{Y} = (X_1^2+\varepsilon_1, X_2+\varepsilon_2, X_2\varepsilon_3)^T$ where the third response has heteroskedastic variance, the subspace $\mathcal{S}_1$ is the same as the previous examples, but its basis matrix is still not the global optimum of (4.2).

The main, high-level, reason for the inconsistency lies in the original estimator's strong reliance on parametric modeling assumptions. Specifically, equation (4.2) only utilizes second order moments between $\mathbf{X}$ and $\mathbf{Y}$ because under the linear model with normal errors in (4.1), second order moments are "sufficient" to identify the underlying subspaces. However, when Y_1 contains non-linear terms or when Y_3 has heteroskedastic variance, the responses are no longer normal and thus, higher-order moments are necessary to identify the underlying dimension-reduced subspaces that satisfy the inner envelope conditions.

Our proposed semiparametric generalization below aims to remove these dependencies on a particular parametric model and arrive at a more robust and efficient approach to inner envelope.

4.3 A Semiparametric Approach to Inner Envelope

4.3.1 Target parameters and assumptions

At a high level, our semiparametric approach focuses on a "model-free" estimation the fundamental quantities underlying the original inner envelope estimator, the three orthogonal spaces introduced in the prior section, $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$ with dimensions u, d and $r-u-d$. However, a general challenge in estimating subspaces is that the subspace requires an identifiable parametrization. For example, a basis matrix $\mathbf{A} \in \mathbb{R}^{k\times\tau}$ for a τ-dimensional subspace $\mathcal{A}$ in $\mathbb{R}^k$ is not unique even if the subspace $\mathcal{A}$ is unique since for any full rank matrix $\mathbf{U} \in \mathbb{R}^{\tau\times\tau}$, $\mathbf{AU}$ also spans the same subspace $\mathcal{A}$. Additionally, the elements of a basis matrix are often variationally dependent, which can complicate the estimation procedure.

Our solution to remedy these issues are as follows. First, we propose a new, representation of the inner envelope space with the following lemma; see Ma and Zhu (2013) for a related result.

Lemma 4.3.1 *Consider a τ-dimensional subspace $\mathcal{A}$ in $\mathbb{R}^k$ where $\tau \in \{1,\ldots,k\}$. Then, there exists an one-to-one mapping ψ_1 such that $\mathbf{a} = \psi_1(\mathcal{A}) \in \mathbb{R}^{\tau(k-\tau)}$ is a vector with all variational independent elements. Also, there exists an one-to-one mapping ψ_2 such that $\mathbf{A} = \psi_2(\mathcal{A}) \in \mathbb{R}^{k\times\tau}$ is a semi-orthogonal matrix. The construction of ψ_1 and ψ_2 are given in Section 4.9.*

With Lemma 4.3.1, we can uniquely parametrize the inner envelope spaces with semi-orthogonal matrices $\mathbf{\Gamma} = \psi_2(\mathcal{S}_1)$ and $\mathbf{\Gamma}_0\mathbf{B} = \psi_2(\mathcal{S}_2)$; note that since $\mathcal{S}_3$ is orthogonal to $\mathcal{S}_1$ and $\mathcal{S}_2$, we can uniquely represent $\mathcal{S}_3$ as $\mathbf{\Gamma}_0\mathbf{B}_0$. Relatedly, given the orthogonality conditions, we can focus on estimating semi-orthogonal matrices $\mathbf{\Gamma}$ and $\mathbf{B}$.

Second, let $\mathbf{\Gamma}$ and $\mathbf{B}$ be the unique parametrizations of the subspaces from Lemma 4.3.1. Using these parameters, we can restate Conditions 4.2.3 and 4.2.4 as

Condition 4.3.1 $\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}_i \perp\!\!\!\perp \mathbf{X}_i$,

Condition 4.3.2 $\mathbf{\Gamma}^T\mathbf{Y}_i \perp\!\!\!\perp \mathbf{\Gamma}_0^T\mathbf{Y}_i \mid \mathbf{X}_i$

We additionally require $\mathcal{S}_3$ to be the largest subspace that satisfies Condition 4.3.1 and $\mathcal{S}_1$ to be the largest space that satisfies Condition 4.3.2 given $\mathcal{S}_3$. As mentioned in Section 4.2.2, this additional requirement generalizes the implicitly assumed constraint in the original inner envelope where the subspace $\mathcal{S}_3$ has a pre-determined, fixed dimension of $r-p$.

Third, the elements in $\mathbf{\Gamma}$ and $\mathbf{B}$ are not variationally independent, which makes them difficult to work with for theory and for some well-known optimization algorithms. Therefore, we collapse the matrices $\mathbf{\Gamma}$ and $\mathbf{B}$ into variationally independent vectors by using Lemma 4.3.1 again.

Specifically, we can reparametrized the two matrices as $\boldsymbol{\gamma} = \psi_1(\psi_2^{-1}(\mathbf{\Gamma}))$, $\mathbf{b} = \psi_1(\psi_2^{-1}(\mathbf{B}))$ and $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \mathbf{b}^T)^T \in \mathbb{R}^{q\times 1}$ with $q = (r-u)\times u + (r-u-d)\times d$. Also, $\boldsymbol{\theta}$ uniquely parametrizes the subspaces span($\mathbf{\Gamma}$) and span($\mathbf{B}$), which in turn uniquely parametrizes the subspaces $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$. Given the uniqueness of the vector representation $\boldsymbol{\theta}$ and the matrix representation $\mathbf{\Gamma}, \mathbf{B}$, we use the two representations interchangeability throughout the text.

4.3.2 Generalized method of moments estimators

In this section, we propose a simple estimator of the subspaces parametrized by $\boldsymbol{\theta}$ based on the generalized method of moments (GMM).

To do this, consider the function $\mathbf{f}(\mathbf{Y}_i, \mathbf{X}_i; \boldsymbol{\theta})$ that satisfies the following condition:

$$\begin{aligned} &\mathbb{E}\{\mathbf{f}(\mathbf{Y}_i, \mathbf{X}_i; \boldsymbol{\theta})\} = \mathbf{0}, \quad \text{where} \\ &\mathbf{f}(\mathbf{Y}_i, \mathbf{X}_i; \boldsymbol{\theta}) = \{\mathbf{g}(\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}_i) - \mathbb{E}\{\mathbf{g}(\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}_i)\}\}\{\mathbf{h}(\mathbf{\Gamma}^T\mathbf{Y}_i, \mathbf{X}_i) - \mathbb{E}\{\mathbf{h}(\mathbf{\Gamma}^T\mathbf{Y}_i, \mathbf{X}_i)\}\}. \end{aligned} \tag{4.3}$$

Here, $\mathbf{g} : \mathbb{R}^{r-u-d} \to \mathbb{R}^{d_g\times 1}$ and $\mathbf{h} : \mathbb{R}^{u+p} \to \mathbb{R}^{1\times d_h}$, $d_g d_h \geq q$, are continuously differentiable functions where the covariance matrix $\text{Var}\{\mathbf{f}(\mathbf{Y}_i, \mathbf{X}_i; \boldsymbol{\theta})\}$ exists. There are many functions $\mathbf{g}$ and $\mathbf{h}$ that satisfy (4.3). For example, if $\mathbf{g}$, and $\mathbf{h}$ are identity functions, i.e. $\mathbf{g}(\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}_i) = \mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}_i$ and $\mathbf{h}(\mathbf{\Gamma}^T\mathbf{Y}_i, \mathbf{X}_i) = \{(\mathbf{\Gamma}^T\mathbf{Y}_i)^T, \mathbf{X}_i^T\}$, then $\mathbb{E}(\mathbf{f}) = \mathbf{0}$ because

$$\begin{aligned} &\mathbb{E}[\{\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}_i - \mathbb{E}(\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}_i)\}\{\mathbf{\Gamma}^T\mathbf{Y}_i - \mathbb{E}(\mathbf{\Gamma}^T\mathbf{Y}_i)\}^T] \\ =&\text{Cov}(\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}_i, \mathbf{\Gamma}^T\mathbf{Y}_i) = \mathbf{0}, \\ &\mathbb{E}[\{\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}_i - \mathbb{E}(\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}_i)\}\{\mathbf{X}_i - \mathbb{E}(\mathbf{X}_i)\}^T] \\ =&\text{Cov}(\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}_i, \mathbf{X}_i) = \mathbf{0}, \end{aligned}$$

where the latter equalities use the fact that Conditions 4.3.1 and 4.3.2 imply $\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}_i \perp\!\!\!\perp (\mathbf{\Gamma}^T\mathbf{Y}_i, \mathbf{X}_i)$.

Theorem 4.3.1 shows that solving the empirical counterparts to (4.3) will lead to $\sqrt{n}$-consistent and asymptotically normal estimators of $\boldsymbol{\theta}$.

Theorem 4.3.1 *Suppose* $\mathbf{f}(\mathbf{Y}_i, \mathbf{X}_i; \boldsymbol{\theta})$ *is a continuously differentiable function that satisfy equation* (4.3)*. Consider the following GMM estimator:*

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \| \sum_{i=1}^{N} \mathbf{f}(\mathbf{Y}_i, \mathbf{X}_i; \boldsymbol{\theta}) \|_2^2.$$

Under the regularity conditions (S1)–(S6) in Section A of the Supplements, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N\big\{\mathbf{0}, (\mathbf{C}_1^T \mathbf{C}_1)^{-1} \mathbf{C}_1^T \mathbf{D}_1 \mathbf{C}_1 (\mathbf{C}_1^T \mathbf{C}_1)^{-1}\big\},$$

where $\mathbf{C}_1 = \mathbb{E}\{\partial \mathrm{vec}(\mathbf{f}) / \partial \boldsymbol{\theta}^{\mathrm{T}}\}$ *and* $\mathbf{D}_1 = \mathrm{Var}\{\mathbf{f}(\mathbf{Y}_i, \mathbf{X}_i; \boldsymbol{\theta})\}$.

The regularity conditions (S1)-(S6) concern bounded moments, differentiability of $\mathbf{f}$, and identifiability of the population GMM equation. These are standard conditions for GMM estimators to achieve consistency and asymptotic normality; see Chapter 2.2.3 in Newey and McFadden (1994) for a textbook exposition.

Compared to existing methods to estimate the inner envelope (or the envelope) Cook et al. (2010, 2016); Cook and Zhang (2015a,b), the GMM approach does not explicitly rely a parametric model between $\mathbf{Y}$ and $\mathbf{X}$ nor a parametric, distributional assumption on $\boldsymbol{\varepsilon}$ to achieve consistency and asymptotic normality. Also, solving the estimating equations in Theorem 4.3.1 is far simpler than solving a non-convex problem, say in (4.2), and can be done with existing statistical packages for GMMs (Chaussé, 2010).

One limitation of Theorem 4.3.1 is that it does not specify which functions $\mathbf{g}$ and $\mathbf{h}$ to use. That is, while all functions satisfying the conditions in Theorem 4.3.1 will lead to an estimator that is consistent and asymptotically normal, some functions may lead to smaller asymptotic variance than others. The next section explores this question formally by deriving the semiparametric efficient lower bound for $\boldsymbol{\theta}$ and showing an estimator that can achieve this lower bound.

4.4 Semiparametric Efficiency

While the GMM estimator is simple and computationally tractable compared to existing methods, it may not lead to the most efficient estimator of $\boldsymbol{\theta}$ under Conditions 4.3.1 and 4.3.2. In this section, we propose a semiparametric efficient estimator of $\boldsymbol{\theta}$ by directly deriving the semiparametric efficient score under Conditions 4.3.1 and 4.3.2.

4.4.1 Efficient Score

We start by characterizing the orthogonal nuisance tangent space; the derivations are detailed in the Appendix.

Under Conditions 4.3.1 and 4.3.2, the joint density of $(\mathbf{Y}, \mathbf{X})$ can be decomposed as

$$\eta(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}) = \eta_1(\boldsymbol{\Gamma}^T\mathbf{Y} \mid \mathbf{X}; \boldsymbol{\theta})\eta_2(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})\eta_3(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}; \boldsymbol{\theta})\eta_4(\mathbf{X}), \tag{4.4}$$

where η_1 is the conditional probability density function (pdf) of $\boldsymbol{\Gamma}^T\mathbf{Y}$ given $\mathbf{X}$, η_2 is the conditional pdf of $\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}$ given $(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X})$, η_3 and η_4 are the pdfs of $\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}$ and $\mathbf{X}$, respectively. For simplicity, we use $\eta_{1,2,3}$ to denote the collection of density functions η_1, η_2 and η_3.

The orthogonal nuisance tangent space of $\boldsymbol{\theta}$ in (4.4) is

$$\begin{aligned}\Lambda^\perp = &\{\mathbf{f}(\mathbf{Y}, \mathbf{X}) \in \mathbb{R}^q : \mathbb{E}(\mathbf{f} \mid \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) \text{ is a function of } \mathbf{X};\ \mathbb{E}(\mathbf{f} \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) \\ &\quad \text{is a function of } \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X};\ \mathbb{E}(\mathbf{f} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) = \mathbf{0};\ \mathbb{E}(\mathbf{f} \mid \mathbf{X}) = \mathbf{0}\}.\end{aligned}$$

We remark that the function $\mathbf{f}(\mathbf{Y}_i, \mathbf{X}_i; \boldsymbol{\theta})$ inside the GMM estimator and defined in (4.3) satisfies $\mathbf{M}\mathbf{f}(\mathbf{Y}_i, \mathbf{X}_i; \boldsymbol{\theta}) \in \Lambda^\perp$ for all $\mathbf{M} \in \mathbb{R}^{q \times d_g d_h}$.

Let $\boldsymbol{\Delta}_1(\mathbf{Y}, \mathbf{X}) = \mathbf{Y} - \mathbb{E}(\mathbf{Y} \mid \mathbf{X})$, and $\boldsymbol{\Delta}_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) = \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) - \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})$. Given the orthogonal nuisance tangent space $\Lambda^\perp$, the efficient score of $\boldsymbol{\theta}$ has the form $S_{\text{eff}} = (S_{\text{eff},\boldsymbol{\gamma}}, S_{\text{eff},\mathbf{b}})$ with

$$\begin{aligned}S_{\text{eff},\boldsymbol{\gamma}} = &\operatorname{vec}^T\left\{\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\Delta}_1\frac{\partial \log \eta_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T}\right\}\frac{\partial \operatorname{vec}(\boldsymbol{\Gamma})}{\partial \boldsymbol{\gamma}^T} + \operatorname{vec}^T\left\{(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} + \boldsymbol{\Delta}_2)\frac{\partial \log \eta_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\mathbf{B}_0^T\right\}\frac{\partial \operatorname{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T} \\ &+ \operatorname{vec}^T\left[\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Delta}_1\left\{\frac{\partial \log \eta_2}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\mathbf{B}^T + \frac{\partial \log \eta_2}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\mathbf{B}_0^T\right\}\right]\frac{\partial \operatorname{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T}, \\ S_{\text{eff},\mathbf{b}} = &\operatorname{vec}^T\left\{\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_2\frac{\partial \log \eta_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\right\}\frac{\partial \operatorname{vec}(\mathbf{B}_0)}{\partial \mathbf{b}^T}.\end{aligned}$$

Interestingly, the efficient score of $S_{\text{eff},\mathbf{b}}$ only relies on the density of η_3. This implies that once $\boldsymbol{\gamma}$ is known, we only need the marginal density η_3 to estimate $\mathbf{b}$ and the additional information contained in η_2 does not improve the efficiency of $\mathbf{B}$. In contrast, the efficient score of $\boldsymbol{\gamma}$ involves all three densities $\eta_{1,2,3}$.

If all the nuisance functions, specifically the partial derivatives of the log likelihoods, i.e. $\partial \log \eta_1/\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T$, $\partial \log \eta_2/\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T$, $\partial \log \eta_2/\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T$, $\partial \log \eta_3/\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T$, and $\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2$ are known, we can obtain an efficient estimator of $\boldsymbol{\theta}$ by solving the score equation $\mathbb{E}\{S_{\text{eff}}(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})\} = \mathbf{0}$ for $\boldsymbol{\theta}$. However, in practice, they are unknown and

must be estimated. The next two subsections discuss different approaches of estimating these nuisance functions, specifically the aforementioned partial derivatives, which will lead to either a globally or locally efficient estimator of $\boldsymbol{\theta}$.

4.4.2 Globally efficient estimator

The globally efficient estimator uses kernel-based, nonparametric estimators of $\eta_{1,2,3}$, $\boldsymbol{\Delta}_1$, and $\boldsymbol{\Delta}_2$ as plug-ins estimators and solves for $\boldsymbol{\theta}$. Specifically, let K_h be a kernel function that satisfies Assumption A5, say a uniform kernel or an Epanechnikov kernel. The density η_1 is estimated by a kernel density estimator of the form

$$\hat{\eta}_1(\boldsymbol{\Gamma}^\intercal \mathbf{y}, \mathbf{x}) = \frac{\sum_{i=1}^n K_h(\boldsymbol{\Gamma}^\intercal \mathbf{Y}_i - \boldsymbol{\Gamma}^T \mathbf{y}) K_h(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^n K_h(\mathbf{X}_i - \mathbf{x})}.$$

Based on the estimator $\hat{\eta}_1(\boldsymbol{\Gamma}^\intercal \mathbf{y}, \mathbf{x})$, $\partial \log \hat{\eta}_1 / \partial (\boldsymbol{\Gamma}^T \mathbf{Y})^T$ can be calculated in closed form, i.e.

$$\frac{\partial \log \hat{\eta}_1}{\partial (\boldsymbol{\Gamma}^T \mathbf{y})^T} = \frac{\hat{\eta}_1'(\boldsymbol{\Gamma}^\intercal \mathbf{y}, \mathbf{x})}{\hat{\eta}_1(\boldsymbol{\Gamma}^\intercal \mathbf{y}, \mathbf{x})} = \frac{\sum_{i=1}^n K_h'(\boldsymbol{\Gamma}^\intercal \mathbf{Y}_i - \boldsymbol{\Gamma}^T \mathbf{y}) K_h(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^n K_h(\boldsymbol{\Gamma}^\intercal \mathbf{Y}_i - \boldsymbol{\Gamma}^T \mathbf{y}) K_h(\mathbf{X}_i - \mathbf{x})}, \tag{4.5}$$

where K_h' is the first order derivative of K_h. Estimation of the other two densities η_2 and η_3 and their partial derivatives are similar to η_1 and stated in Section B.1 of the Appendix.

Similarly, to estimate $\boldsymbol{\Delta}_1$, we use nonparametric kernel regressions of the form

$$\hat{\boldsymbol{\Delta}}_1(\mathbf{y}, \mathbf{x}) = \mathbf{y} - \frac{\sum_{i=1}^N \mathbf{Y}_i K_h(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^N K_h(\mathbf{X}_i - \mathbf{x})}. \tag{4.6}$$

The estimation of $\boldsymbol{\Delta}_2$ is similar and given in Section B.1 of the Supplements. For additional details on kernel-based nonparametric estimators, see Hayfield and Racine (2008).

Let $\hat{S}_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \hat{\eta}_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta})$ denote the sample version of the efficient score where we replace the unknown nuisance functions with their nonparametrically estimated counterparts. The globally efficient estimator is $\hat{\boldsymbol{\theta}}$ that solves the following estimating equation:

$$\frac{1}{n} \sum_{i=1}^n \hat{S}_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \hat{\eta}_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{4.7}$$

Theorem 4.4.1 shows that under Assumptions A1-A5, the estimator $\hat{\boldsymbol{\theta}}$ obtained from equation (4.7) is $\sqrt{n}$-consistent and efficient.

Assumption A1 *(The true conditional densities $\eta_{1,2,3}$) The true conditional densities $\eta_{1,2,3}$ are bounded away from 0 and ∞. The third order derivatives of $\log \eta_{1,2,3}$ around $\boldsymbol{\theta}$ are locally Lipschitz-continuous. Also, there exists a compact set $\boldsymbol{\Theta}$ which contains the true parameter value $\boldsymbol{\theta}$.*

Assumption A2 *(Identifiability of S_{eff}) The solution to the estimating equation $\mathbb{E}\{S_{\text{eff}}(\mathbf{Y}, \mathbf{X}, \eta_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})\} = \mathbf{0}$ is unique.*

Assumption A3 *(Estimation of $\hat{\eta}_{1,2,3}$) The estimators $\hat{\eta}_{1,2,3}$ are obtained through a kernel density estimator with a common bandwidth h. The bandwidth satisfies $nh^8 \to 0$ and $nh^{2(1+s)} \to \infty$ as $n \to \infty$, where $s = \max(r + p - u, p + u)$.*

Assumption A4 *(Smoothness of $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$) The second order derivatives of $\mathbb{E}(\mathbf{Y} \mid \mathbf{X})$ and $\mathbb{E}(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X})$ with respect to $\mathbf{X}$ and $(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X})$ are continuous for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.*

Assumption A5 *(Kernel function) Consider a univariate kernel $K_h(x)$, $0 \neq \int x^2 K_h(x) < \infty$, which is bounded, symmetric, has a compact support on $[-h, h]$ and has a bounded second derivative. The multivariate kernel $K_h(\mathbf{x})$ for a d-dimensional $\mathbf{x}$ has the form $K_h(\mathbf{x}) = \prod_{j=1}^{d} K_h(x_j)$ for $\mathbf{x} = (x_1, \ldots, x_d)^T$.*

Theorem 4.4.1 *Under Assumptions A1–A5 and Conditions 4.3.1–4.3.2, the estimator $\hat{\boldsymbol{\theta}}$ obtained from solving (4.7) achieves the semiparametric efficiency bound $\mathcal{V}_{\boldsymbol{\theta}} = \mathbb{E}(S_{\text{eff}} S_{\text{eff}}^T)^{-1}$, i.e. as $n \to \infty$, $\hat{\boldsymbol{\theta}}$ satisfies*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N\big(\mathbf{0}, \mathcal{V}_{\boldsymbol{\theta}}\big).$$

Assumptions A1–A5 are used to achieve consistency of kernel-based nonparametric estimators at sufficiently fast rates. These assumptions are not new and have been used in past works; see Chapter 1 in Li and Racine (2007) for one textbook example. Also, under the original inner envelope framework where the linear model in (4.1) is the true model, Assumptions A1, A2, and A5 concerning the data generating model would hold (Su and Cook, 2012).

4.4.3 Local efficiency and a robust score S^*_{eff}

In some settings, investigators may have working models of the densities $\eta_{1,2,3}$, denoted as $\eta^*_{1,2,3}$. For example, suppose the investigator imposes the working models for $\eta^*_{1,2,3}$ based on the multivariate normal model in (4.1). Then, under (4.1), the densities of $\eta^*_{1,2,3}$ follow multivariate normal distributions

$$\begin{aligned}
&\eta^*_1(\boldsymbol{\Gamma}^T\mathbf{Y}_i, \mathbf{X}_i) \sim N(\boldsymbol{\Gamma}^T\mathbf{Y}_i - \boldsymbol{\zeta}_1^T\mathbf{X}_i, \boldsymbol{\Omega}),\\
&\eta^*_2(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i, \mathbf{X}_i) \sim N(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i - \boldsymbol{\zeta}_2^T\mathbf{X}_i - \boldsymbol{\mu}_2^T\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i, \boldsymbol{\Sigma}_2),\\
&\eta^*_3(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i) \sim N(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i, \mathbf{B}_0^T\boldsymbol{\Omega}_0\mathbf{B}_0),
\end{aligned}$$

where $\boldsymbol{\mu}_2 = (\mathbf{B}_0^T\boldsymbol{\Omega}_0\mathbf{B}_0)^{-1}\mathbf{B}_0^T\boldsymbol{\Omega}_0\mathbf{B}$ and $\boldsymbol{\Sigma}_2 = \mathbf{B}^T\boldsymbol{\Omega}_0^{\frac{1}{2}}\mathbf{P}_{\boldsymbol{\Omega}_0^{1/2}\mathbf{B}}\boldsymbol{\Omega}_0^{\frac{1}{2}}\mathbf{B}$; see Su and Cook (2012) for details. Note that the terms $\boldsymbol{\zeta}_1 \in \mathbb{R}^{u\times p}$ and $\boldsymbol{\zeta}_2 \in \mathbb{R}^{(p-u)\times p}$ are the coordinates of $\boldsymbol{\beta}$ under the basis matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0\mathbf{B}$, respectively, and the terms $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are the covariance matrices of $\boldsymbol{\Gamma}^T\mathbf{Y}_i$ and $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i$ respectively. Also, based on the form of $\eta^*_{1,2,3}$, their partial derivatives in S_{eff} can be written in closed-form as

$$\begin{aligned}
&\frac{\partial \log \eta^*_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y}_i)^T} = -(\boldsymbol{\Gamma}^T\mathbf{Y}_i - \boldsymbol{\zeta}_1^T\mathbf{X}_i)^T\boldsymbol{\Omega}^{-1},\\
&\frac{\partial \log \eta^*_2}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T} = -(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i - \boldsymbol{\zeta}_2^T\mathbf{X}_i - \boldsymbol{\mu}_2^T\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T\boldsymbol{\Sigma}_2^{-1},\\
&\frac{\partial \log \eta^*_2}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T} = (\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i - \boldsymbol{\zeta}_2^T\mathbf{X}_i - \boldsymbol{\mu}_2^T\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2^T,\\
&\frac{\partial \log \eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T} = -(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T(\mathbf{B}_0^T\boldsymbol{\Omega}_0\mathbf{B}_0)^{-1}.
\end{aligned}$$

Then, under the multivariate normal working model, we only have to estimate finite-dimensional nuisance parameters $\boldsymbol{\zeta}_1$, $\boldsymbol{\zeta}_2$, $\boldsymbol{\Omega}$, and $\boldsymbol{\Omega}_0$ to characterize the densities $\eta_{1,2,3}$ and these nuisance parameters can be estimated at $\sqrt{n}$- rates. Also, if the working models for $\eta^*_{1,2,3}$ are correct, $S_{\text{eff}}(\mathbf{Y}, \mathbf{X}, \eta^*_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta}) \in \Lambda^\perp$ and the estimator based on solving for $\boldsymbol{\theta}$ in the estimating equation $\mathbb{E}\{S_{\text{eff}}(\mathbf{Y}, \mathbf{X}, \eta^*_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})\} = \mathbf{0}$ is semiparametrically efficient.

However, more often than not, the working models $\eta^*_{1,2,3}$ are likely incorrect, which implies $S_{\text{eff}}(\mathbf{Y}, \mathbf{X}, \eta^*_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta}) \notin \Lambda^\perp$ and the solution to the estimating equation $\mathbb{E}\{S_{\text{eff}}(\mathbf{Y}, \mathbf{X}, \eta^*_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})\} = \mathbf{0}$ will lead to an inconsistent estimator of $\boldsymbol{\theta}$. To overcome the sensitivity of the efficient score to potential mis-specifications of the working models, we propose an alternative, robust score in $\Lambda^\perp$, denoted as

$S^*_{\text{eff}} = (S^{*\mathsf{T}}_{\text{eff},\boldsymbol{\gamma}}, S^{*\mathsf{T}}_{\text{eff},\mathbf{b}})^T$ and stated below:

$$
\begin{aligned}
S^*_{\text{eff},\boldsymbol{\gamma}} = \text{vec}^T\Bigg[\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\Delta}_1\bigg\{\frac{\partial \log \eta^*_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T} - \mathbb{E}\bigg(\frac{\partial \log \eta^*_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T}\;\Big|\;\mathbf{X}\bigg)\bigg\}\Bigg]\frac{\partial \text{vec}(\boldsymbol{\Gamma})}{\partial \boldsymbol{\gamma}^T} \\
+ \text{vec}^T\Bigg[(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} + \boldsymbol{\Delta}_2)\bigg\{\frac{\partial \log \eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T} - \mathbb{E}\bigg(\frac{\partial \log \eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\bigg)\bigg\}\mathbf{B}_0^T\Bigg]\frac{\partial \text{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T} \\
+ \text{vec}^T\Bigg[\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Delta}_1\bigg\{\frac{\partial \log \eta^*_2}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T} - \mathbb{E}\bigg(\frac{\partial \log \eta^*_2}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\;\Big|\;\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\bigg)\bigg\}\mathbf{B}^T\Bigg]\frac{\partial \text{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T}, \\
+ \text{vec}^T\Bigg[\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Delta}_1\bigg\{\frac{\partial \log \eta^*_2}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T} - \mathbb{E}\bigg(\frac{\partial \log \eta^*_2}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\;\Big|\;\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\bigg)\bigg\}\mathbf{B}_0^T\Bigg]\frac{\partial \text{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T}, \\
S^*_{\text{eff},\mathbf{b}} = \text{vec}^T\Bigg[\boldsymbol{\Gamma}_0^T\boldsymbol{\varepsilon}_2\bigg\{\frac{\partial \log \eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T} - \mathbb{E}\bigg(\frac{\partial \log \eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\bigg)\bigg\}\Bigg]\frac{\partial \text{vec}(\mathbf{B}_0)}{\partial \mathbf{b}^T}.
\end{aligned}
$$

An appealing feature of the new score $S^*_{\text{eff}}(\mathbf{Y}, \mathbf{X}, \eta^*_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})$ is that when the working models of $\eta^*_{1,2,3}$ are correctly specified, we have $S^*_{\text{eff}} = S_{\text{eff}}$ and the resulting estimator of $\boldsymbol{\theta}$ based on solving $\mathbb{E}\{S^*_{\text{eff}}(\mathbf{Y}, \mathbf{X}, \eta^*_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})\} = \mathbf{0}$ is semiparametrically efficient. But, if any of the working models are misspecified, S^*_{eff} is still an element in $\Lambda^{\perp}$ and thus, the estimator will still be consistent and asymptotically normal. In other words, S^*_{eff} will always guarantee a consistent estimator of $\boldsymbol{\theta}$ irrespective of the choice of the working models and the estimator will be efficient if the working models are correct.

The new, robust score S^*_{eff} contains new nuisance parameters in the form of conditional expectations of the partial log of $\eta^*_{1,2,3}$. Depending on the choice of the working models, these conditional expectations may or may not have to be estimated. For example, under the aforementioned working models based on the multivariate normal distribution, S^*_{eff} simplifies to

$$
\begin{aligned}
S^*_{\text{eff},\boldsymbol{\gamma}} = &- \text{vec}^T(\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\Delta}_1\boldsymbol{\Delta}_1^T\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1})\frac{\partial \text{vec}(\boldsymbol{\Gamma})}{\partial \boldsymbol{\gamma}^T} - \text{vec}^T\big\{\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Delta}_1\boldsymbol{\Delta}_1^T\boldsymbol{\Gamma}_0\mathbf{B}\boldsymbol{\Sigma}_2^{-1}(\mathbf{B}^T - \boldsymbol{\mu}_2^T\mathbf{B}_0^T) \\
&+ (\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2)\mathbf{Y}\boldsymbol{\Gamma}_0\mathbf{B}_0(\mathbf{B}_0^T\boldsymbol{\Omega}\mathbf{B}_0)^{-1}\mathbf{B}_0^T\big\}\frac{\partial \text{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T}, \\
S^*_{\text{eff,b}} = &- \text{vec}^T\{\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_2\mathbf{Y}\boldsymbol{\Gamma}_0\mathbf{B}_0(\mathbf{B}_0^T\boldsymbol{\Omega}\mathbf{B}_0)^{-1}\}\frac{\partial \text{vec}(\mathbf{B}_0)}{\partial \mathbf{b}^T}.
\end{aligned}
$$

Notice that the conditional expectations of the partial logs of $\eta^*_{1,2,3}$ are not present and the only unknown quantities in S^*_{eff} are $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$, which can be estimated using the same nonparametric regression estimators from the globally efficient estimator in Section 4.4.2.

However, if the working models are more complex, investigators may have to

estimate these conditional expectations. Here, we propose to estimate them in the same manner as estimating $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ with a kernel regression estimator. For example, for $\mathbb{E}\{\partial \log \eta_1^*/\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T \mid \mathbf{X} = \mathbf{x}\}$, we propose the following nonparametric estimator

$$\hat{\mathbb{E}}\left(\frac{\partial \log \eta_1^*}{\partial(\boldsymbol{\Gamma}^\intercal \mathbf{Y}_i)^T} \;\middle|\; \mathbf{x}\right) = \frac{\sum_{i=1}^N \partial \log \eta_1^*/\partial(\boldsymbol{\Gamma}^\intercal \mathbf{Y}_i)^T K_h(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^N K_h(\mathbf{X}_i - \mathbf{x})}.$$

For additional details on the nonparametric estimators of $\mathbb{E}\{\partial \log \eta_2^*/\partial(\mathbf{B}^\intercal\boldsymbol{\Gamma}_0^\intercal\mathbf{Y})^T \mid \mathbf{B}_0^\intercal\boldsymbol{\Gamma}_0^\intercal\mathbf{y}, \mathbf{X}\}$, $\mathbb{E}\{\partial \log \eta_2^*/\partial(\mathbf{B}_0^\intercal\boldsymbol{\Gamma}_0^\intercal\mathbf{Y})^T \mid \mathbf{B}_0^\intercal\boldsymbol{\Gamma}_0^\intercal\mathbf{Y}, \mathbf{X}\}$, see Section B.1 of the Supplements. Practically speaking, we recommend investigators start with the multivariate normal working model as it not only mirrors the original, parametric inner envelope estimator, but also it leads to a simpler S_{eff}^*.

Let $\hat{S}_{\text{eff}}^*(\mathbf{Y}, \mathbf{X}, \hat{\eta}_{1,2,3}^*, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta})$ denote the sample version of the score with the posited densities $\eta_{1,2,3}^*$ and the estimated $\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2$ from the globally efficient estimator in (4.6). We define $\hat{\boldsymbol{\theta}}$ to be the solution to the following estimating equation:

$$\frac{1}{n}\sum_{i=1}^n \hat{S}_{\text{eff}}^*(\mathbf{Y}_i, \mathbf{X}_i, \hat{\eta}_{1,2,3}^*, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{4.8}$$

Theorem 4.4.2 shows that under Assumptions B1-B4, the solution to (4.8) will always be asymptotically normal and be locally efficient if all the working densities are correctly specified.

Assumption B1 *(Working models $\eta_{1,2,3}^*$) The working models $\eta_{1,2,3}^*$ are bounded away from 0 and infinity. The Hessians of $\log \eta_1^*$, $\log \eta_2^*$, $\log \eta_3^*$ are positive definite and bounded on a compact set $\boldsymbol{\Theta}$ that contains $\boldsymbol{\theta}$. The third order derivatives of $\log \eta_{1,2,3}^*$ around $\boldsymbol{\theta}$ are locally Lipschitz-continuous.*

Assumption B2 *(Identifiability of S_{eff}^*) The solution to the estimating equation $\mathbb{E}\{S_{\text{eff}}^*(\mathbf{Y}, \mathbf{X}, \eta_{1,2,3}^*, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})\} = \mathbf{0}$ is unique.*

Assumption B3 *(Smoothness of conditional expectations) The second order derivatives of $\mathbb{E}\{\partial \log \eta_1^*/\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T \mid \mathbf{X}\}$, $\mathbb{E}\{\partial \log \eta_2^*/\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\}$ and $\mathbb{E}\{\partial \log \eta_2^*/\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\}$ are uniformly continuous for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. If they are estimated via kernel regressions in Section B.1 of the Supplements, the bandwidth h satisfies $nh^8 \to 0$ and $nh^{2(p+r-u-d)} \to \infty$.*

Assumption B4 *(Data density) The data density for $(\mathbf{X}, \mathbf{Y})$ is bounded away from 0 and infinity and has twice continuously differentiable derivatives.*

Theorem 4.4.2 *Suppose Assumptions B1–B4, A4–A5 and Conditions 4.3.1–4.3.2 hold. Then, the estimator in* (4.8) *is consistent and asymptotically normal, i.e.*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N\{\mathbf{0}, \mathbf{C}_2^{-1}\mathbf{D}_2(\mathbf{C}_2^{-1})^T\},$$

where

$$\mathbf{C}_2 = \mathbb{E}\left\{\frac{\partial S^*_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta^*_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\}, \quad \mathbf{D}_2 = \mathbb{E}\left\{S^*_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta^*_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})^{\otimes 2}\right\}.$$

Additionally, if the working models are correctly specified, i.e. $\eta_1^* = \eta_1$, $\eta_2^* = \eta_2$ *and* $\eta_3^* = \eta_3$, *the estimator is locally efficient.*

We make some remarks about the assumptions underlying Theorem 4.4.2. First, the aforementioned multivariate normal model automatically satisfy Assumptions B1–B3 and there is no need to nonparametrically estimate the conditional expectations of the partial log likelihoods. Second, the consistency and asymptotic normality of the locally efficient estimator requires weaker regularity condition for the bandwidth rate than the globally efficient estimator. Specifically, the bandwidth rate $nh^{2(1+s)} \to \infty$ from Assumption A3 implies $nh^{2(p+r-u-d)} \to \infty$ in Assumption B3.

4.5 Computational and Other Considerations

4.5.1 An Alternating Algorithm to Solve Estimating Equations

Both the globally and locally efficient estimators in equations (4.7) and (4.8), respectively, require solving potentially non-linear estimating equations. While there are many general-purpose, optimization procedures to solve such equations (e.g., Chapter 11 in Nocedal and Wright (2006)), we propose a procedure based on an alternating algorithm The optimization algorithm for the globally efficient estimator is detailed in Algorithm 1 and the optimization algorithm for the locally efficient estimator is detailed in Algorithm 2.

We make some remarks about both algorithms. First, both algorithms alternate between estimating $\boldsymbol{\theta}$ and the relevant nuisance parameters in the scores S_{eff} and S^*_{eff}. The rationale behind alternating between these two estimation steps is that investigators can directly use popular, off-the-shelf software for nonparametric

Algorithm 1: An alternating algorithm to solve equation (4.7)

Inputs: (1) data $(\mathbf{Y}_i, \mathbf{X}_i)$, $i = 1, \ldots, n$; (2) kernel function K_h; (3) convergence threshold $\delta > 0$.
1. Initialize $\boldsymbol{\theta}$ using the GMM estimator in Section 4.3 and denote it as $\boldsymbol{\theta}^{(0)}$.
2. Estimate $\boldsymbol{\Delta}_1$ using equation (4.6).
while $|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}| > \delta$
3. Obtain $\boldsymbol{\Gamma}^{(t)}$, $\boldsymbol{\Gamma}_0^{(t)}$, $\mathbf{B}^{(t)}$, $\mathbf{B}_0^{(t)}$ from $\boldsymbol{\theta}^{(t)}$ using Lemma 4.3.1.
4. Using the estimates in step 3, use the kernel estimators to nonparametrically estimate $\eta_{1,2,3}$ and $\boldsymbol{\Delta}_2$.
5. Using the estimates in step 4, solve the estimating equation in (4.7) for $\boldsymbol{\theta}$ and denote it as $\boldsymbol{\theta}^{(t+1)}$.
End while
Output: Final estimator $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(t+1)}$.

Algorithm 2: An alternating algorithm to solve equation (4.8)

Inputs: (1) data $(\mathbf{Y}_i, \mathbf{X}_i)$, $i = 1, \ldots, n$; (2) working models of $\eta^*_{1,2,3}$; (3) kernel function K_h; (4) convergence threshold $\delta > 0$.
1. Initialize $\boldsymbol{\theta}$ using the GMM estimator in Section 4.3 and denote it as $\boldsymbol{\theta}^{(0)}$.
2. Estimate $\boldsymbol{\Delta}_1$ using equation (4.6).
while $|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}| > \delta$
2. Obtain $\boldsymbol{\Gamma}^{(t)}$, $\boldsymbol{\Gamma}_0^{(t)}$, $\mathbf{B}^{(t)}$, $\mathbf{B}_0^{(t)}$ from $\boldsymbol{\theta}^{(t)}$ using Lemma 4.3.1.
3. Using the estimates in step 2, use the kernel estimators to estimate $\boldsymbol{\Delta}_2$.
4. Using the estimates in step 2, estimate relevant parameters in the working models $\eta^*_{1,2,3}$.
5. If necessary, estimate $\mathbb{E}\{\partial \log \eta_1^* / \partial (\boldsymbol{\Gamma}^{(t)\mathsf{T}} \mathbf{Y})^T \mid \mathbf{X}\}$,
$\mathbb{E}\{\partial \log \eta_2^* / \partial (\mathbf{B}^{(t)\mathsf{T}} \boldsymbol{\Gamma}_0^{(t)\mathsf{T}} \mathbf{Y})^T \mid \mathbf{B}_0^{(t)\mathsf{T}} \boldsymbol{\Gamma}_0^{(t)\mathsf{T}} \mathbf{y}, \mathbf{X}\}$,
$\mathbb{E}\{\partial \log \eta_2^* / \partial (\mathbf{B}_0^{(t)\mathsf{T}} \boldsymbol{\Gamma}_0^{(t)\mathsf{T}} \mathbf{Y})^T \mid \mathbf{B}_0^{(t)\mathsf{T}} \boldsymbol{\Gamma}_0^{(t)\mathsf{T}} \mathbf{Y}, \mathbf{X}\}$ via equations (S7)–(S10).
6. Using the estimates from steps 3 to steps 5, solve the estimating equation in (4.8) and denote it as $\boldsymbol{\theta}^{(t+1)}$.
End while
Output: The final estimator $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(t+1)}$.

kernel-based estimation. For example, for nonparametric, kernel regression estimators of $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$, investigators can use the R function "npreg" in the R package "np" Hayfield and Racine (2008). For the densities $\eta_{1,2,3}$ and their derivatives, investigators can use the R function "kde" and "kdde" in the R package "ks" Duong (2007). For choosing the bandwidth parameter, investigators can directly use the R functions "npregbw" and "Hpi" in the above two packages, which use cross validation to choose the bandwidth parameters; note that while there is a rate-driven choice of the bandwidth parameter h from our theory (see Theorem 4.4.1 and Theorem 4.4.2), in practice, cross validation is used. Second, to evaluate the partial derivatives of the basis matrices with respect to their vector representations (e.g. $\partial \text{vec}(\boldsymbol{\Gamma}_0)/\partial \boldsymbol{\gamma}^T$), we can use Lemma 4.3.1, which provides a continuous mapping between ($\boldsymbol{\Gamma}, \mathbf{B}$) and $\boldsymbol{\gamma}, \mathbf{b}$, and finite differencing Nocedal and Wright (2006). Third, both algorithms initialize with the GMM estimator, but other initializations are possible. In practice, we recommend using the GMM estimator with the identity functions $\mathbf{g}$ and $\mathbf{h}$ for simplicity.

4.5.2 Dimension Selection of the Inner Envelope

An important step in using any dimension-reduction procedure is selecting the reduced dimension subspace. For the inner envelope, we need to estimate the dimensions of $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ and a common approach to estimating the dimensions involves using Bayesian information criterion. But, this approach relies on a parametic model between the responses and the outcome whereas our semiparametric generalization does not impose such a model. Instead, we use a nonparametric, bootstrapped-based procedure in Dong and Li (2010); Ye and Weiss (2003) adapted to our inner envelope setting to estimate the dimension of the subspaces.

Formally, for each possible dimension of $\mathcal{S}_i$, we calculate the GMM estimators of the corresponding $\boldsymbol{\theta}$ with Theorem 4.3.1. We also take nonparametric bootstrapped samples of $(\mathbf{Y}, \mathbf{X})$ and obtain B bootstrapped estimates of $\boldsymbol{\theta}$.

$$(\hat{u}, \hat{d}) = \underset{u,d}{\text{argmax}} \frac{1}{B} \sum_{b=1}^{B} \{q^2(\hat{\mathcal{S}}_{1,u}, \hat{\mathcal{S}}^b_{1,u}) + q^2(\hat{\mathcal{S}}_{2,d}, \hat{\mathcal{S}}^b_{2,d}) + q^2(\hat{\mathcal{S}}_{3,r-u-d}, \hat{\mathcal{S}}^b_{3,r-u-d})\}, \quad (4.9)$$

where we assume the dimensions of $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ are greater or equal to 1. Here, q^2 is the Hotelling's vector correlation coefficient Hotelling (1936b) between two spaces.

That is, for any subspaces $\mathcal{A}$ and $\mathcal{B} \subset \mathbb{R}^r$, $q^2(\mathcal{A}, \mathcal{B})$ is defined as

$$q^2(\mathcal{A}, \mathcal{B}) = \det(\mathbf{B}^T \mathbf{A}\mathbf{A}^T \mathbf{B}),$$

where $\mathbf{A}$ and $\mathbf{B}$ are any basis matrices for the spaces $\mathcal{A}$ and $\mathcal{B}$. The correlation coefficient q^2 is bounded between 0 and 1 where higher value of q^2 indicates higher correlation between the two subspaces. In the extreme case where $q^2(\mathbf{A}, \mathbf{B}) = 1$, $\text{span}(\mathbf{A}) = \text{span}(\mathbf{B})$.

Roughly speaking, equation (4.9) selects the dimensions where the intra-correlation between the bootstrapped estimates of the subspaces and the estimated subspace are small. If the true dimension of a subspace is k^*, but we estimated the subspace assuming it is of dimension $k \neq k^*$, the intra-correlation between the bootstrapped estimates assuming k will be larger than that from assuming $k = k^*$ and thus, the selected dimension from (4.9) will be closer to the true dimension. For additional details behind using intra-correlation of bootstrapped estimates of dimension-reduced subspaces to estimate dimensions, see Dong and Li (2010); Ye and Weiss (2003).

4.6 Simulations Study

We conduct three simulation studies to numerically assess our proposed methods. The first two simulation studies concern linear and non-linear models with $r = 4$ responses and $p = 2$ regressors. We set the true dimension of $\mathcal{S}_1$,$\mathcal{S}_2$ and $\mathcal{S}_3$ to be 1, 1, 2 respectively, and consider five different sample sizes, $n = 100, 300, 500, 750, 1000$. We compare our methods with existing approaches and compare the performance of each method by computing $\text{dist}(\mathcal{S}_j, \hat{\mathcal{S}}_j) = \|\mathbf{P}_{\mathcal{S}_j} - \mathbf{P}_{\hat{\mathcal{S}}_j}\|_F$, with a smaller distance implying a better method. Additionally, for the linear simulation model in Section 4.6.1, we compute $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F^2$ and for the non-linear simulation model in Section 4.6.2, we compare the out-of-sample predictive root mean squared error. Finally, for the last simulation study, we generate our simulation model based on the well-known iris data Fisher (1936).

4.6.1 Linear model with normal errors

Consider the following model

$$\mathbf{Y}_i = \boldsymbol{\Gamma} f_1(\mathbf{X}_i) + \boldsymbol{\Gamma}_0 \mathbf{B} f_2(\mathbf{X}_i) + \boldsymbol{\varepsilon}_i, \quad X_{1i}, X_{2i} \sim U[-5, 5] \tag{4.10}$$

Let $\boldsymbol{\Gamma} = (0.50, 0.50, 0.50, 0.50)^T \in \mathbb{R}^{4\times 1}$, $\mathbf{B} = (1, 2, 3)^T/\sqrt{14} \in \mathbb{R}^{3\times 1}$, $\boldsymbol{\Gamma}_0 = (\boldsymbol{\Gamma}_{01}, \boldsymbol{\Gamma}_{02}, \boldsymbol{\Gamma}_{03}) \in \mathbb{R}^{4\times 3}$ with $\boldsymbol{\Gamma}_{01} = (0.50, -0.83, 0.17, 0.17)^T$, $\boldsymbol{\Gamma}_{02} = (0.50, 0.17, -0.83, 0.17)^T$, $\boldsymbol{\Gamma}_{03} = (0.50, 0.17, 0.17, -0.83)^T$. For this simulation study, we set f_1, f_2, and $\boldsymbol{\varepsilon}_i$ as follows

$$f_1(\mathbf{X}_i) = X_{1i}, \quad f_2(\mathbf{X}_i) = X_{1i} + X_{2i}, \quad \boldsymbol{\varepsilon}_i = \boldsymbol{\Gamma}\varepsilon_{1i} + \boldsymbol{\Gamma}_0\mathbf{B}(0.2, 0.2)\boldsymbol{\varepsilon}_{2i} + \boldsymbol{\Gamma}_0\mathbf{B}_0\boldsymbol{\varepsilon}_{2i},$$
$$\varepsilon_{1i} \sim N(0, 1), \quad \boldsymbol{\varepsilon}_{2i} \sim N\left(\mathbf{0}, 100\mathbf{I}_2\right), \quad i = 1, \ldots, n.$$

With the above specifications, the simulation model is equivalent to the linear model in equation (4.1) where $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}_1^T + \boldsymbol{\Gamma}_0\mathbf{B}\boldsymbol{\eta}_2^T$, $\boldsymbol{\eta}_1^T = (1, 0)$, $\boldsymbol{\eta}_2^T = (1, 1)$, $\boldsymbol{\Omega}_1 = 72.5$, $\boldsymbol{\Omega}_0 = (\boldsymbol{\omega}_{01}, \boldsymbol{\omega}_{02}, \boldsymbol{\omega}_{03})$, $\boldsymbol{\omega}_{01} = (0.25, 0.25, 0.25)^T$, $\boldsymbol{\omega}_{02} = (0.25, 68.25, -31.75)^T$, and $\boldsymbol{\omega}_{03} = (0.25, -31.75, 68.25)^T$. Also, Conditions 4.3.1 and 4.3.2 hold with the subspace dimensions $u = 1$ and $d = 1$. We apply our proposed methods in Sections 4.3 and 4.4 as well as the original inner envelope estimator developed under a normal linear model. We remark that for the GMM estimator, the functions $\mathbf{g}$ and $\mathbf{h}$ are chosen to be identity functions.

Figure 4.1 shows the results for $\text{dist}(\mathcal{S}_j, \hat{\mathcal{S}}_j)$ across different methods; for detailed numerical results, see Table C.2 of the Supplements. We see that the locally efficient estimator with correctly specified density performs better than the globally efficient estimator and the GMM estimator performs the worst. Also, as expected, the original inner envelope estimator performed the best since the simulation model is linear and the errors are normally distributed. Finally, we remark that across different sample sizes, the correct inner envelope dimension (i.e. $u = 1$, $d = 1$) was selected 97% of the time.

Next, we compare the estimates of the regression parameter $\boldsymbol{\beta}$. As discussed in Section 4.2, once we have an estimate of the basis matrices $\hat{\boldsymbol{\Gamma}}$ and $\hat{\mathbf{B}}$, we can estimate $\hat{\boldsymbol{\beta}}$ via projections; see the Appendix for details. Also, since the model is linear, we include four additional estimators for $\boldsymbol{\beta}$: the OLS estimator, the original envelope estimator, the response partial least squares (PLS) estimator Cook (2018a) and the oracle estimator that assumes that the inner envelope spaces are known a priori. Figure 4.2 shows $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F^2$ across different methods; see Table C.2 in the Supplements for additional details. Compared to the OLS estimator, there are gains from using either the original inner envelope estimator or our proposed estimators. Specifically, the GMM estimator has a smaller $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F^2$ than the OLS estimator, the PLS estimator, and the original envelope estimator. But, the GMM estimator is worse than the original inner envelope estimator, the locally efficient estimator and the globally efficient estimator. These observations agree with what's expected from

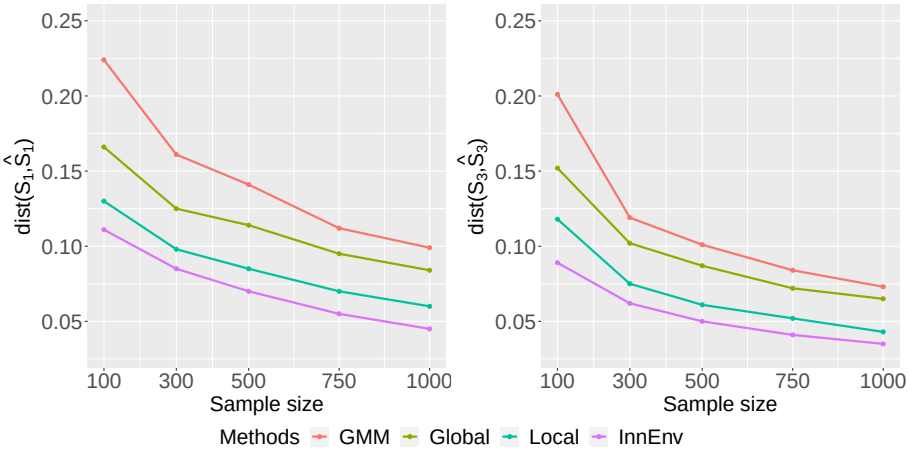


Figure 4.1: Distance between the true space $\mathcal{S}_j$ and the estimated space $\hat{\mathcal{S}}_j$, denoted as $\text{dist}(\mathcal{S}_j, \hat{\mathcal{S}}_j)$, from the linear simulation model in Section 4.6.1. GMM represents the GMM estimator, global represents the globally efficient estimator, local represents the locally efficient estimator, and InnEnv represents the original inner envelope estimator.

theory as the GMM estimator utilizes the inner envelope space compared to the OLS estimator, the PLS estimator, and the original envelope estimator, leading to better performance. But, the GMM estimator is not designed to be semiparametrically efficient compared to the local, global, and the original inner envelope estimator.

Finally, in Section C.3 of the Appendix, we calculate the out-of-sample predictive root mean squared error (RMSE) when $n = 500$. In short, because the underlying data generating model is linear, the predictive performance of the different methods mirror the performance of estimating $\boldsymbol{\beta}$.

4.6.2 Non-linear model with non-normal errors

For this simulation study, we consider the same general model in (4.10) and the subspace matrices $\boldsymbol{\Gamma}, \mathbf{B}$, except we consider the following f_1, f_2, and $\boldsymbol{\varepsilon}_i$:

$$
\begin{aligned}
&f_1(\mathbf{X}_i) = X_{1i}^2 \text{sgn}(X_{2i}), \qquad f_2(\mathbf{X}_i) = 20 \sin\{0.5(X_{1i} + X_{2i})\}, \\
&\boldsymbol{\varepsilon}_i = \boldsymbol{\Gamma}\varepsilon_{1i} + \boldsymbol{\Gamma}_0\mathbf{B} \cdot \mathbf{0.1}^T \boldsymbol{\varepsilon}_{2i} + \boldsymbol{\Gamma}_0\mathbf{B}_0\boldsymbol{\varepsilon}_{2i}, \\
&\varepsilon_{1i} \sim t_5(0, 1), \quad \boldsymbol{\varepsilon}_{2i} \sim t_5(\mathbf{0}, 100\mathbf{I}_2)
\end{aligned}
$$

Here, $t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate t distribution with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$ and degrees of freedom ν, and $\mathbf{0.1}$ denotes the vector $\mathbf{1} \times 0.1$. Overall, the above

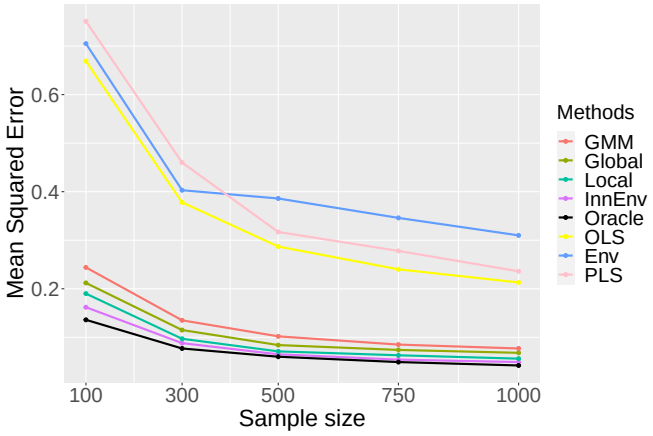


Figure 4.2: Mean squared error of estimating $\boldsymbol{\beta}$ (i.e. $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F^2$) in Section 4.6.1. GMM represents the GMM estimator, global represents the globally efficient estimator, local represents the locally efficient estimator, InnEnv represents the original inner envelope estimator, oracle represents the oracle OLS estimator where the inner envelope structure is known a priori, OLS represents the naive ordinary least squares estimator, Env represents the original envelope estimator, and PLS represents the partial least squares estimator.

specification creates a non-linear, non-normal, and heteroskedatsic model between the responses and the regressors. We also remark that Conditions 4.3.1 and 4.3.2 are satisfied for the above non-linear model.

Figure 4.3 shows the results of $\text{dist}(\mathcal{S}_j, \hat{\mathcal{S}}_j)$ under different methods; for additional details, see Table C.2 in the Supplements. Because the underlying model is non-linear and has non-normal errors, the original inner envelope estimator has a high dist even as n increases. In contrast, our proposed estimators which do not rely on linearity or normality show that the underlying subspaces are being estimated correctly with dist shrinking towards zero as n increases. Also, between the locally efficient estimator and the globally efficient estimator, the locally efficient estimator uses wrong normal working models for the densities η_1, η_2, and η_3 leading to worse performance than the globally efficient estimator, which estimate these densities nonparametrically via kernel regression. Finally, we remark that the correct inner envelope dimension $u = 1$, $d = 1$ was selected 95% of the time. Next, we evaluate the out-of-sample predictive RMSE when $n = 500$. To do this, we create pseudo-outcomes $\tilde{\mathbf{Y}}_i = (\hat{\boldsymbol{\Gamma}}^T\mathbf{Y}_i, \hat{\mathbf{B}}^T\hat{\boldsymbol{\Gamma}}_0^T\mathbf{Y}_i)$ based on the estimated subspaces from above and run a supervised learning algorithm with $\tilde{\mathbf{Y}}_i$ as the outcome and $\mathbf{X}_i$ as the predictor. Note that after training the supervised learning model and obtaining the predictions for the pseudo-outcome, we can transform the pseudo-outcome back to

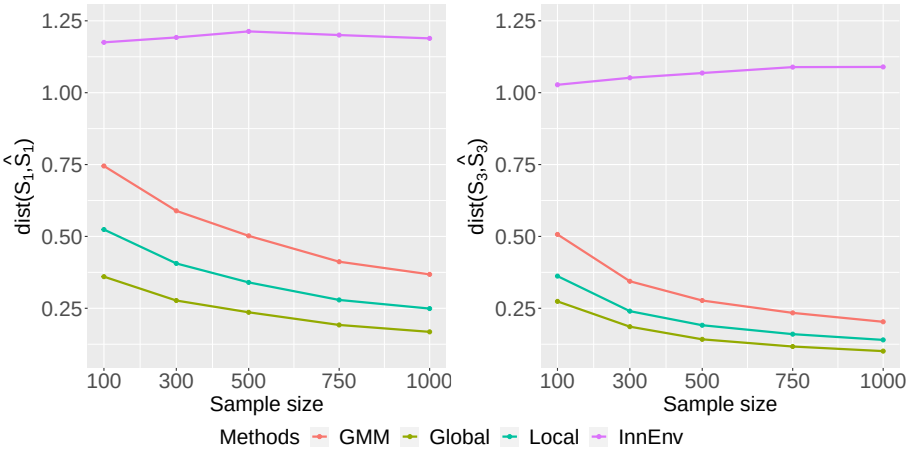


Figure 4.3: Distance between the true space $\mathcal{S}_j$ and the estimated space $\hat{\mathcal{S}}_j$, denoted as $\text{dist}(\mathcal{S}_j, \hat{\mathcal{S}}_j)$ for the non-linear model in Section 4.6.2. GMM represents the GMM estimator, global represents the globally efficient estimator, local represents the locally efficient estimator, and InnEnv represents the original inner envelope estimator.

the original outcome via $\hat{\mathbf{Y}}_i = (\hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\Gamma}}_0\hat{\mathbf{B}})\tilde{\mathbf{Y}}_i$. We randomly split 80% of the data into training data $(\mathbf{Y}_{train}, \mathbf{X}_{train})$ $(i = 1, \ldots, m)$ and the rest into test data $(\mathbf{Y}_{test}, \mathbf{X}_{test})$ $(i = m+1, \ldots, n)$ and the predictive RMSE is evaluated on the test data. Finally, for our supervised learning algorithms, we use XGBoost (Chen and Guestrin, 2016) with the default hyperparameters in the R package (Chen et al., 2015).

The predictive RMSE for "Local", "Global", and "GMM" methods in Figure 4.3 are 18.94, 18.83, and 19.21, respectively. Also, the oracle predictive RMSE, which is the RMSE from predictions that use the true subspaces, is 18.64 and the naive predictive RMSE, which is the RMSE from predictions that use the original responses $\mathbf{Y}$ is 22.75. We see that incorporating the inner envelope structure into supervised learning methods reduces out-of-sample predictive RMSE compared to the naive method that directly use the original outcome. Also, the locally efficient and globally efficient methods have a predictive RMSE that is close to the oracle method, and broadly speaking, the predictive performance roughly follows the performance of estimating the subspaces in Figure 4.3.

4.6.3 Synthetic dataset based on the iris data

Our third simulation study mirrors is based on the the classic *iris* dataset by Fisher Fisher (1936), which has been used by other envelope-based method (Su and Cook,

2012). Briefly, the data contains 150 samples of iris species (setosa, versicolor, and virginica) along with their flower characteristics (sepal length, sepal width, and petal length). We take the species, dichotomized to two dummy variables, as predictors We also standardize the flower characteristics to mean zero and one standard deviation. Also, as a sanity check, we added two, random artificial responses $(Z_1, Z_2) \overset{i.i.d}{\sim} N(\mathbf{0}, \mathbf{I}_2)$ to the original set of responses. Our algorithms should identify these two responses as part of the subspace $\mathcal{S}_3$. In total, we have six responses $\mathbf{Y} \in \mathbb{R}^6$ and two predictors $\mathbf{X} \in \mathbb{R}^2$.

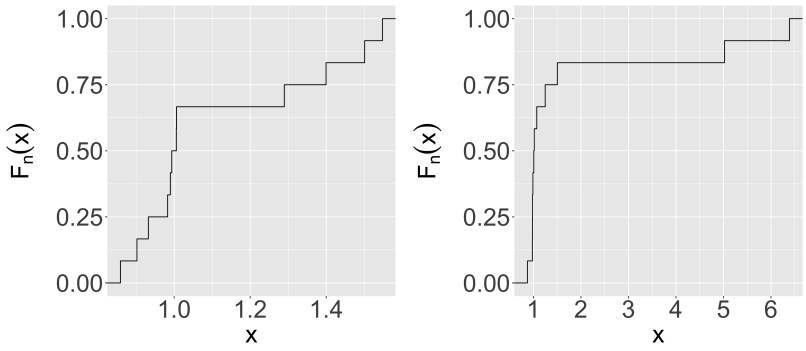
We first fit a multivariate linear regression of $\mathbf{Y}$ against $\mathbf{X}$. We also conduct a Shapiro-Wilk normality test for the error terms based on the linear model. The test suggests that the responses are non-normal, with p-value less than 10^{-12}, and the original inner envelope method may be inappropriate in this setting.

Next, we run our globally efficient estimator using the selected dimension of $u = \dim(\mathcal{S}_1) = 1$ and $d = \dim(\mathcal{S}_3) = 4$ from the nonparametric bootstrap method; note that the original envelope method selected $u = 4$ as the dimension, suggesting that the envelope does not discover structure beyond the two noise responses we added artifically. Also, we verified whether the estimated $\mathcal{S}_3$ contains the subspace spanned by the two artificially added random responses. At a high level, this verification involves checking the distances between the estimated S_3 and the space spanned by the two responses and we found that our estimated $\mathcal{S}_3$ contains the subspace; see the Appendix for details.

Next, we compare the estimated regression parameter $\boldsymbol{\beta}$ under different methods. We also conduct nonparametric bootstrap 100 times to obtain estimates of the standard errors of $\hat{\boldsymbol{\beta}}_{ij,\text{ols}}$, $\hat{\boldsymbol{\beta}}_{ij,\text{env}}$, $\hat{\boldsymbol{\beta}}_{ij,\text{InnEnv}}$, $\hat{\boldsymbol{\beta}}_{ij,\text{global}}$ and $\hat{\boldsymbol{\beta}}_{ij,\text{local}}$, where we posit the normal densities for the locally efficient estimator. The point estimates, bootstrap standard errors and p-values are given in Table C.5 in the Supplements. On average, the mean ratio of the standard error of the globally efficient estimator over that of the OLS estimator is 1.83. The mean ratios of the standard errors for the locally efficient estimator and the original inner envelope estimator compared to that of the OLS estimator are 1.70 and 1.76. Finally, the mean ratio of the standard error of the envelope estimator over that of the OLS estimator is 1.12. Figure 4.4 visualizes these results by comparing the empirical cumulative distribution functions (ECDF) of $\text{sd}(\hat{\boldsymbol{\beta}}_{ij,\text{ols}})/\text{sd}(\hat{\boldsymbol{\beta}}_{ij,\text{env}})$ and $\text{sd}(\hat{\boldsymbol{\beta}}_{ij,\text{ols}})/\text{sd}(\hat{\boldsymbol{\beta}}_{ij,\text{global}})$ for each element of $\boldsymbol{\beta}$ matrix. Roughly speaking, these results imply that to achieve the same power to test the null hypothesis that β_{ij} is zero versus a fixed, alternative hypothesis, a Wald test based on the globally efficient estimator only requires about half of the original

sample size (54.6%) compared to the Wald test based on the OLS estimator. The Appendix contains additional results from the analysis.

Figure 4.4: The empirical cumulative distribution function of $\text{sd}(\hat{\boldsymbol{\beta}}_{ij,\text{ols}})/\text{sd}(\hat{\boldsymbol{\beta}}_{ij,\text{env}})$ (left plot) and $\text{sd}(\hat{\boldsymbol{\beta}}_{ij,\text{ols}})/\text{sd}(\hat{\boldsymbol{\beta}}_{ij,\text{global}})$ (right plot) for each element of $\boldsymbol{\beta}$ matrix from the *iris* dataset.



(a) Envelope (b) Inner envelope

4.7 Real Data Analysis

Cardiovascular disease is a major cause of morbidity and mortality in patients with type 2 diabetes. The Action to Control Cardiovascular Risk in Diabetes (ACCORD) study in 2008 aimed to determine if the rate of cardiovascular disease (CVD) can be reduced in people with type 2 diabetes using intensive glycemic control, intensive blood pressure control, and multiple lipid management. Specifically, the ACCORD study randomized patients to either intensive glycemic control (HbA1c ¡6%) or standard glycemic control (HbA1c 7-7.9%). Participant were between the age of 40 to 82 who have been involved in the study for 2 to 7 years. Apart from diabetes, they also had a high risk of heart attack and stroke where each participant either has at least two risk factors for CVD and diabetes or has been diagnosed with CVD before the start of the study.

For our analysis, we are interested in investigating whether intensive glycemic control is associated with better outcomes after adjusting for baseline covariates. There are in total 6766 observations, 9 responses $\mathbf{Y} \in \mathbb{R}^9$ measuring the efficacy of intensive glycemic contorl, and 3 regressors $\mathbf{X} \in \mathbb{R}^3$. The response variables are treatment satisfaction, depression scale, aggregate physical activity score, aggregate mental score, symptom and distress score, systolic blood pressure (SBP), diastolic

blood pressure, and heart rate. Similar to the simulation studies, we standardized the responses. The predictors are body weight, age and the treatment indicator.

We first build a multivariate linear model to check if the residuals follows a normal distribution using the Shapiro–Wilk normality test. The p-values for each of the responses are significant under both tests, which suggests the normality assumption is violated. Next, we apply our proposed methods and the original envelope method. We remark that the envelope method chose a dimension of $u = 9$, which equals the total number of responses and the envelope method may have limited practical value. In contrast, our method found a non-trivial inner envelope structure with dimensions $u = \dim(\mathcal{S}_1) = 2$, and $d = \dim(\mathcal{S}_3) = 3$.

Once the dimension are selected, we assess the out-of-sample RMSE where we split 80% of the dataset into training data and 20% into test data, with 5416 and 1350 observations. Like Section 4.6, We calculate the predictive RMSE of the test data using different machine learning algorithms where the responses are either the original responses or the responses utilizing the inner envelope structure. The results are shown in Table 4.1.

	XGBoost	Random Forest	Linear
Using inner envelope structure	0.9918	0.9854	0.9834
Using original response	0.9949	0.9856	0.9835

Table 4.1: Predictive root-mean squared error (RMSE) for the test data in the ACCORD study

Comparing the rows of Table 4.1, the predictive RMSE is slightly smaller for all the methods if we use the inner envelope structure. Also, comparing the columns of Table 4.1, the linear model seems to have a better predictive performance compared to more complex methods. Given this, we compare the estimation of the regression parameter $\boldsymbol{\beta}$ supposing that the underlying model is linear. Specifically, we use the nonparametric bootstrap to obtain estimates of the standard errors of the OLS estimator and our globally efficient semiparametric inner envelope estimator of $\boldsymbol{\beta}$. The point estimates, bootstrap standard errors and p-values for the parameters in a linear model corresponding to the treatment are given in Table 4.2.

Our method found that the treatment had a significant effect on mental score, SBP and heart rate. In contrast, OLS found that the treatment was not significantly associated with these responses. This may be because our method is more efficient than the OLS, leading to more power to reject the null hypothesis of no effect.

Table 4.2: The point estimates, bootstrap standard errors and p-values for the regression parameter corresponding treatment for the ACCORD study. Asterisks correspond to p-values that are less than or equal to 0.05.

	Our Method			OLS		
Corresponding to Treatment	$\hat{\boldsymbol{\beta}}$	$\hat{\text{SE}}$	p-value	$\hat{\boldsymbol{\beta}}$	$\hat{\text{SE}}$	p-value
Treatment Satisfaction	0.029	0.023	0.22	0.027	0.023	0.24
Depression Scale	0.054	0.026	0.04*	0.054	0.026	0.04*
Physical Score	0.016	0.024	0.50	0.014	0.024	0.55
Mental Score	-0.045	0.020	0.03*	-0.043	0.026	0.10
Interference Score	-0.011	0.015	0.46	-0.009	0.027	0.72
Symptom & Distress Score	0.039	0.024	0.11	0.039	0.025	0.13
SBP	-0.012	0.005	0.02*	-0.034	0.022	0.12
DBP	0.004	0.024	0.86	0.002	0.025	0.94
Heart Rate	0.039	0.020	0.05*	0.038	0.023	0.11

Indeed, the estimated standard errors of our method are generally smaller compared to that from the OLS estimator. For example, the mean ratio of the standard error using our method over that using the OLS estimator is 1.55.

4.8 Summary and Discussion

In this paper, we proposed a semiparametric approach to the inner envelope, a dimension reduction method proposed by Su and Cook (2012) for the linear multivariate regression models. We derived the orthogonal nuisance tangent space, score function and efficient scores to estimate the inner envelope, all without having to make parametric modeling assumptions between the response, the covariates, and or the error terms. We also prooposed a simple GMM estimator from a set of estimating equations based on moment conditions.

We briefly take a moment to highlight some key limitations of our work. First, we assume throughout the theoretical results that the dimension of the inner envelope is fixed, even though in practice it is estimated from data. While we have shown that our bootstrap-based approach can reliably select the dimension numerically, we leave it as future work to derive the statistical properties of the estimated inner envelope with an estimated dimension. Second, we have not explored joint dimension reductions of both the responses and the predictors, which may bring further efficiency gains on the parameters of interest.

4.9 Proofs

4.9.1 Proof of Lemma 4.3.1

Take any matrix $\mathbf{A}_1 \in \mathbb{R}^{k\times\tau}$ that is full column rank and spans a τ-dimensional subspace $\mathcal{A}$ in a k-dimensional space. We can represent $\mathbf{A}_1$ by $\mathbf{A}_1 = \left(\mathbf{A}_u^T, \mathbf{A}_l^T\right)^T$, where $\mathbf{A}_u \in \mathbb{R}^{\tau\times\tau}$ and $\mathbf{A}_l \in \mathbb{R}^{(k-\tau)\times\tau}$. Without loss of generality, we can assume that $\mathbf{A}_u$ is invertible. Then, $\mathcal{A}$ is uniquely represented by $\mathbf{A}_{\text{repr}} = \mathbf{A}_1\mathbf{A}_u^{-1} = \left\{\mathbf{I}_u, (\mathbf{A}_l\mathbf{A}_u^{-1})^T\right\}^T$, and the lower $(p-u)\times u$ submatrix $\mathbf{A}_l\mathbf{A}_u^{-1}$ uniquely parameterizes $\mathcal{A}$. Let $\mathbf{a} = \text{vec}(\mathbf{A}_l\mathbf{A}_u^{-1})$ denote the vector concatenation of the lower part of $\mathbf{A}_{\text{repr}}$. Since the mapping between $\mathbf{A}_{repr}$ and $\mathcal{A}$ is one-to-one, there exists a one-to-one mapping ψ_1 such that $\mathbf{a} = \psi_1(\mathcal{A})$.

Because $\mathbf{A}_{\text{repr}}$ is not a semi-orthogonal matrix, we use Gram-Schmidt procedure to obtain a unique semi-orthogonal matrix $\mathbf{A}$ from $\mathbf{A}_{\text{repr}}$. Hence, there exists $\tilde{\psi}$ such that $\mathbf{A} = \psi_2(\mathcal{A})$. One can show that ψ_2 is also a one-to-one mapping. The relationship between $\mathbf{a}$, $\mathbf{A}_{\text{repr}}$ and $\mathbf{A}$ are shown in the Figure 4.5. Notice that if we decompose $\mathbf{A}$ as $\mathbf{A} = (\mathbf{A}_{\text{up}}^T, \mathbf{A}_{\text{low}}^T)^T$ as before, we also have $\mathbf{A}_{\text{repr}} = \mathbf{A}\mathbf{A}_{\text{up}}^{-1}$. Therefore, we have unique representations of the space $\mathcal{A}$ by a variational independent parameter $\mathbf{a}$ and by an orthogonal matrix $\mathbf{A}$.

$$\mathbf{a} = \begin{pmatrix} a_{\tau+1,1} \\ \vdots \\ a_{k,1} \\ \vdots \\ a_{\tau+1,\tau} \\ \vdots \\ a_{k,\tau} \end{pmatrix} \underset{\text{Vectorize lower submatrix}}{\overset{\text{Stack } \mathbf{a} \text{ and concatenate } \mathbf{I}_\tau}{\rightleftarrows}} \mathbf{A}_{\text{repr}} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \\ a_{\tau+1,1} & \cdots & a_{\tau+1,\tau} \\ \vdots & \ddots & \vdots \\ a_{k,1} & \cdots & a_{k,\tau} \end{pmatrix} \underset{\mathbf{A}_{\text{repr}}=\mathbf{A}\mathbf{A}_{\text{up}}^{-1}}{\overset{\text{Gram-Schmidt}}{\rightleftarrows}} \mathbf{A} = \begin{pmatrix} \mathbf{A}_{\text{up}} \\ \mathbf{A}_{\text{low}} \end{pmatrix}$$

Figure 4.5: Unique parameterization of any space $\mathcal{A}$.

4.9.2 Proof of Theorem 4.4.2

Proof. We prove the consistency and asymptotic normality by checking conditions in Theorem 2.1 (Lemma 6 in the Supplements) and 3.1 (Lemma 7 in the Supplements) from Newey and McFadden (1994), and the efficiency by checking the asymptotic variance achieves the semiparametric efficiency bound. We write

$X_n = o_p(1)$ if $X_n \xrightarrow{p} 0$ and $X_n = O_p(1)$ if for all $\epsilon > 0$ there exists M such that $\sup_n \mathrm{P}(\|X_n > M\|) < \epsilon$. Throughout the proof of this theorem, we let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$.

We firstly prove the $\hat{\boldsymbol{\theta}}$ obtained by solving equation (4.8) satisfies $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

We prove by checking conditions of Theorem 2.1 in Newey and McFadden (1994). Consider the following two functions:

$$Q_0(\boldsymbol{\theta}) = \frac{1}{2}\left\{\frac{1}{n}\sum_{i=1}^{n} S^*_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta^*_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})\right\}^2$$

$$\hat{Q}_n(\boldsymbol{\theta}) = \frac{1}{2}\left\{\frac{1}{n}\sum_{i=1}^{n} \hat{S}^*_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \hat{\eta}^*_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta})\right\}^2.$$

Let $\hat{\boldsymbol{\theta}}_0$ be a minimizer of $Q_0(\boldsymbol{\theta})$. By regularity condition (B2), $\hat{\boldsymbol{\theta}}_0$ is unique and $\hat{\boldsymbol{\theta}}_0 \xrightarrow{p} \boldsymbol{\theta}_0$. Also, $\hat{\boldsymbol{\theta}}$ is the minimizer of $\hat{Q}_n(\boldsymbol{\theta})$. Because the parameter space $\boldsymbol{\Theta}$ is compact and $Q_0(\boldsymbol{\theta})$ is continuous, in order to apply Theorem 2.1 in Newey and McFadden (1994), we only need to show condition (iv) holds. That is, $\hat{Q}_n(\boldsymbol{\theta})$ converges uniformly in probability to $Q_0(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. By Lemma 2.8 in Newey and McFadden (1994), we only need to show (1) $\hat{Q}_n(\boldsymbol{\theta}) \xrightarrow{p} Q_0(\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$; and (2) $\hat{Q}_n(\boldsymbol{\theta})$ is stochastic equicontinuous. By the continuous mapping theorem, we only need to show

$$\hat{S}^*_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta^*_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta}) \xrightarrow{p} S^*_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta^*_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta}).$$

Since $\hat{S}^*_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta^*_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2)$ is a continuous function of $\hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2, \hat{\mathbb{E}}\{\partial \log \eta^*_1/\partial(\boldsymbol{\Gamma}^T\mathbf{Y}_i)^T \mid \mathbf{X}_i\}, \hat{\mathbb{E}}\{\partial \log \eta^*_2/\partial(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i, \mathbf{X}_i\}$ and $\hat{\mathbb{E}}\{\partial \log \eta^*_3/\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T\}$; and by the properties of nonparametric regression, $\hat{\boldsymbol{\Delta}}_1 \xrightarrow{p} \boldsymbol{\Delta}_1$, $\hat{\boldsymbol{\Delta}}_2(\boldsymbol{\theta}) \xrightarrow{p} \boldsymbol{\Delta}_2(\boldsymbol{\theta})$, where $\hat{\boldsymbol{\Delta}}_{i1} = \mathbf{Y}_i - \hat{\mathbb{E}}(\mathbf{Y}_i \mid \mathbf{X}_i)$, $\hat{\boldsymbol{\Delta}}_{i2}(\boldsymbol{\theta}) = \mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}}\{\hat{\mathbb{E}}(\mathbf{Y}_i \mid \mathbf{B}_0\boldsymbol{\Gamma}_0^T\mathbf{Y}_i, \mathbf{X}_i) - \hat{\mathbb{E}}(\mathbf{Y}_i \mid \mathbf{B}_0\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)\}$, and

$$\begin{aligned}
&\hat{\mathbb{E}}\left\{\frac{\partial \log \eta^*_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y}_i)^T} \,\middle|\, \mathbf{X}_i\right\} \xrightarrow{p} \mathbb{E}\left\{\frac{\partial \log \eta^*_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y}_i)^T} \,\middle|\, \mathbf{X}_i\right\},\\
&\hat{\mathbb{E}}\left\{\frac{\partial \log \eta^*_2}{\partial(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T} \,\middle|\, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i, \mathbf{X}_i\right\} \xrightarrow{p} \mathbb{E}\left\{\frac{\partial \log \eta^*_2}{\partial(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T} \,\middle|\, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i, \mathbf{X}_i\right\},\\
&\hat{\mathbb{E}}\left\{\frac{\partial \log \eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right\} \xrightarrow{p} \mathbb{E}\left\{\frac{\partial \log \eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right\};
\end{aligned}$$

by continuous mapping theorem,

$$\hat{S}^*_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta^*_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta}) \xrightarrow{p} S^*_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta^*_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})$$

for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. To prove the stochastic equicontinuity of $\hat{Q}_n(\boldsymbol{\theta})$, by Lemma 2.9 of Newey and McFadden (1994), we only need to show $\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$, there exists $B_n = O_p(1)$ such that $|\hat{Q}_n(\boldsymbol{\theta}_1) - \hat{Q}_n(\boldsymbol{\theta}_2)| \leq B_n \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$. By regularity conditions (A5) and (B1), $\hat{Q}_n(\boldsymbol{\theta})$ is differentiable. By mean value theorem, there exists $\bar{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ such that $|\hat{Q}_n(\boldsymbol{\theta}_1) - \hat{Q}_n(\boldsymbol{\theta}_2)| \leq \|\hat{Q}'_n(\bar{\boldsymbol{\theta}})\| \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$. By Theorem 2.9 in Li and Racine (2007) and regularity condition (B3), $\partial \hat{\boldsymbol{\Delta}}_2(\boldsymbol{\theta})/\partial \boldsymbol{\theta} - \partial \boldsymbol{\Delta}_2(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = O_p(n^{1/2} h^{1+(p+r-u-d)/2}) = o_p(1)$. Because $\log \eta^*_{1,2,3}$ are twice differentiable, we have

$$\begin{aligned}
&\hat{\mathbb{E}}\left\{ \frac{\partial^2 \log \eta^*_1}{\partial \boldsymbol{\theta} \partial (\boldsymbol{\Gamma}^T \mathbf{Y}_i)^T} \;\middle|\; \mathbf{X}_i \right\} \xrightarrow{p} \mathbb{E}\left\{ \frac{\partial^2 \log \eta^*_1}{\partial \boldsymbol{\theta} \partial (\boldsymbol{\Gamma}^T \mathbf{Y}_i)^T} \;\middle|\; \mathbf{X}_i \right\}, \\
&\hat{\mathbb{E}}\left\{ \frac{\partial^2 \log \eta^*_2}{\partial \boldsymbol{\theta} \partial (\boldsymbol{\Gamma}_0^T \mathbf{Y}_i)^T} \;\middle|\; \mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y}_i, \mathbf{X}_i \right\} \xrightarrow{p} \mathbb{E}\left\{ \frac{\partial^2 \log \eta^*_2}{\partial \boldsymbol{\theta} \partial (\boldsymbol{\Gamma}_0^T \mathbf{Y}_i)^T} \;\middle|\; \mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y}_i, \mathbf{X}_i \right\}, \\
&\hat{\mathbb{E}}\left\{ \frac{\partial^2 \log \eta^*_3}{\partial \boldsymbol{\theta} \partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y}_i)^T} \right\} \xrightarrow{p} \mathbb{E}\left\{ \frac{\partial^2 \log \eta^*_3}{\partial \boldsymbol{\theta} \partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y}_i)^T} \right\}.
\end{aligned}$$

Therefore, by continuous mapping theorem, $\hat{Q}'_n(\boldsymbol{\theta}) \xrightarrow{p} Q'_0(\boldsymbol{\theta})$. Hence, $B_n = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{Q}'_n(\boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q'_0(\boldsymbol{\theta}) + \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |Q'_0(\boldsymbol{\theta}) - \hat{Q}'_n(\boldsymbol{\theta})|$. Since $Q'_0(\boldsymbol{\theta})$ is a continuous function in a compact set, $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q'_0(\boldsymbol{\theta})$ is $O_p(1)$. Suppose $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |Q'_0(\boldsymbol{\theta}) - \hat{Q}'_n(\boldsymbol{\theta})| = \infty$, then there exists a sequence $\{\boldsymbol{\theta}_k\}$ such that $|Q'_0(\boldsymbol{\theta}_k) - \hat{Q}'_n(\boldsymbol{\theta}_k)| \to \infty$. That is, for any $C > 0$, there exists $N \in \mathbb{N}^+$ such that for any $n \geq N$, $|Q'_0(\boldsymbol{\theta}_n) - \hat{Q}'_n(\boldsymbol{\theta}_n)| > C$, which contradicts with the fact that $\hat{Q}'_n(\boldsymbol{\theta}_n) \xrightarrow{p} Q'_0(\boldsymbol{\theta}_n)$. Hence, $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |Q'_0(\boldsymbol{\theta}) - \hat{Q}'_n(\boldsymbol{\theta})| = o_p(1)$. Therefore, all the conditions in Theorem 2.1 from Newey and McFadden (1994) hold, and we have $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

Then, in order to show the asymptotic normality of $\hat{\boldsymbol{\theta}}$, we only need to verify conditions (i)–(v) in Theorem 3.1 from Newey and McFadden (1994). Conditions (i)–(ii) are already satisfied. We then prove $\sqrt{n} \partial \hat{Q}_n(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}$ is asymptotically normal.

Because

$$\sqrt{n} \frac{\partial \hat{Q}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{S}^*_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \hat{\eta}^*_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta}_0) \cdot \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\theta}} \hat{S}^*_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \hat{\eta}^*_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta}_0),$$

and

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\theta}} \hat{S}^*_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \hat{\eta}^*_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta}_0) \xrightarrow{p} \mathbb{E}\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} S^*_{\text{eff}}(\mathbf{Y}, \mathbf{X}, \eta^*_1, \eta^*_2, \eta^*_3; \boldsymbol{\theta}_0) \right\},$$

by Slutsky's Theorem, we only need to show $n^{-1/2}\sum_{i=1}^{n}\hat{S}^*_{\text{eff}}(\mathbf{Y}_i,\mathbf{X}_i,\hat{\eta}^*_{1,2,3},\hat{\boldsymbol{\Delta}}_1,\hat{\boldsymbol{\Delta}}_2;\boldsymbol{\theta}_0)$ converges to a normal distribution. Also, because

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}S^*_{\text{eff}}(\mathbf{Y}_i,\mathbf{X}_i,\eta^*_{1,2,3},\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2;\boldsymbol{\theta}_0)\xrightarrow{d}N\Big[\mathbf{0},\mathbb{E}\big\{S^*_{\text{eff}}(\mathbf{Y}_i,\mathbf{X}_i,\eta^*_1,\eta^*_2,\eta^*_3;\boldsymbol{\theta}_0)^{\otimes 2}\big\}\Big],$$

we only need to show

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{S}^*_{\text{eff}}(\mathbf{Y}_i,\mathbf{X}_i,\hat{\eta}^*_{1,2,3},\hat{\boldsymbol{\Delta}}_1,\hat{\boldsymbol{\Delta}}_2;\boldsymbol{\theta}_0)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}S^*_{\text{eff}}(\mathbf{Y}_i,\mathbf{X}_i,\eta^*_{1,2,3},\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2;\boldsymbol{\theta}_0)\xrightarrow{p}0.$$

The score function $S^*_{\text{eff}}=(S^*_{\text{eff},\boldsymbol{\gamma}},S^*_{\text{eff},\mathbf{b}})$ has two components. For simplicity, we only show the convergence of the second component $S^*_{\text{eff},\mathbf{b}}$. The convergence of the first component can be proved using the same technique. Let $\boldsymbol{\Gamma}$ and $\mathbf{B}$ denote the orthogonal basis derived from $\boldsymbol{\theta}_0$. Then,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{S}^*_{\text{eff},\mathbf{b}}(\mathbf{Y}_i,\mathbf{X}_i,\hat{\eta}^*_{1,2,3},\hat{\boldsymbol{\Delta}}_1,\hat{\boldsymbol{\Delta}}_2;\boldsymbol{\theta}_0)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}S^*_{\text{eff},\mathbf{b}}(\mathbf{Y}_i,\mathbf{X}_i,\eta^*_{1,2,3},\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2;\boldsymbol{\theta}_0)\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\text{vec}^T\left[\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Delta}}_{i2}(\boldsymbol{\theta}_0)\left\{\frac{\partial\log\eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}-\hat{\mathbb{E}}\left(\frac{\partial\log\eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)\right\}\right]\frac{\partial\text{vec}(\mathbf{B}_0)}{\partial\mathbf{b}^T}\\
&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\text{vec}^T\left[\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)\left\{\frac{\partial\log\eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}-\mathbb{E}\left(\frac{\partial\log\eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)\right\}\right]\frac{\partial\text{vec}(\mathbf{B}_0)}{\partial\mathbf{b}^T}\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\text{vec}^T\left[\boldsymbol{\Gamma}_0^T\left\{\hat{\boldsymbol{\Delta}}_{i2}(\boldsymbol{\theta}_0)-\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)\right\}\left\{\frac{\partial\log\eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}-\mathbb{E}\left(\frac{\partial\log\eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)\right\}\right]\frac{\partial\text{vec}(\mathbf{B}_0)}{\partial\mathbf{b}^T}\\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\text{vec}^T\left[\boldsymbol{\Gamma}_0^T\left\{\hat{\boldsymbol{\Delta}}_{i2}(\boldsymbol{\theta}_0)-\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)\right\}\left\{\mathbb{E}\left(\frac{\partial\log\eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)-\hat{\mathbb{E}}\left(\frac{\partial\log\eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)\right\}\right]\frac{\partial\text{vec}(\mathbf{B}_0)}{\partial\mathbf{b}^T}\\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\text{vec}^T\left[\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)\left\{\mathbb{E}\left(\frac{\partial\log\eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)-\hat{\mathbb{E}}\left(\frac{\partial\log\eta^*_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)\right\}\right]\frac{\partial\text{vec}(\mathbf{B}_0)}{\partial\mathbf{b}^T}
\end{aligned}$$

By Lemma 5, the first term can be bounded by $O_p(h^2+n^{-1/2}h^{p+r-u-d}\log n+n^{1/2}h^4)$. Under regularity condition (B3), the second term is $o_p(1)$.

By Theorem 2.6 in Li and Racine (2007), under regularity condition (A4),

$$\sup_{\mathbf{X}_i,\mathbf{Y}_i}|\hat{\boldsymbol{\Delta}}_{i2}(\boldsymbol{\theta}_0)-\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)|=O_p\left\{\left(\frac{\log n}{nh^{p+r-u-d}}\right)^{1/2}+h^2\right\}$$

is $o_p(1)$ under regularity condition (B3).

By the central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{vec}^T\left[\left\{\hat{\boldsymbol{\Delta}}_{i2}(\boldsymbol{\theta}_0)-\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)\right\}\left\{\mathbb{E}\left(\frac{\partial\log\eta_3^*}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)-\hat{\mathbb{E}}\left(\frac{\partial\log\eta_3^*}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)\right\}\right]\frac{\partial\mathrm{vec}(\mathbf{B}_0)}{\partial\mathbf{b}^T}$$
$$=\mathrm{vec}^T\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\boldsymbol{\Delta}}_{i2}(\boldsymbol{\theta}_0)-\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)\right\}\cdot\sqrt{n}\left\{\mathbb{E}\left(\frac{\partial\log\eta_3^*}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)-\hat{\mathbb{E}}\left(\frac{\partial\log\eta_3^*}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)\right\}\right]\frac{\partial\mathrm{vec}(\mathbf{B}_0)}{\partial\mathbf{b}^T}$$
$$=o_p(1)\cdot O_p(1)=o_p(1).$$

Hence, the second term is also $o_p(1)$.

Also,

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)\left\{\mathbb{E}\left(\frac{\partial\log\eta_3^*}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)-\hat{\mathbb{E}}\left(\frac{\partial\log\eta_3^*}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)\right\}\\ =&o_p(1)\cdot n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)\\ =&o_p(1),\end{aligned}$$

where the last equation is because $\boldsymbol{\Delta}_{i2}$ are i.i.d. mean $\mathbf{0}$ random variables. Hence, the third term is also $o_p(1)$.

Therefore,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{S}^*_{\mathrm{eff},\mathbf{b}}(\mathbf{Y}_i,\mathbf{X}_i,\hat{\eta}^*_{1,2,3},\hat{\boldsymbol{\Delta}}_1,\hat{\boldsymbol{\Delta}}_2;\boldsymbol{\theta}_0)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}S^*_{\mathrm{eff},\mathbf{b}}(\mathbf{Y}_i,\mathbf{X}_i,\eta^*_{1,2,3},\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2;\boldsymbol{\theta}_0)\xrightarrow{p}0.$$

Hence, by Slutsky's Theorem,

$$\sqrt{n}\frac{\partial\hat{Q}_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\xrightarrow{d}N\{\mathbf{0},\mathbf{C}_2^T\mathbf{D}_2\mathbf{C}_2\},$$

where

$$\mathbf{C}_2=\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}S^*_{\mathrm{eff}}(\mathbf{Y},\mathbf{X},\eta_1^*,\eta_2^*,\eta_3^*;\boldsymbol{\theta}_0)\right\},\quad\mathbf{D}_2=\mathbb{E}\left\{S^*_{\mathrm{eff}}(\mathbf{Y}_i,\mathbf{X}_i,\eta_1^*,\eta_2^*,\eta_3^*;\boldsymbol{\theta}_0)^{\otimes2}\right\}.$$

Next, we verify the conditions (iv)–(v) that are related to $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\hat{Q}_n(\boldsymbol{\theta})$. Notice that

$$\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\hat{Q}_n(\boldsymbol{\theta})\xrightarrow{p}\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}S^*_{\mathrm{eff}}(\mathbf{Y}_i,\mathbf{X}_i,\eta^*_{1,2,3},\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2;\boldsymbol{\theta})\right\}\cdot\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\theta}^T}S^*_{\mathrm{eff}}(\mathbf{Y}_i,\mathbf{X}_i,\eta^*_{1,2,3},\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2;\boldsymbol{\theta})\right\}$$
$$+\mathbb{E}\left\{S^*_{\mathrm{eff}}(\mathbf{Y}_i,\mathbf{X}_i,\eta^*_{1,2,3},\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2;\boldsymbol{\theta})\right\}\cdot\mathbb{E}\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}S^*_{\mathrm{eff}}(\mathbf{Y}_i,\mathbf{X}_i,\eta^*_{1,2,3},\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2;\boldsymbol{\theta})\right\}=H(\boldsymbol{\theta}).$$

Following the same argument as proving the uniform convergence in probability for $\partial\hat{Q}_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, because the third order derivative of $\log \eta^*_{1,2,3}$ exists, we have $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\hat{Q}_n(\boldsymbol{\theta}) \xrightarrow{p} H(\boldsymbol{\theta})$ uniformly for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Hence, condition (iv) holds. Because

$$\frac{1}{n}\sum_{i=1}^{n} S^*_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta^*_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{0},$$

we have $H(\boldsymbol{\theta}_0) = \mathbf{C}_2\mathbf{C}_2^T$. Since $\mathbf{C}_2$ is nonsingular, $H(\boldsymbol{\theta}_0)$ is nonsingular. Hence, condition (v) holds. Therefore, by Theorem 3.1 in Newey and McFadden (1994),

$$\sqrt{n}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) \xrightarrow{p} N\{\mathbf{0}, H(\boldsymbol{\theta}_0)^{-1}\mathbf{C}_2^T\mathbf{D}_2\mathbf{C}_2H(\boldsymbol{\theta}_0)^{-1}\} = N\{\mathbf{0}, \mathbf{C}_2^{-1}\mathbf{D}_2(\mathbf{C}_2^T)^{-1}\}.$$

■

Chapter 5

Improving Instrumental Variable Estimation by Removing Redundant Instruments

In observational studies, a primary challenge in estimating treatment effects is the presence of unmeasured confounding, that is, the outcomes between treatment groups may differ, not only because of the treatment effect but also because of unmeasured factors that may affect the treatment selection. For example, in the classic example of the effect of schooling on earnings (Angrist and Keueger, 1991), a comparison between earnings among people with different schooling may be biased because pre-schooling levels of ability such as Intelligence Quotient (IQ), affects both schooling choices and earnings given education levels. Instrumental variables (IV) are commonly used to recover the effect of exposure in the presence of unmeasured confounding between exposure and outcome. A valid IV is defined to satisfy 3 key conditions: (a) it is associated with treatment; (b) it is independent of any unmeasured confounder of the exposure-outcome relationship, and (c) it has no direct effect on the outcome. With a valid IV, one can extract exogenous variation in the treatment that is unconfounded with the outcome, and to leverage this bias-free component to make a causal inference about the treatment effect (Robins, 1989; Angrist et al., 1996; Heckman, 1997).

The development of the IV approach can be traced back to Wright (1928) and Goldberger (1972) under linear structural equations in econometrics. Various parametric and semiparametric IV estimators have been developed (Robins, 1989, 1994; Imbens and Angrist, 1994; Angrist et al., 1996; Vansteelandt and Goetghebeur, 2003;

Robins and Rotnitzky, 2004; Hernán and Robins, 2006; Tan, 2010; Sun et al., 2017; Matsouaka and Tchetgen Tchetgen, 2017; Liu et al., 2017). Most of the existing IV methods prefers the IV to have a strong correlation with the exposure. For example, it is well recognized that in finite samples, the two-stage least squares (2SLS) estimator is biased and the bias increases as the association between IV and exposure decreases (Bound et al., 1995; Staiger and Stock, 1997; Imbens and Rosenbaum, 2005; Small and Rosenbaum, 2008). An IV that is strongly associated with exposure is hard to find. For example, in the aforementioned study of Angrist and Keueger (1991), they interact the birth quarter with 50 state and 9 year-of-birth dummies. In a followup work, Chamberlain and Imbens (2004) consider three-way interaction of quarter, year and state of birth to get 504 instruments. However, the majority of those constructed instruments are only weakly related to or independent of the years of schooling.

Rather than finding a new IV that extract more unconfounded components in the exposure, we explore the alternative option to reduce the redundant IVs[1]. An instrumental variable is weak if its correlation with the exposure is small, while the size of being "small" depends on the inference problem at hand, and on the sample size (Andrews and Stock, 2018; Bound et al., 1995). Thus, a natural description of redundant IVs is that those are not correlated with exposure. Such a characterization motivated the literature of IV selection. For example, Donald and Newey (2001) derived a mean-square error criteria that can be minimized to choose the instrument set and Hall and Peixe (2003) improved this method in terms of the finite sample performance. All these IV selection methods are selecting IVs coordinate-wisely.

As a start, we use a sufficient dimension reduction method (SDR) to select linear combinations of IV in the commonly used two stage least square (2SLS) and indirect least square (ILS) regression models. Such a selection is more general than the coordinate-wise selection because all IVs may be correlated with exposure but there may exist some linear combination of IVs that has a null correlation with the exposure. Specifically, we use the envelope method proposed by (Cook et al., 2013). By assuming that there exists some combinations of the variables that is independent of the regression of interest, the envelope method can estimate the material and immaterial information and thus improve the estimation of the parameter of interest. Specifically, it may achieve substantial efficiency gains when the variables have high correlations. Ever since the first paper on the envelope

[1]Another strand of work is to invert tests to derive conservative bounds for the causal effect (Anderson and Rubin, 1949; Kleibergen, 2002, 2005; Moreira, 2003)

method, an increasing number of research has been carried out to generalize envelope in various settings (Cook et al., 2010; Su and Cook, 2011, 2012; Cook and Su, 2013; Cook et al., 2013, 2015; Cook and Zhang, 2015a,b; Su et al., 2016; Li and Zhang, 2017) including response envelope, partial envelope, inner envelope, scaled envelope, predictor envelope, reduced rank envelope, simultaneous envelope, model-free envelope, sparse envelope and tensor envelope. Various algorithms (Cook et al., 2016; Cook and Zhang, 2016, 2018) have been proposed to effectively and efficiently estimate the envelope models. A review of the envelope methods can be found in Cook (2018a).

We explore the reduction of IVs at the first stage of the 2SLS estimation and at the second stage regression of ILS procedure. Interestingly, we found that the first stage reduction only results in a reduction in finite sample bias but not in asymptotic efficiency gain. While the reduction at the second stage has the potential to achieve substantial efficiency gain. This motivates us to generalize the description of informative IVs not only to those that have an association with the exposure but also to those that affect the outcome through the exposure. With such a generalized description of useful IVs, we can eliminate more immaterial IVs and thus achieve potentially substantial asymptotic efficiency gain of the causal effect estimate.

This chapter has several contributions: Firstly, we considered a novel way to improve the IV estimation by removing redundant IVs using SDR. Such a method circumvents the search of a stronger IV and can be more efficient than the standard method. Secondly, we generalized the notion of redundant IV by considering the linear combination of some IVs and the correlation among IVs and that between IV and response. Thirdly, to the best of our knowledge, this is the first paper to consider the envelope methods in a model with endogeneity or unmeasured confounding.

The rest of the chapter is organized as follows. In Section 5.1, we briefly review the definition of a valid IV and several variations of the envelope method. In Section 5.2, we consider 2SLS with predictor envelope at the first stage and ILS with predictor envelope at the second stage. We carry out extensive simulations in Section 5.3. We illustrated our method in the Wisconsin Longitudinal Study to estimate the effect of education on wage in Section 5.4. We conclude with a discussion in Section 5.5. All proofs are given in the Appendix.

5.1 Preliminaries

5.1.1 Notations and Assumptions

Consider a population of n individuals, where individuals have identical and independent observations. For individual i, let $X_i \in \mathbb{R}^p$ denote a multivariate exposure, $Z_i \in \mathbb{R}^k$ denote a multivariate pre-exposure variable. For the identification purpose, we always assume $k \geq p$, that is, the number of IVs is no less than that of the exposure. Let $Y_i \in \mathbb{R}^r$ denote the observed multivariate response and let $U_i \in \mathbb{R}^l$ denote the unmeasured confounder between X_i and Y_i. We omit the observed covariates as our methods can easily be generalized. Additionally, we suppress the subscript i if referring to the variables of a general person, e.g., we use X_i to denote the exposure for a person. Let $X = (X_1, \ldots, X_n)^T$, and define Y, Z, and U similarly. Throughout this paper, we assume without loss of generality that all the variables have mean 0. For random vectors W_k, $n = 1, 2, \ldots$, in a normed space, we denote $X_k = O_p(Y_k)$ if for every $\epsilon > 0$, there is a $M > 0$ such that $P(\|X_k\|/\|Y_k\| > M) < \epsilon$. Also, we denote $X_k = o_p(Y_k)$ if X_k/Y_k converges to 0 in probability.

Formally, an IV must satisfy the following assumptions:

Assumption 5.1.1 (IV relevance) $Z_i \perp\!\!\!\perp X_i \mid U_i$;

Assumption 5.1.2 (IV unconfoundedness) $Z_i \perp\!\!\!\perp U_i$;

Assumption 5.1.3 (Exclusion restriction) $Z_i \perp\!\!\!\perp Y_i \mid X_i, U_i$.

Assumption 5.1.1 states that IV and exposure have a non-null association even if the association is not causal. Assumption 5.1.2 states that the IV is not affected by the exposure-outcome confounder U. Assumption 5.1.3 states that Z does not have a direct effect on the outcome Y. If Assumptions 5.1.1–5.1.3 are satisfied for all individuals, then Z is a valid IV.

We consider the linear model,

$$\begin{aligned} X_i &= \delta Z_i + \varepsilon_{X_i}, \\ Y_i &= \beta X_i + \varepsilon_{Y_i}, \end{aligned} \tag{5.1}$$

where $\varepsilon_{X_i} \in \mathbb{R}^p$ and $\varepsilon_{Y_i} \in \mathbb{R}^r$ are both functions of U_i, correlated and are mean zero. $Z_i \in \mathbb{R}^k$ is a valid IV. We further assume δ is full row rank so that δZ may contain enough information of X.

The 2SLS method is a standard method to parse the unmeasured confounding using instrumental variables. Specifically, the estimation is carried out in two stages:

the exposure X is projected onto the IV Z at the first stage, and the outcome Y is regressed on the predicted exposure $\widehat{X} = \widehat{\delta}_{2SLS} Z$ at the second stage, where $\widehat{\delta}_{2SLS} = (Z^T Z)^{-1} Z^T X$ is the least square estimate obtained at the first stage. A detailed algorithm is given in the Appendix. The idea of 2SLS is to project the confounded exposure X onto the space of unconfounded IV Z so that the confounding is removed. It can be shown that $\widehat{\beta}_{2SLS} = Y^T P_Z X (X^T P_Z X)^{-1}$ is consistent for β, where $P_Z = Z(Z^T Z)^{-1} Z^T$ is the projection matrix onto Z.

It is well known that the strength of an IV is important. Rather than finding a "stronger" IV that extracts more unconfounded components in the exposure, we explore the alternative option of reducing the IV redundancy. We generalize the envelope method (Cook et al., 2010) to the scenario with unmeasured confounding to achieve substantial efficiency gain of the causal effect estimate by eliminating immaterial IVs.

5.1.2 Review of the Envelope method

The envelope model was first proposed by Cook et al. (2010) for response reduction in the multivariate linear model. The overarching goal of the envelope model is to identify the material and immaterial information of regression and eliminate the immaterial part to increase efficiency in estimation and prediction. Here, we briefly review the predictor envelope proposed by Cook et al. (2013).

With a little abuse of notation, we express the linear regression model considered by Cook et al. (2010) as

$$Y_i = \beta X_i + \varepsilon_i, \tag{5.2}$$

where predictor X_i has variance $\Sigma_X \in \mathbb{S}^{p\times p}$, $\mathbb{S}^{p\times p}$ denotes the collection of all symmetric positive definite matrices of size p, and $X_i \perp\!\!\!\perp \varepsilon_i$. The response can either be univariate or multivariate. Unlike the second equation in Model (5.1), the error term $\boldsymbol{\varepsilon}_i$ and the predictor X_i in the standard linear regression model (5.2) are independent. Cook et al. (2013) assumes that there exists an orthogonal matrix (Φ, Φ_0) with $\Phi \in \mathbb{R}^{p\times u}$, $\Phi_0 \in \mathbb{R}^{p\times(p-u)}$ such that

Condition 5.1.1 $cov(Y_i, \Phi_0^T X_i \mid \Phi^T X_i) = 0$,

Condition 5.1.2 $cov(\Phi^T X_i, \Phi_0^T X_i) = 0$.

Condition 5.1.1 states that there is no linear relationship between Y_i and $\Phi_0^T X_i$ given $\Phi^T X_i$. Condition 5.1.2 ensures no marginal linear relationship between the

reduced predictor and its complement. The predictor envelope is defined as the smallest space of span(Φ) that satisfies Conditions 5.1.1–5.1.2. The dimension u of the envelope basis Φ is defined as the envelope dimension. The envelope dimension has the range $0 \leq u \leq p$ and it can be estimated using information-based criteria such as BIC, or cross-validation. Although the parameter Φ is not point-wise identifiable, the span of Φ is uniquely identified and is denoted as $\mathcal{E}_{\Sigma_X}\{\text{span}(\beta^T)\}$. Throughout this paper, we simplify the envelope notation as $\mathcal{E}_{\Sigma_X}(\beta^T)$. It has been shown that under Conditions 5.1.1–5.1.2, Model (5.2) can be written as

$$Y_i = \eta^T \Phi^T X_i + \varepsilon_i, \tag{5.3}$$

where X has variance $\Sigma_X = \Phi\Omega\Phi^T + \Phi_0\Omega_0\Phi_0^T$, where $\Omega = \Phi^T\Sigma_X\Phi$ and $\Omega_0 = \Phi_0^T\Sigma_X\Phi_0$.

The envelope $\mathcal{E}_{\Sigma_X}(\beta^T)$ can be estimated by minimizing the objective function:

$$\underset{\text{span}(G)\in\mathcal{G}(u,p)}{\arg\min} \log|G^T\widehat{\Sigma}_{X|Y}G| + \log|G^T\widehat{\Sigma}_X^{-1}G|. \tag{5.4}$$

Because Φ is a semi-orthogonal matrix, this is a Grassmann manifold optimization problem. Cook et al. (2016) re-parameterized this objective function to convert this problem to a non-Grassmann manifold one. Without loss of generality, Cook et al. (2016) assume the first u rows of Φ is non-singular, then we have $\Phi = (\Phi_1^T, \Phi_2^T)^T = (I_u, A^T)^T\Phi_1 = G_A\Phi_1$, where $A = \Phi_2\Phi_1^{-1}$ with no constraints and $G_A = (I_u, A^T)^T$. Then the objective function (5.4) can be converted to

$$\underset{A\in\mathbb{R}^{(k-u)\times u}}{\arg\min} -2\log|G_A^TG_A| + \log|G_A^T\widehat{\Sigma}_{X|Y}G_A| + \log|G_A^T\widehat{\Sigma}_X^{-1}G_A|, \tag{5.5}$$

which is a non-Grassmann manifold optimization problem. Cook et al. (2016) discussed an algorithm to solve this optimization problem. The envelope estimator can be defined as $\widehat{\beta}_{env} = P_{\widehat{\Phi}(S_X)}\widehat{\beta}_{ols}$, where $span(\widehat{\Phi}) = span(\widehat{G}_A)$. From the following propositions, under some normality assumptions, Cook et al. (2010) showed that the envelope estimator has an asymptotic variance no greater than that of the standard ordinary least square (OLS) estimator in multivariate regression. When X_i, Y_i are i.i.d with finite fourth moments, $\sqrt{n}\{\text{vec}(\widehat{\beta}) - \text{vec}(\beta)\}$ converges to a normal random vector with mean 0. Especially, when X_i, Y_i are joint normal, the asymptotic covariance matrix is $\Phi\Omega^{-1}\Phi^T \otimes \Sigma + (\Phi_0 \otimes \eta^T)M^\dagger(\Phi_0^T \otimes \eta)$, where $\Sigma = cov(\varepsilon)$, $M^\dagger$ is the generalized inverse of $M = \Omega_0 \otimes \eta\Sigma^{-1}\eta^T + \Omega_0 \otimes \Omega^{-1} - 2I_{p-u} \otimes I_u$. Additionally,

if the eigenvalues of material part are much larger than that of the immaterial part, the predictor envelope estimator can achieve substantial asymptotic efficiency gain.

5.2 Envelope-IV method

5.2.1 First stage predictor envelope

Motivated by the definition of a weak IV, a natural way of reducing the dimension of IVs is to impose the predictor envelope on the first-stage regression. As stated in Algorithm 6, the first stage of the 2SLS regresses the endogenous exposure X onto the exogenous IV Z. Therefore, there is not endogeneity at this stage and we can directly impose the standard predictor envelope to reduce the dimension of the IV. Suppose X and ε_X are independent of k. Consider the following for an orthogonal matrix (Φ, Φ_0) with $\Phi \in \mathbb{R}^{k\times u}$, $\Phi_0 \in \mathbb{R}^{k\times(k-u)}$, $0 \le u \le k$, where u is given, such that

Condition 5.2.1 $cov(X_i, \Phi_0^T Z_i \mid \Phi^T Z_i) = 0$,

Condition 5.2.2 $cov(\Phi^T Z_i, \Phi_0^T Z_i) = 0$.

Under these two conditions, first stage regression can be written as: $X_i = \eta^T \Phi^T Z_i + \varepsilon_i$, where X_i has variance $\Sigma_X = \Phi\Omega\Phi^T + \Phi_0\Omega_0\Phi_0^T$, $\Omega \in \mathbb{R}^{u\times u}$ and $\Omega_0 \in \mathbb{R}^{(k-u)\times(k-u)}$. We can obtain envelope basis Φ by solving objective function replacing response and predictor from (5.5) as

$$\underset{A\in\mathbb{R}^{(k-u)\times u}}{\arg\min} \; -2\log|G_A^T G_A| + \log|G_A^T \widehat{\Sigma}_{Z|X} G_A| + \log|G_A^T \widehat{\Sigma}_Z^{-1} G_A|. \tag{5.6}$$

Then we have first stage envelope estimator $\widehat{\delta}_{1st,env}^T = P_{\widehat{\Phi}(\widehat{\Sigma}_Z)}\widehat{\Sigma}_Z^{-1}\widehat{\Sigma}_{Z,X} = \widehat{\Phi}(\widehat{\Phi}^T\widehat{\Sigma}_Z\widehat{\Phi})^{-1}\widehat{\Phi}^T\widehat{\Sigma}_{Z,X}$, and $\widehat{\beta}_{1st,env} = \widehat{\Sigma}_{Y,Z}\widehat{\delta}_{1st,env}^T\left(\widehat{\delta}_{1st,env}\widehat{\Sigma}_Z\widehat{\delta}_{1st,env}^T\right)^{-1}$, where $\widehat{\Sigma}_Z = Z^TZ/n$, $\widehat{\Sigma}_{Y,Z} = Y^TZ/n$ and $\widehat{\Sigma}_{Z,X} = Z^TX/n$.

We now investigate the asymptotic behavior of $\widehat{\beta}_{2SLS}$ and $\widehat{\beta}_{1st,env}$. Cook et al. (2013) showed that the envelope estimators can achieve substantial efficiency gain. However, as we show below, the first stage envelope estimator is warranted to have the same efficiency as the 2SLS estimator.

Proposition 5.2.1 *Under Model* (5.1)*, suppose* $(X_i^T, Z_i^T)^T$ *are i.i.d. with finite fourth moment,* $i = 1, \ldots, n$*. Then,* $\sqrt{n}\{vec(\widehat{\beta}_{2SLS} - vec(\beta))\}$ *converges to a normal*

distribution with mean 0 *and variance* V_{2SLS}*, where* $V_{2SLS} = (\delta\Sigma_Z\delta^T)^{-1}\otimes\Sigma_{\varepsilon_Y}$*, where* $\Sigma_Z = cov(Z)$ *and* $\Sigma_{\varepsilon_Y} = cov(\varepsilon_Y)$*.*

Proposition 5.2.2 *Under Model* (5.1)*, assume* u *is known and* $(X_i^T, Z_i^T)^T$ *are i.i.d with finite fourth moments. Then,* $\sqrt{n}\{vec(\widehat{\beta}_{1st,env}) - vec(\beta)\}$ *converges to a normal distribution with mean* 0 *and variance* $V_{1st,env} = (\delta\Sigma_Z\delta^T)^{-1}\otimes\Sigma_{\varepsilon_Y} = V_{2SLS}$*.*

Proposition 5.2.1 and 5.2.2 show that first stage envelope estimator behaves the same as the 2SLS estimator in terms of asymptotic efficiency even though it uses less IVs. This result is different from the traditional envelope result where the resulting estimator may achieve a substantial efficiency gain. The first stage envelope estimator cannot improve on the efficiency because the causal effect of interest β is a function of the parameter that we envelope. However, the first stage envelope estimator may reduce the finite sample bias as we show below.

For finite samples, Buse (1992) shows that under Model (5.1), the bias of the 2SLS estimator $E(\widehat{\beta}_{2SLS} - \beta)$ is approximately $(k-p-1)(\delta\Sigma_Z\delta^T)^{-1}\Sigma_{\varepsilon_X,\varepsilon_Y}/n$ where $P_Z = Z(Z^TZ)^{-1}Z^T$ and $\Sigma_{\varepsilon_X,\varepsilon_Y} = cov(\varepsilon_X, \varepsilon_Y)$. That is, the bias of 2SLS $E(\widehat{\beta}_{2SLS}-\beta)$ is approximately proportional to $(k-p-1)/n$. Hence, if the dimension of IV can be reduced, the finite sample bias is also smaller, when the approximation is close to the true bias, that is when the concentration parameter $\mu^2 = \delta Z^TZ\delta/\sigma^2_{\varepsilon_X}$ is moderately large, where $\sigma^2_{\varepsilon_X} = cov(\varepsilon_X)$. When the approximate bias is small compared with the variance, the MSE can be potentially reduced. The following Lemma shows that the ratio of the approximate bias of first stage envelope method and 2SLS can be $(u-p-1)/(k-p-1)$.

Lemma 5.2.1 *Under Model* (5.1)*, suppose envelope dimension* u *and basis* $\Phi \in \mathbb{R}^{k\times u}$ *are given. We have for arbitrary small* $\epsilon > 0$*,*

$$\widehat{\beta}_{1st,env} - \beta = \widehat{b} + o_p(n^{-1+\epsilon})$$

where $E(\widehat{b}) = (u-p-1)(\delta\Sigma_Z\delta^T)^{-1}\Sigma_{\varepsilon_X,\varepsilon_Y}/n + o_p(n^{-1})$*.*

Both Proposition 5.2.1 and Lemma 5.2.1 are assuming the true envelope dimension is given. In reality, we need to estimate the envelope dimension from data. To estimate it under a normality assumption, Cook (2018a) proposed to use likelihood ratio test (LRT), BIC to choose the dimension of first stage envelope. Without the normality assumption on the IVs and the errors, we select the envelope dimension by adapting the vector correlation coefficient proposed by Hotelling (1936a) evaluating

the distance of two space. Suppose A and B are semi-orthogonal $m \times n$ matrices, the vector correlation coefficient is defined as $q^2 = |B^T A A^T B|$, with $0 \leq q^2 \leq 1$ and high q^2 implies A and B are close. When the prefixed u is larger than the true dimension of envelope, the estimated envelope space contains the true envelope basis and other random basis orthogonal to the true envelope, causing q^2 away from 1. As a result, we carry out the dimension selection procedure as follows. First, we fix an envelope dimension from $k-1$ to 1 and calculate the corresponding envelope basis Φ. Then $m-$folds bootstrap gives envelope basis Φ_i for i-th bootstrap and the empirical $\widehat{q}^2(u) = \sum_{i=1}^{m} q^2(\Phi, \Phi_i)/m$. We choose the dimension u when $\widehat{q}^2(u)$ achieves a threshold, which we will use 0.9 as the threshold.

5.2.2 Second stage predictor envelope in ILS

First stage envelope method can reduce redundant IVs that does not contribute to the exposure X, which leads to the reduction of finite sample bias. However, the asymptotic variance of the first stage envelope estimator remains the same as the 2SLS estimator. Inspired by ILS, which we review below, we can avoid the endogeneity of the regression of the outcome on the exposure and adapt envelope method to reduce the dimension of target parameter β, which will lead to efficiency gain.

In Model (5.1), substituting the first equation into the second one, we have $Y_i = \beta X_i + \varepsilon_{Y_i} = \beta\delta Z_i + (\beta\varepsilon_{X_i} + \varepsilon_{Y_i}) = \lambda Z_i + \widetilde{\varepsilon}_{Y_i}$, where $\lambda = \beta\delta$ and $\widetilde{\varepsilon}_{Y_i} = \beta\varepsilon_{X_i} + \varepsilon_{Y_i}$. Hence, to estimate the causal effect, the ILS estimator is also obtained in two stages (see a summarized algorithm in the Appendix): The first stage is the same as 2SLS, from which we obtain $\widehat{\delta}_{2SLS} = X^T Z(Z^T Z)^{-1}$; At the second stage, we estimate λ by regressing Y on Z, $\widehat{\lambda}_{ILS} = Y^T Z(Z^T Z)^{-1}$. Then we calculate $\widehat{\beta}$ from $\widehat{\lambda}_{ILS} = \beta\widehat{\delta}_{2SLS}$, that is, $\widehat{\beta}_{ILS} = \widehat{\lambda}_{ILS} M \widehat{\delta}_{2SLS}^T \{\widehat{\delta}_{2SLS} M \widehat{\delta}_{2SLS}^T\}^{-1}$, where $M \in \mathbb{R}^{k\times k}$ can be any positive definite matrix. When $M = Z^T Z$, the ILS estimator $\widehat{\beta}_{2SLS} = Y^T Z \widehat{\delta}_{2SLS}^T \{\widehat{\delta}_{2SLS}(Z^T Z)\widehat{\delta}_{2SLS}^T\}^{-1}$ is equivalent to traditional 2SLS. For the rest of this paper, we always use such a choice of M so that ILS is the same as 2SLS.

A nice feature of ILS estimation is that the second stage encodes the relationship between IV and outcome. Thus, if we reduce the dimension of the IV at the second stage of the ILS regression, we remove not only those IVs that have no association with the X, but also those do not affect Y through X. Again, all the redundancy is defined using linear combination rather than coordinate-wisely. Similar to the first stage envelope, the envelope method is applicable in the second stage ILS because

there is no endogeneity between IV Z and errors $(\varepsilon_X, \varepsilon_Y)$.

Specifically, we apply predictor envelope at the second stage of ILS to obtain the envelope $\mathcal{E}_{\Sigma_Z}(\lambda^T)$ as follows. We consider Condition 5.2.2 together with the following conditions

Condition 5.2.3 $cov(Y_i, \Gamma_0^T Z_i \mid \Gamma^T Z_i) = 0$,

Condition 5.2.4 $cov(\Gamma^T Z_i, \Gamma_0^T Z_i) = 0$.

Under Model (5.1), Condition 5.2.3 implies Condition 5.2.1. As a result, the predictor envelope imposed at the second stage is contained in that of the first stage, i.e., $\mathcal{E}_{\Sigma_Z}(\lambda^T) \subseteq \mathcal{E}_{\Sigma_Z}(\delta^T)$. This means that by imposing the envelope at the second stage of the ILS regression, we can potentially remove more redundancy in the IVs as compared to that at the first stage, and thus achieves more accurate and efficient estimation. The second stage envelope basis can be estimated by solving similar objective function as (5.5): $\widehat{\Gamma} = \text{argmin}_{A \in \mathbb{R}^{(k-q)\times q}} -2\log|G_A^T G_A| + \log|G_A^T \widehat{\Sigma}_{Z|Y} G_A| + \log|G_A^T \widehat{\Sigma}_Z^{-1} G_A|$. Suppose $\Gamma \in \mathbb{R}^{k\times q}$ is the basis of $\mathcal{E}_{\Sigma_Z}(\lambda^T)$, and Γ_0 is the basis of its orthogonal subspace, then $\Sigma_Z = \Gamma\Delta\Gamma^T + \Phi_0\Delta_0{\Gamma_0}^T$, and $\lambda^T = \Gamma\eta$. Consequently, we obtain $\widehat{\lambda}_{2nd,env} = Y^T Z\widehat{\Gamma}(\widehat{\Gamma}^T Z^T Z\widehat{\Gamma})^{-1}\widehat{\Gamma}^T$ and $\widehat{\beta}_{2nd,env} = \widehat{\lambda}_{2nd,env}\widehat{\Sigma}_Z\widehat{\delta}_{2SLS}^T(\widehat{\delta}_{2SLS}\widehat{\Sigma}_Z\widehat{\delta}_{2SLS}^T)^{-1}$, where $\widehat{\Sigma}_Z = Z^T Z/n$ and $\widehat{\delta}_{2SLS} = X^T Z(Z^T Z)^{-1}$. Comparing with 2SLS, second stage envelope estimator changes the method of estimating $\widehat{\lambda}$.

We first consider the asymptotic behavior of second stage envelope with joint normality of (Y, X, Z). The following proposition shows that $\sqrt{n}\{vec(\widehat{\beta}_{2nd,env}) - vec(\beta)\}$ converges to a normal random vector and gives its asymptotic variance.

Proposition 5.2.3 *Under Model* (5.1)*, and suppose Assumptions 5.1.1–5.1.3 hold. Assume that* (Y_i, Z_i) *are independent and identically distributed with a normal distribution and assume the envelope dimension* u *is known. Then* $\sqrt{n}\{vec(\widehat{\beta}_{2nd,env}) - vec(\beta)\}$ *converges to a normal random vector with mean* 0 *and variance* $V_{2nd,env} = V_{2nd,env,\Gamma} + V_{cost}$ *in distribution, where* $V_{2nd,env,\Gamma} = (\delta\Sigma_Z\delta^T)^{-1}\delta\Gamma\Delta\Gamma^T\delta^T(\delta\Sigma_Z\delta^T)^{-1} \otimes \Sigma_{\varepsilon_Y} + (\delta\Sigma_Z\delta^T)^{-1}\delta\Gamma_0\Delta_0\Gamma_0^T\delta^T(\delta\Sigma_Z\delta^T)^{-1} \otimes \beta\Sigma_{\varepsilon_X}\beta^T$, $V_{cost} = (\delta\Sigma_Z\delta^T)^{-1}\delta\Sigma_Z(BTB^T - ATA^T)\Sigma_Z\delta^T(\delta\Sigma_Z\delta^T)^{-1}$, $A = \beta\Sigma_{\varepsilon_X}\beta^T + \beta\Sigma_{\varepsilon_X,\varepsilon_Y}$, $B = \Sigma_{\varepsilon_Y} + \Sigma_{\varepsilon_Y,\varepsilon_X}\beta^T$, *and* $T = \{\Sigma_{Y|Z}^{-1}\eta \otimes \Gamma_0\}(\eta^T\Sigma_{Y|Z}^{-1}\eta \otimes \Omega_0 + \Omega \otimes \Omega_0^{-1} + \Omega^{-1} \otimes \Omega_0 - 2I_{u(k-u)})^{\dagger}\{\eta^T\Sigma_{Y|Z}^{-1} \otimes \Gamma_0^T\}$.

The asymptotic variance of second stage envelope consists of two parts: the first part $V_{2nd,env,\Gamma}$ is the asymptotic variance given the envelope basis which will be derived in Proposition 5.2.4; the second part is the cost of estimating envelope.

Suppose we now have the basis Γ. Proposition 5.2.4 gives the asymptotic behavior of $\widehat{\beta}_{2nd,env}$ given the basis Γ.

Proposition 5.2.4 *Under Model* (5.1)*, and suppose Assumptions 5.1.1–5.1.3 hold, and further assume semi-orthogonal basis* Φ *of* $\mathcal{E}_{\Sigma_Z}(\lambda^T)$ *is known. Suppose* $(\varepsilon_{X_i}^T, \varepsilon_{Y_i}^T)^T$*,* $i = 1, \ldots, n$*, are i.i.d. with finite variance. Then* $\sqrt{n}(\widehat{\beta}_{2nd,env} - \beta)$ *converges in distribution to a normal distribution with mean* 0 *and variance* $V_{2nd,env,\Gamma}$*, where* $V_{2nd,env,\Gamma} = (\delta\Sigma_Z\delta^T)^{-1}\delta\Gamma\Delta\Gamma^T\delta^T(\delta\Sigma_Z\delta^T)^{-1}\otimes\Sigma_{\varepsilon_Y}+(\delta\Sigma_Z\delta^T)^{-1}\delta\Gamma_0\Delta_0\Gamma_0^T\delta^T(\delta\Sigma_Z\delta^T)^{-1}\otimes\beta\Sigma_{\varepsilon_X}\beta^T$.

Corollary 5.2.1 *Under the conditions in Proposition 5.2.4,* $V_{2SLS} - V_{2nd,env,\Gamma} = (\delta\Sigma_Z\delta^T)^{-1}\delta\Gamma_0\Delta_0\Gamma_0^T\delta^T(\delta\Sigma_Z\delta^T)^{-1} \otimes (\Sigma_{\varepsilon_Y} - \beta\Sigma_{\varepsilon_X}\beta^T)$*. Furthermore,* $V_{2SLS} - V_{2nd,env,\Gamma} > 0$ *if and only if* $\Sigma_{\varepsilon_Y} - \beta\Sigma_{\varepsilon_X}\beta^T > 0$*.*

Corollary 5.2.1 provides the formula of the efficiency gain of the second stage envelope estimator as compared with the standard 2SLS estimator. Such a difference in asymptotic variance is semi-positive definite when the effect $S = cov(\varepsilon_{Y_i}) - cov(\beta\varepsilon_{X_i}) \geq 0$, so the second stage envelope estimator is asymptotically more efficient or as efficient as the standard IV estimator when this condition holds. The scale parameter β will be called as effect size in this article.

In reality, the regression of Y and X is fixed based on what effect we are interested in. When all the IVs contain little information to explain X, ε_X can be large. If we have some informative IVs, $\varepsilon_X = X - E(X|Z)$ is relatively small compared with non-informative IVs. Note that $cov(\beta\varepsilon_X)$ is the scaled unexplained term of X by Z. When IVs are informative, $cov(\varepsilon_Y) - cov(\beta\varepsilon_X) \geq 0$ is more likely to hold.

Whether $S > 0$ holds or not is crucial for us to decide whether second stage envelope method is useful or not. We have two approach to test $S > 0$. The first one is bootstrap. We can conduct a confidence interval based on estimated S from bootstrap and therefore decide whether $S > 0$ or not. Another approach is to calculate the asymptotic variance of S and derive a confidence interval based on the fact that S is asymptotic normal when n is large. The asymptotic variance is obtained by delta method. Note that $\widehat{S}$ is a function of $\widehat{t}$, where $\widehat{t} = (vech(\widehat{\Sigma}_Y)^T, vec(\widehat{\Sigma}_{X,Y})^T, vec(\widehat{\Sigma}_{Z,Y})^T, vec(\widehat{\Sigma}_{Z,X})^T, vech(\widehat{\Sigma}_Z)^T)^T$, and $var(\widehat{t})$ is part in $var(vec(\widehat{\Sigma}_C)) = E(\sum_{i=1}^n C_iC_i^T \otimes C_iC_i)/n^2 - vec(\Sigma_C)vec(\Sigma_C)^T/n$, where C_i is (Y_i^T, X_i^T, Z_i^T). As a result, delta method gives the asymptotic variance of $\widehat{S}$.

Corollary 5.2.2 *Under the conditions in Proposition 5.2.4, when* $\mathcal{E}_{\Sigma_Z}(\lambda^T)$=$\mathcal{E}_{\Sigma_Z}(\delta^T)$*, we have*

$avar\{\sqrt{n}vec(\widehat{\beta}_{2SLS})\} = avar\{\sqrt{n}vec(\widehat{\beta}_{2nd,env})\}$. *Especially, when* $p = dim(X) = 1$, *we have* $\mathcal{E}_{\Sigma_Z}(\lambda^T) = \mathcal{E}_{\Sigma_Z}(\delta^T)$.

When $\mathcal{E}_{\Sigma_Z}(\lambda^T) = \mathcal{E}_{\Sigma_Z}(\delta^T)$, we only reduce the dimension of δ, the effect of exposure X on the IVs Z. For instance, when $p = dim(X) = 1$, there is no redundant dimension in β. Thus, second stage envelope method does not reduce the dimension of the target β and there is no asymptotic efficiency gain.

Similar to first stage envelope, Proposition 5.2.3 and 5.2.4 also requires that the true envelope dimension is known. We adapt the same way in Section 5.2.1 to estimate the dimension of second stage envelope.

5.3 Simulations

We investigate the sample bias, variance and MSE of 2SLS, the first stage envelope and second stage envelope estimators in two different scenarios: 1) large sample size with a positive effect; 2) moderate sample size with a negative effect size. Also, we only use partial significant IVs for the two envelope methods to show that it is important to utilize all available IVs instead of searching for more IVs.

5.3.1 Scenario 1

Under Model (5.1), we generate a sample of size $n = 4000$. Set $r = 1$, $p = 2$, $k = 5$, $\beta = (0.5, 0.5)$ and $\delta = \Gamma^T$, where Γ is the first two columns of a randomly generated orthonormal matrix $(\Gamma, \Gamma_0) \in \mathbb{R}^{k\times k}$. Suppose we have k available IV and $k' = 10$ invisible IV. For the i^{th} individual, we generate the available IV, Z_i, from $N(0, \Sigma_Z)$, where $\Sigma_Z = \Gamma\Delta\Gamma^T + \Gamma_0\Delta_0\Gamma_0^T$, $\Delta = ((3,2)^T, (2,3)^T)$ and $\Delta_0 = 0.3I_3$. The invisible IV $Z_{invisible,i}$ are generated from standard k' dimension normal distribution. Then, we generate the errors $(\varepsilon_X, \varepsilon_Y)$ following $N(0, ((1,0,4)^T, (0,16,16)^T, (4,16,33)^T)^T)$. Set $X = \delta Z + \delta' Z_{invisible} + \varepsilon_X$ and $Y = \beta X + \varepsilon_Y$, where $\delta' \in \mathbb{R}^{p\times k'}$ consisting of randomly generated $U(0,1)$. Under this setting, the dimensions of first and second stage envelope are 2 and 1 respectively. Also, the envelope basis is $\Gamma_{1st,env} = \Gamma$ and $\Gamma_{2nd,env} = \Gamma(1/\sqrt{2}, 1/\sqrt{2})^T$. We apply 2SLS, first stage envelope and second stage envelope using partial IVs with dimensions are estimated using 50 bootstrap samples. We also compare the performance of 2SLS using all the IVs with estimators using only partial IVs. The simulations are repeated 1000 times. The bootstrap method correctly selects the dimension of the first and second stage envelope 100% and 99.8% among 1000 repetitions.

	Full, 2SLS	2SLS	1st, env	2nd, env
Bias	1.25	0.27	0.19	0.22
tr(Var)	26.53	40.48	40.83	13.83
MSE	28.07	40.51	40.82	13.87

Table 5.1: Large sample with positive effect. (Bias is defined as $\|E(\widehat{\beta} - \beta)\|_2$.)

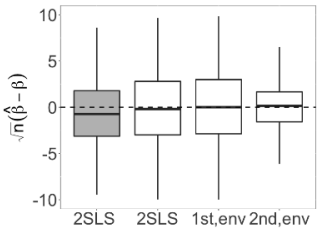
We conduct bootstrap with 300 repetitions to give a 95% confidence interval of $S = cov(\varepsilon_Y) - cov(\beta\varepsilon_X)$ to test whether second stage envelope can potentially reduce the variance or not. There are 100% tests in 1000 repetitions choosing the correct hypothesis that second stage efficiency condition holds. The empirical variances of S calculated by bootstrap, delta method and repetition are 10.01, 10.16 and 10.10 with true S being 17.25.

The boxplots of 2SLS with all the IVs, 2SLS, first stage envelope and second stage envelope estimators are given in Figure 5.1. In this scenario, the sample size is relatively large, so by Proposition 5.2.4, we expect the reduction in the bias of the first stage envelope estimator is not a leading part in the MSE and second stage envelope estimator can reduce the variance of 2SLS. From Table 5.1, first stage envelope estimator reduces the bias, but because variance is the leading term, it does not improve the MSE of 2SLS. While second stage envelope significantly reduce the variance and thus achieve efficiency gain in MSE. Note that second stage envelope with available IVs can be better than 2SLS with all the IVs in terms of MSE, it is important to use the material linear combinations of IVs via second stage envelope instead of searching for all possible IVs.

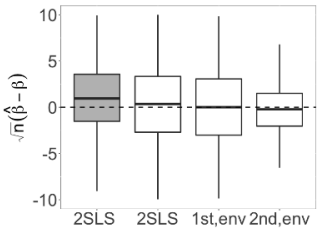
This result shows that for large sample, if condition $cov(\varepsilon_Y) - cov(\beta\varepsilon_X) > 0$ in Corollary 5.2.1 holds that is if error in the predictor is relatively small compare with error in the response (transformed though β to ensure they are under the same unit), second stage envelope method can significantly reduce the variance compare with 2SLS. Also, because variance plays an important role in MSE for large sample, second stage envelope can potentially reduce MSE.

5.3.2 Scenario 2

In scenario 2, we examine the performance of envelope estimators when the sample size is moderate and the effect is negative. We change the setting of scenario 1. Set $n = 300$, $k = 15$, $k' = 15$, $\beta = (1, 1)^T$ and variance of errors $(\varepsilon_X, \varepsilon_Y)$ to be $N(0, ((7, 0, -7)^T, (0, 7, 7)^T, (-7, 7, 15)^T)^T)$. We carry out 1000 repetitions. The



(a) First coordinate



(b) Second coordinate

Figure 5.1: Large sample: Boxplots of $\sqrt{n}(\widehat{\beta} - \beta)$ for the 2SLS with all the IVs, 2SLS, the first stage envelope and the second stage envelope estimators with positive effect. (Boxplot of 2SLS with all the IVs is ploted in gray.)

	Full, 2SLS	2SLS	1st, env	2nd, env
Bias	5.78	5.48	0.61	2.91
tr(Var)	3.86	8.39	12.01	21.86
MSE	37.30	38.42	12.38	30.32

Table 5.2: Moderate sample with negative effect. (Bias is defined as $\|E(\widehat{\beta} - \beta)\|_2$.)

q^2 method selects correct first stage envelope dimension in 81.3% repetitions and correct second stage envelope dimension in 99.9% repetitions.

We conduct bootstrap with 300 repetitions to give a 95% confidence interval of $S = cov(\varepsilon_Y) - cov(\beta\varepsilon_X)$ to test whether second stage envelope can potentially reduce the variance or not. All tests in 1000 repetitions choosing the correct hypothesis that second stage efficiency condition fails. The empirical variances of S calculated by bootstrap, delta method and repetition are 20.62, 31.50 and 24.88 with true S being -1.

The boxplots of 2SLS with all the IVs, 2SLS, first stage envelope and second stage envelope estimators are given in Figure 5.2. In this scenario, the sample size is moderate, we expect the reduction in the bias of the first stage envelope estimator can be significant in MSE. From Table 5.2, first stage envelope estimator significantly reduces the bias and enlarge the variance in the same time. The trade off improves the performance of 2SLS in terms of MSE. Because the effect is negative, second stage envelope estimator has a larger variance comparing with 2SLS. Second stage envelope estimator uses fewer dimension of feature space, thus the bias is also reduced.

(a) First coordinate

(b) Second coordinate

Figure 5.2: Moderate sample: Boxplots of $\sqrt{n}(\widehat{\beta} - \beta)$ for the 2SLS with all the IVs, 2SLS, the first stage envelope and the second stage envelope estimators with positive effect. (Boxplot of 2SLS with all the IVs is ploted in gray.)

5.4 Real Data

Wisconsin Longitudinal Study (WLS) (Herd et al., 2014) is a longitudinal survey of a random sample of men and women who graduated from Wisconsin high schools in 1957. The survey data provides record of social background, schooling, family formation and labor market experience of the respondents. The goal is to use the WLS to replicate a well-studied question in labor economics, the effect of education, measured in terms of years of schooling, on wages, measured in log units; see Card (2001); Blundell et al. (2005); Card (1999) for reviews.

Specifically, in our analysis, we use the log of wage in 1974 as the outcome and years of education as the treatment. Additionally, we studied the main effect of education, and the interaction effect of education and sex as the exposures.

For instruments, we use college availability (a discrete variable with nine levels), number of children in the family and the order of birth as IVs. College availability measures how far did the individual live away from a college while growing up and has been used in the past as an instrument for studying the return on education (Card, 1993).

In this way, we have $X \in \mathbb{R}^2$, $Y \in \mathbb{R}^1$, and $Z \in \mathbb{R}^{10}$. At the same time, we control for the following variables: sex, home population, dad and mom's education, parents income, size of the high school attend by the student, the type of high school, parents income class compared in the neighborhood and father's occupation.

We did a complete-case analysis on our dataset, that is, we removed any records that contains at least one missing value. There are 4775 observations left after we handled missing data.

Firstly, we compare OLS, 2SLS and LIML estimators. Their point estimates are

shown in Table 5.3. Although the point estimates of OLS and 2SLS have the same sign, their magnitude differs a lot. Also, the 2SLS estimator is close to the LIML estimator which suggests the reliability of 2SLS in this case.

	educ			educ×sex		
	$\widehat{\beta}$	$\widehat{\text{SE}}$	t-stat	$\widehat{\beta}$	$\widehat{\text{SE}}$	t-stat
OLS	0.075	0.006	11.90	-0.017	0.006	-2.99
LIML	0.141	0.065	2.17	-0.323	0.142	-2.27
2SLS	0.131	0.062	2.11	-0.298	0.127	-2.35
1st,env	0.137	0.053	2.49	-0.324	0.123	-2.63
2nd,env	0.134	0.030	4.47	-0.306	0.068	-4.50

Table 5.3: The point estimates, standard error and t-stat for the effect of education and its interactions on wage for OLS, LIML, OLS, and first and second stage envelope methods

Next, we apply our proposed first and second stage envelope methods. Based on bootstrap envelope dimension selection, the dimensions for the first and second stage envelope are all 6, which reduce the dimensions by 3. The point estimate of $\widehat{\beta}_{2SLS}$, $\widehat{\beta}_{1st,env}$ and $\widehat{\beta}_{2nd,env}$ are $(0.131, -0.298)$, $(0.137, -0.324)$, and $(0.134, -0.306)$ respectively for education, and the interaction effect of education and sex. The point estimates are close for the three different methods. The estimated 95% confidence interval of second stage efficiency condition in Corollary 5.2.1 calculated by 300 repetitions bootstrap is $(0.592, 0.762)$ which is above 0. This suggests our second stage envelope has the potential for efficiency gains.

We then evaluate the standard deviations of $\widehat{\beta}_{2SLS}$, $\widehat{\beta}_{1st,env}$ and $\widehat{\beta}_{2nd,env}$ using bootstrap with 200 repetitions. The standard deviation for 2SLS, first and second stage envelope is given in Table 5.3. The first stage envelope is roughly having the same standard deviation as the OLS, and the second stage envelope have 50.7% and 46.4% reduction on the three coordinates. In addition, we have the expected errors being 1.03 (2SLS), 0.96 (first stage envelope) and 0.82 (second stage envelope) through a 5-fold cross-validation. The expected errors drop from 1.02 to 0.98 as the first stage envelope have a smaller bias as compare to the 2SLS.

5.5 Discussion

In this project, we generalized the envelope method to the setting with unmeasured confounding. We explored two possible ways of imposing the envelope conditions.

While the first stage envelope does not yield any asymptotic efficiency gain on the standard IV estimator, the second stage envelope can achieve substantial efficiency gain. An inspiring message we get is that the notion of redundant IV or weak IV should not only be defined with regard to the relationship between X and Z, but should also be whether IV affects Y through X.

Although the majority models we consider in this paper is linear models, our methods can be extended to a more general model as we briefly explained in Section 5.2.2.

Throughout the paper, the coefficients are assumed fixed, that is, they do not change with sample size. Another scenario that has been considered in the literature of weak IVs is that the strength of IV decrease as the sample size increases. For examples, Bekker (1994) developed asymptotic approximations of the bias when the number of instruments k is proportional to the sample size n, $\delta Z^T Z \delta^T/(n-p)$ and $\varepsilon_X^T \varepsilon_X/(n-p)$ fixed in Model (5.1). Staiger and Stock (1997) considered when k is fixed, $\sqrt{n}\delta$ is constant. Our methods can potentially be extended to such a situation. We leave the extension for future exploration.

Appendix A

Appendix for Chapter 2

A.1 Proof of Propositions

Proof of Proposition 2.4.1

Let $\hat{\boldsymbol{\beta}}_{std,obs}$, $\hat{\boldsymbol{\Sigma}}_{std,obs}$ denote the maximizer of the observed data likelihood under model (2.1). Let $\hat{\boldsymbol{\beta}}_{env,obs}$, $\hat{\boldsymbol{\Sigma}}_{env,obs}$ denote the maximizer of the observed data likelihood assuming (i) and (ii) under model (2.1).

Similar as the notations in the manuscript, we omit the vectorization notations here. Let $\mathbf{V}_0 = \text{avar}(\hat{\boldsymbol{\beta}}_{env,obs}, \hat{\boldsymbol{\Sigma}}_{env,obs})$ and $\mathbf{V} = \text{avar}(\hat{\boldsymbol{\beta}}_{std,obs}, \hat{\boldsymbol{\Sigma}}_{std,obs})$ denote the asymptotic covariance matrices of the estimators for the EM envelope parameters and the standard EM parameters. Also, let $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0)$ and $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$ denote the parameter under the envelope model and the standard model. Assume $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ satisfy the following condition A.1.1.

Condition A.1.1 *L_{obs} is unimodal, i.e, the probability distribution has a single maximum, in the parameter space $\boldsymbol{\Phi}$ with only one point $\boldsymbol{\phi}_0$ such that $\partial Q(\boldsymbol{\phi}|\boldsymbol{\phi}_t)/\partial\boldsymbol{\phi}|_{\phi=\phi_0} = 0$, and that $\partial Q(\boldsymbol{\phi}|\boldsymbol{\phi}_t)/\partial\boldsymbol{\phi}$ is continuous in $\boldsymbol{\phi}$ and $\boldsymbol{\phi}_t$,*

Let $\{\boldsymbol{\theta}_t\}$ and $\{\tilde{\boldsymbol{\theta}}_t\}$ denote the EM sequences, i.e., the parameters sequences we obtain from each EM iteration, of the envelope model and the standard model. Under the regularity condition A.1.1 and by Corollary 1 of Wu (1983), the two EM sequences$\{\boldsymbol{\theta}_t\}$ and $\{\tilde{\boldsymbol{\theta}}_t\}$ converge to their unique maximizer of L_{obs}. Hence, in order to prove $\mathbf{V}_{env} \leq \mathbf{V}_{std}$, it suffices to prove $\mathbf{V}_0 \leq \mathbf{V}$. We can find function $\mathbf{h}$ such that

$$\mathbf{h}(\boldsymbol{\theta}) = \begin{pmatrix} \text{vec}(\boldsymbol{\beta}) \\ \text{vech}(\boldsymbol{\Sigma}) \end{pmatrix} = \begin{pmatrix} \text{vec}(\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T) \\ \text{vech}(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) \end{pmatrix}.$$

Becaues of the over-parameterization of $\boldsymbol{\theta}$, the gradient matrix $\mathbf{G} = \dfrac{\partial \mathbf{h}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$ is not of full rank. By Proposition 4.1 in Shapiro (1986), we have

$$\mathbf{V}_0 = \mathbf{G}(\mathbf{G}^T\mathbf{V}^{-1}\mathbf{G})^\dagger\mathbf{G}^T.$$

Hence,

$$\mathbf{V} - \mathbf{V}_0 = \mathbf{V} - \mathbf{G}(\mathbf{G}^T\mathbf{V}^{-1}\mathbf{G})^\dagger\mathbf{G}^T = \mathbf{V}^{\frac{1}{2}}[\mathbf{I} - \mathbf{V}^{-\frac{1}{2}}\mathbf{G}(\mathbf{G}^T\mathbf{V}^{-1}\mathbf{G})^\dagger\mathbf{G}^T\mathbf{V}^{-\frac{1}{2}}]\mathbf{V}^{\frac{1}{2}}.$$

Since $\mathbf{I} - \mathbf{V}^{-\frac{1}{2}}\mathbf{G}(\mathbf{G}^T\mathbf{V}^{-1}\mathbf{G})^\dagger\mathbf{G}^T\mathbf{V}^{-\frac{1}{2}}$ is the projection matrix onto the orthogonal complement of span($\mathbf{V}^{-\frac{1}{2}}\mathbf{G}$), it is positive semi-definite. Hence, $\mathbf{V}_0 \leq \mathbf{V}$.

Proof of Proposition 2.4.2

Since the EM envelope model is overparameterized, we only need to show the standard EM estimator using misspecified likelihood is $\sqrt{n}$-consistent and asymptotically normal. Once we prove that, we can follow the same argument as in the proof of Proposition 2.4.1, using Proposition 4.1 of Shapiro (1986), to prove that our method has efficiency gain over the standard estimator.

We assume the regularity conditions in Condition A.1.1 hold, so that the EM sequence $\hat{\boldsymbol{\theta}}_{em\_std}$ converge to the observed data MLE $\hat{\boldsymbol{\theta}}_{obs\_std}$. Additionally, we assume the following regularity conditions: the error $\boldsymbol{\varepsilon}_i$ and covariates $\mathbf{X}_i$ has finite $(4+\delta)$-th moment, $\liminf_n \lambda_-\{n^{-1}\mathrm{Var}(s_n(\boldsymbol{\theta}))\} > 0$ and $\liminf_n \lambda_-\{n^{-1}\mathbf{M}_n(\boldsymbol{\theta})\} > 0$, where $\lambda_-(\mathbf{A})$ denote the smallest eigenvalue of $\mathbf{A}$, $s_n(\boldsymbol{\theta}) = \dfrac{\partial l}{\partial \boldsymbol{\theta}}$, $\mathbf{M}_n(\boldsymbol{\theta}) = -\mathbb{E}\{\dfrac{\partial^2 l}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\}$, and l is the log-likelihood when the error and covariates are normally distributed.

We aim to use Proposition 5.5 and Theorem 5.14 in Shao (2003) to prove consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}_{obs\_std}$. Since we treat $\mathbf{X}_i$ as normally distributed, the estimator $\hat{\boldsymbol{\theta}}_{obs\_std}$ is obtained by maximizing the following misspecified likelihood:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int\int (2\pi)^{-\frac{r+p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} |\boldsymbol{\Sigma}_x|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta})\}$$
$$\cdot \exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_x)^T\boldsymbol{\Sigma}_x^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_x)\} d\mathbf{x}_{i,mis} d\mathbf{y}_{i,mis}.$$

From Section A.2 and notations therein, we have

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} |\mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\Sigma}}\mathbf{B}_i^T\mathbf{S}_i^T|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{D}_{i,obs}-\mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\mu}})^T(\mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\Sigma}}\mathbf{B}_i^T\mathbf{S}_i^T)^{-1}(\mathbf{D}_{i,obs}-\mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\mu}})\}.$$

By denoting $\mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_{i,obs}$ and $\mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\Sigma}}\mathbf{B}_i^T\mathbf{S}_i^T = \boldsymbol{\Sigma}_{i,obs}$, we have

$$L(\boldsymbol{\theta}) \propto |\boldsymbol{\Sigma}_{i,obs}|^{-\frac{1}{2}} \exp\{(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T\boldsymbol{\Sigma}_{i,obs}^{-1}(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})\}.$$

The estimator $\hat{\boldsymbol{\theta}}_{obs\_std}$ is the solution to the following generalized estimating equation (GEE)

$$\frac{\partial l}{\partial \boldsymbol{\theta}^T} = \sum_{i=1}^{n} = \frac{\partial l}{\partial \boldsymbol{\theta}^T} = \sum_{i=1}^{n} \psi_i^T(\mathbf{D}_{i,obs}, \boldsymbol{\theta}) = 0,$$

where l_i is the misspecified log-likelihood of each observation, and $\psi_i(\mathbf{D}_{i,obs}, \boldsymbol{\theta}) = \dfrac{\partial l_i}{\partial \boldsymbol{\theta}}$.

Let vech$(\cdot)$ denote the half-vectorization operator. For example, vech$(\mathbf{A})$ of a symmetric $n \times n$ matrix $\mathbf{A}$ is the $n(n+1)/2 \times 1$ column vector obtained by vectorizing only the lower triangular part of $\mathbf{A}$. Let $\mathbf{E}_n \in \mathbb{R}^{n^2 \times n(n+1)/2}$ be the expansion matrix such that vec$(\cdot) = \mathbf{E}_n$vech$(\cdot)$ for a symmetric $n \times n$ matrix. Denote $\mathbf{M}_{i1} = \dfrac{\partial \boldsymbol{\mu}_{i,obs}}{\partial \boldsymbol{\mu}_x^T}$, $\mathbf{M}_{i2} = \dfrac{\partial \boldsymbol{\mu}_{i,obs}}{\partial \boldsymbol{\beta}^T}$, $\mathbf{M}_{i3} = \dfrac{\partial \text{vech}(\boldsymbol{\Sigma}_{i,obs})}{\partial \text{vech}(\boldsymbol{\Sigma})^T}$, $\mathbf{M}_{i4} = \dfrac{\partial \text{vech}(\boldsymbol{\Sigma}_{i,obs})}{\partial \boldsymbol{\beta}^T}$, and $\mathbf{M}_{i5} = \dfrac{\partial \text{vech}(\boldsymbol{\Sigma}_{i,obs})}{\partial \text{vech}(\boldsymbol{\Sigma}_x)^T}$. Also, let k_i denote the length of $\mathbf{D}_{i,obs}$.

By matrix calculus, we have

$$\frac{\partial l_i}{\partial \boldsymbol{\mu}_x^T} = (\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T\boldsymbol{\Sigma}_{i,obs}^{-1}\mathbf{M}_{i1},$$

$$\frac{\partial l_i}{\partial \text{vech}(\boldsymbol{\Sigma}_x)^T} = -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1})^T\mathbf{E}_{k_i}\mathbf{M}_{i5}+\frac{1}{2}(\mathbf{D}_{i,obs}-\boldsymbol{\mu}_{i,obs})^T\otimes(\mathbf{D}_{i,obs}-\boldsymbol{\mu}_{i,obs})^T(\boldsymbol{\Sigma}_{i,obs}^{-1}\otimes\boldsymbol{\Sigma}_{i,obs}^{-1})\mathbf{E}_{k_i}\mathbf{M}_{i5},$$

$$\frac{\partial l_i}{\partial \text{vech}(\boldsymbol{\Sigma})^T} = -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1})^T\mathbf{E}_{k_i}\mathbf{M}_{i3}+\frac{1}{2}(\mathbf{D}_{i,obs}-\boldsymbol{\mu}_{i,obs})^T\otimes(\mathbf{D}_{i,obs}-\boldsymbol{\mu}_{i,obs})^T(\boldsymbol{\Sigma}_{i,obs}^{-1}\otimes\boldsymbol{\Sigma}_{i,obs}^{-1})\mathbf{E}_{k_i}\mathbf{M}_{i3},$$

$$\begin{aligned}\frac{\partial l_i}{\partial \boldsymbol{\beta}^T} &= \frac{\partial l_i}{\partial \text{vech}(\boldsymbol{\Sigma}_{i,obs})^T}\frac{\partial \text{vech}(\boldsymbol{\Sigma}_{i,obs})}{\partial \boldsymbol{\beta}^T} + \frac{\partial l_i}{\partial \boldsymbol{\mu}_{i,obs}^T}\frac{\partial \boldsymbol{\mu}_{i,obs}}{\partial \boldsymbol{\beta}^T} \\ &= -\frac{1}{2}\text{vec}(\boldsymbol{\Sigma}^{-1})^T\mathbf{E}_{k_i}\mathbf{M}_{i4} + \frac{1}{2}(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T \otimes (\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T(\boldsymbol{\Sigma}_{i,obs}^{-1} \otimes \boldsymbol{\Sigma}_{i,obs}^{-1})\mathbf{E}_{k_i}\mathbf{M}_{i4} \\ &\quad + (\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T\boldsymbol{\Sigma}_{i,obs}^{-1}\mathbf{M}_{i2}.,\end{aligned}$$

and

$$\psi_i(\mathbf{D}_{i,obs},\boldsymbol{\theta}) = (\frac{\partial l_i}{\partial \boldsymbol{\mu}_x^T}, \frac{\partial l_i}{\partial \text{vech}(\boldsymbol{\Sigma}_x)^T}, \frac{\partial l_i}{\partial \text{vech}(\boldsymbol{\Sigma})^T}, \frac{\partial l_i}{\partial \boldsymbol{\beta}^T})^T$$

Since $\mathbb{E}(\mathbf{D}_{i,obs}) = \boldsymbol{\mu}_{i,obs}$, and $\text{Var}(\mathbf{D}_{i,obs}) = \boldsymbol{\Sigma}_{i,obs}$, we have

$$\mathbb{E}\{(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T \otimes (\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T(\boldsymbol{\Sigma}_{i,obs}^{-1} \otimes \boldsymbol{\Sigma}_{i,obs}^{-1})\} = \text{vec}(\boldsymbol{\Sigma}^{-1})^T.$$

Hence, $\mathbb{E}\{\psi_i(\mathbf{D}_{i,obs},\boldsymbol{\theta}_0)\} = 0$ where the subscript 0 indicates the true parameter value.

In order to use Proposition 5.5 in Shao (2003) to prove consistency, we need to show the conditions in Lemma 5.3 in Shao (2003) hold for any compact subset of the parameter space. That is, for any $c > 0$ and sequence $\{\mathbf{D}_{i,obs}\}_{i=1}^{\infty}$ satisfying $\|\mathbf{D}_{i,obs}\| \leq c$, the sequence of functions $\psi_i(\mathbf{D}_{i,obs},\boldsymbol{\theta})$ is equicontinuous on any compact set of the parameter space.

By taking derivative for $\psi_i(\mathbf{D}_{i,obs},\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, we will see that $\dfrac{\partial \psi_i}{\partial \boldsymbol{\theta}}$ is continuous in $\boldsymbol{\theta}$ and $\mathbf{D}_{i,obs}$. Hence, when the parameter space Θ is compact and $\|\mathbf{D}_{i,obs}\| \leq c$, $\dfrac{\partial \psi_i}{\partial \boldsymbol{\theta}}$ is uniformly bounded. Therefore, $\psi_i(\mathbf{D}_{i,obs},\boldsymbol{\theta})$ is equicontinuous. Moreover, since both $\boldsymbol{\varepsilon}_i$ and $\mathbf{X}_i$ have finite $(4+\delta)$-th moment, we have $\mathbb{E}\{\sup_{\theta\in\Theta}\|\psi_i(\mathbf{D}_{i,obs},\boldsymbol{\theta})\|\}^2 < \infty$, and $\mathbb{E}\|\mathbf{D}_{i,obs}\| < \infty$. The conditions in Lemma 5.3 in Shao (2003) holds.

By referring to Proposition 5.5 in Shao (2003), we also need to prove that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{\psi_i(\mathbf{D}_{i,obs},\boldsymbol{\theta})\} = 0$$

implies $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. For each i, the missing pattern can be arbitrary. Hence $\lim_{n\to\infty}\sum_{i=1}^{n}\mathbb{E}\{\psi_i(\mathbf{D}_{i,obs},\boldsymbol{\theta})\}/n = 0$ implies $\lim_{n\to\infty}\sum_{i=1}^{n}\mathbb{E}\{\psi_i(\mathbf{Y}_i,\mathbf{X}_i,\boldsymbol{\theta})\}/n = 0$, where $\psi_i(\mathbf{Y}_i,\mathbf{X}_i,\boldsymbol{\theta})$ is the score for the full data. Hence $\lim_{n\to\infty}\sum_{i=1}^{n}\mathbb{E}\{\psi_i(\mathbf{Y}_i,\mathbf{X}_i,\boldsymbol{\theta})\}/n = \mathbb{E}\{\psi_i(\mathbf{Y}_i,\mathbf{X}_i,\boldsymbol{\theta})\} = 0$ implies $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Since the observed data MLE $\hat{\boldsymbol{\theta}}_{obs\_std}$ is always $\mathcal{O}(1)$, by Proposition 5.5 in Shao (2003), $\hat{\boldsymbol{\theta}}_{obs\_std} \xrightarrow{p} \boldsymbol{\theta}_0$.

Then, we prove asymptotic normality of $\hat{\boldsymbol{\theta}}_{em\_std}$ using Theorem 5.14 in Shao (2003). Since $\mathbf{D}_{i,obs}$ has finite $(4+\delta)$-th moment, $\sup_i\|\psi_i(\mathbf{D}_{i,obs},\boldsymbol{\theta})\|^{2+\frac{\delta}{2}} < \infty$. Then, if conditions $\liminf_n \lambda_-\{n^{-1}\text{Var}(s_n(\boldsymbol{\theta}))\} > 0$ and $\liminf_n \lambda_-\{n^{-1}\mathbf{M}_n(\boldsymbol{\theta})\} > 0$ holds, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{obs\_std} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0,\mathbf{V}).$$

Since the regularity conditions in Wu (1983) hold, we have $\hat{\boldsymbol{\theta}}_{em\_std} \to \hat{\boldsymbol{\theta}}_{obs\_std}$. There-

fore,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{em\_std} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \mathbf{V}).$$

We proved the $\sqrt{n}$-consistency and asymptotical normality of $\hat{\boldsymbol{\theta}}_{em\_std}$. Because the envelope model is overparameterized, by Proposition 4.1 of Shapiro (1986),

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{em\_env} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \mathbf{V}_0).$$

A.2 The derivations of examples

In the following example, we show that if $(\mathbf{X}_i^T, \mathbf{Y}_i^T)^T$ follows a normal distribution, then $(\mathbf{Y}_{i,obs}^T, \mathbf{X}_{i,obs}^T)^T$ follows a closed form distribution.

Example A.2.1 *Suppose the predictors and responses are normally distributed as* $\mathbf{Y}_i|\mathbf{X}_i \overset{i.i.d}{\sim} N(\boldsymbol{\beta}\mathbf{X}_i, \boldsymbol{\Sigma})$ *and* $\mathbf{X}_i \overset{i.i.d}{\sim} N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. *Then,* $(\mathbf{Y}_{i,obs}^T, \mathbf{X}_{i,obs}^T)^T$ *independently follows a normal distribution* $N(\boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}_i^*)$, *where the explicit form of the parameter* $\boldsymbol{\mu}_i^* = \mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\mu}}$ *and* $\boldsymbol{\Sigma}_i^* = \mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\Sigma}}\mathbf{B}_i^T\mathbf{S}_i^T$ *where* $\mathbf{B}_i$, $\mathbf{S}_i$, $\tilde{\boldsymbol{\mu}}$ *and* $\tilde{\boldsymbol{\Sigma}}$ *are given below.*

Derivation of Example A.2.1. Note that $\mathbf{Y}_i|\mathbf{X}_i \overset{i.i.d}{\sim} N(\boldsymbol{\beta}\mathbf{X}_i, \boldsymbol{\Sigma})$ and $\mathbf{X}_i \overset{i.i.d}{\sim} N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$; hence, $(\mathbf{X}_i^T, \mathbf{Y}_i^T)^T \overset{i.i.d}{\sim} N(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, where $\tilde{\boldsymbol{\mu}} = (\boldsymbol{\mu}_x^T, \boldsymbol{\mu}_x^T\boldsymbol{\beta}^T)^T$, and $\tilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_x\boldsymbol{\beta} \\ \boldsymbol{\beta}^T\boldsymbol{\Sigma}_x & \boldsymbol{\Sigma} + \boldsymbol{\beta}^T\boldsymbol{\Sigma}_x\boldsymbol{\beta} \end{pmatrix}$. Also, there exists a unique permutation matrix $\mathbf{B}_i$, i.e., a square matrix that has exactly one entry of 1 in each row and each column and 0s everywhere, such that $(\mathbf{X}_{i,obs}^T, \mathbf{Y}_{i,obs}^T, \mathbf{X}_{i,mis}^T, \mathbf{Y}_{i,mis}^T)^T = \mathbf{B}_i(\mathbf{X}_i^T, \mathbf{Y}_i^T)^T$; thus, $(\mathbf{X}_{i,obs}^T, \mathbf{Y}_{i,obs}^T, \mathbf{X}_{i,mis}^T, \mathbf{Y}_{i,mis}^T)^T$ independently follows $N(\mathbf{B}_i\tilde{\boldsymbol{\mu}}, \mathbf{B}_i\tilde{\boldsymbol{\Sigma}}\mathbf{B}_i^T)$. Therefore, by the property of normal distribution, $(\mathbf{X}_{i,obs}^T, \mathbf{Y}_{i,obs}^T)^T \sim N(\mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\mu}}, \mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\Sigma}}\mathbf{B}_i^T\mathbf{S}_i^T)$, where $\mathbf{S}_i = \begin{pmatrix} \mathbf{I}_{k_i} & \mathbf{O}_{k_i\times(l-k_i)} \end{pmatrix}$, $\mathbf{O}_{a\times b}$ is a matrix of size $a \times b$ with all elements being 0, k_i is the total length of $(\mathbf{X}_{i,obs}^T, \mathbf{Y}_{i,obs}^T)^T$, and l is the total length of $(\mathbf{X}^T, \mathbf{Y}^T)^T$. Hence, $\boldsymbol{\mu}_i^* = \mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\mu}}$, and $\boldsymbol{\Sigma}_i^* = \mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\Sigma}}\mathbf{B}_i^T\mathbf{S}_i^T$

The updates of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ have been discussed above. Here, we present two examples focusing on the calculation of $\mathbf{A}_{j,t}$ and $\boldsymbol{\rho}$.

Example A.2.2 *Under model (2.1) and assume* $\mathbf{X}_i \overset{i.i.d}{\sim} N_p(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. *Then, the update of parameters are* $\boldsymbol{\mu}_{x,t+1} = \mathbb{E}(\mathbf{X}_i|\mathbf{D}_{i,obs}; \boldsymbol{\theta}_t)/n$ *and* $\boldsymbol{\Sigma}_{x,t+1} = \{\mathbf{A}_{3,t} - 2\mathbb{E}(\mathbf{X}_i|\mathbf{D}_{i,obs}; \boldsymbol{\theta}_t)\boldsymbol{\mu}_{x,t+1}\}/n + \boldsymbol{\mu}_{x,t+1}\boldsymbol{\mu}_{x,t+1}^T$. *The proof and the calculation of* $\mathbf{A}_{j,t}$ *is shown below.*

Example A.2.3 *Under model (2.1), assume* $p = 1$ *and* $X_i \overset{i.i.d}{\sim} \text{Ber}(\pi)$. *The update of parameter is* $\pi_{t+1} = \sum_{i=1}^n \tilde{\pi}_{i,t}/n$. *The form of* $\tilde{\pi}_{i,t}$ *and the formula of* $\mathbf{A}_{j,t}$ *are given below.*

The above examples both benefit from having a closed form distribution of $\mathbf{D}_{mis}|\mathbf{D}_{obs}$ and closed form representations of the elements in $\mathbf{A}_{i,t}$. Even if there is no closed form representations for those matrices, as long as we know the likelihood of the predictors, we can always use methods like the Metropolis-Hastings algorithm to obtain numerical approximations of those matrices.

Derivation of Example A.2.2

The likelihood function of $\mathbf{X}$ can be written as

$$l(\boldsymbol{\rho}|\mathbf{x}) = C' - \frac{n}{2}\log|\boldsymbol{\Sigma}_x| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_x)\boldsymbol{\Sigma}_x^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_x)^T,$$

where $C' = -(np\log 2\pi)/2$. Thus,

$$\begin{aligned}
& \mathbb{E}\{l(\boldsymbol{\rho}|\mathbf{x})|\mathbf{D}_{i,obs};\boldsymbol{\theta}_t\} \\
= \;& C' - \frac{n}{2}\log|\boldsymbol{\Sigma}_x| - \frac{1}{2}\sum_{i=1}^{n}[\mathrm{tr}\{\boldsymbol{\Sigma}_x^{-1}\mathbb{E}(\mathbf{x}_i^T\mathbf{x}_i|\mathbf{D}_{i,obs};\boldsymbol{\theta}_t)\} + 2\boldsymbol{\mu}\boldsymbol{\Sigma}_x^{-1}\mathbb{E}(\mathbf{x}_i^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}_t) - \boldsymbol{\mu}\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}^T] \\
= \;& C' - \frac{n}{2}\log|\boldsymbol{\Sigma}_x| - \frac{1}{2}\{\mathrm{tr}(\boldsymbol{\Sigma}_x^{-1}\mathbf{A}_{3,t}) + 2\boldsymbol{\mu}\boldsymbol{\Sigma}_x^{-1}\mathbf{A}_{4,t} - n\boldsymbol{\mu}\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}^T\},
\end{aligned}$$

where $\mathbf{A}_{i4,t} = \mathbb{E}(\mathbf{X}_i|\boldsymbol{\theta}_t, \mathbf{D}_{i,obs})$ denote the conditional expectation of $\mathbf{X}_i$ given $\mathbf{D}_{i,obs}$. Let $\boldsymbol{\rho}_{t+1} = (\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{x,t+1})$. By Lemma A.3.1, we have $\boldsymbol{\mu}_{t+1} = \mathbf{A}_{4,t}^T/n$, and $\boldsymbol{\Sigma}_{x,t+1} = (\mathbf{A}_{3,t} - 2\mathbf{A}_{4,t}\boldsymbol{\mu}_{t+1})/n + \boldsymbol{\mu}_{t+1}^T\boldsymbol{\mu}_{t+1}$.

Then, we calculate $\mathbf{A}_{1,t}$, $\mathbf{A}_{2,t}$, $\mathbf{A}_{3,t}$. Since $\mathbf{X}_i$ and $\mathbf{Y}_i|\mathbf{X}_i$ are normaly distributed, following a similar derivation as in the Example A.2.1, given $\boldsymbol{\theta}_t$, $(\mathbf{X}_i^T, \mathbf{Y}_i^T)^T$ also follows a normal distribution with mean $(\boldsymbol{\mu}_{x,t}^T, \boldsymbol{\mu}_{x,t}^T\boldsymbol{\beta}_t^T)^T$ and covariance matrix

$$\tilde{\boldsymbol{\Sigma}}_t = \begin{pmatrix} \boldsymbol{\Sigma}_{x,t} & \boldsymbol{\Sigma}_{x,t}\boldsymbol{\beta}_t \\ \boldsymbol{\beta}_t^T\boldsymbol{\Sigma}_{x,t} & \boldsymbol{\Sigma}_t + \boldsymbol{\beta}_t^T\boldsymbol{\Sigma}_{x,t}\boldsymbol{\beta}_t \end{pmatrix}.$$

For simplicity, for the derivation of the parameter updates below, we only focus on the t^{th} step, and thus omit all the subscript t for the parameter updates. For different individuals, missing value occurs at different locations, so we rearrange $\mathbf{X}_i$, $\mathbf{Y}_i$ to separate missing variables from the observed variables. Write $(\mathbf{D}_{i,mis}^T, \mathbf{D}_{i,obs}^T)^T = \mathbf{B}_i(\mathbf{X}_i^T, \mathbf{Y}_i^T)^T$, where $\mathbf{B}_i$ is a permutation matrix. Thus, $(\mathbf{D}_{i,mis}^T, \mathbf{D}_{i,obs}^T)^T$ independently follows $N\{(\boldsymbol{\mu}_{i,1}^T, \boldsymbol{\mu}_{i,2}^T)^T, \begin{pmatrix} \boldsymbol{\Sigma}_{i1} & \boldsymbol{\Sigma}_{i2} \\ \boldsymbol{\Sigma}_{i2}^T & \boldsymbol{\Sigma}_{i3} \end{pmatrix}\}$, where $(\boldsymbol{\mu}_{i,1}^T, \boldsymbol{\mu}_{i,2}^T)^T = \mathbf{B}_i(\boldsymbol{\mu}_x^T, \boldsymbol{\mu}_x^T\boldsymbol{\beta}^T)^T$,

and $\begin{pmatrix} \boldsymbol{\Sigma}_{i1} & \boldsymbol{\Sigma}_{i2} \\ \boldsymbol{\Sigma}_{i2}^T & \boldsymbol{\Sigma}_{i3} \end{pmatrix} = \mathbf{B}_i \tilde{\boldsymbol{\Sigma}} \mathbf{B}_i^T$. Hence, $\mathbf{D}_{i,mis}|\mathbf{D}_{i,obs}$ independently follows $N\{\boldsymbol{\mu}_{i,1} + \boldsymbol{\Sigma}_{i2}\boldsymbol{\Sigma}_{i3}^{-1}(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,2}), \boldsymbol{\Sigma}_{i1} - \boldsymbol{\Sigma}_{i2}\boldsymbol{\Sigma}_{i3}^{-1}\boldsymbol{\Sigma}_{i2}^T\}$. Therefore,

$$\mathbb{E}(\mathbf{D}_{i,mis}|\mathbf{D}_{i,obs};\boldsymbol{\theta}) = \boldsymbol{\mu}_{i,1} + \boldsymbol{\Sigma}_{i2}\boldsymbol{\Sigma}_{i3}^{-1}(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,2}),$$

$$\mathbb{E}(\mathbf{D}_{i,mis}\mathbf{D}_{i,obs}^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}) = \mathbb{E}(\mathbf{D}_{i,mis}|\mathbf{D}_{i,obs};\boldsymbol{\theta})\mathbf{D}_{i,obs}^T = \{\boldsymbol{\mu}_{i,1}+\boldsymbol{\Sigma}_{i2}\boldsymbol{\Sigma}_{i3}^{-1}(\mathbf{D}_{i,obs}-\boldsymbol{\mu}_{i,2})\}\mathbf{D}_{i,obs}^T,$$

and

$$\begin{aligned} & \mathbb{E}(\mathbf{D}_{i,mis}\mathbf{D}_{i,mis}^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}) \\ = \; & \mathbb{E}(\mathbf{D}_{i,mis}|\mathbf{D}_{i,obs};\boldsymbol{\theta})\mathbb{E}(\mathbf{D}_{i,mis}|\mathbf{D}_{i,obs};\boldsymbol{\theta})^T + \mathrm{Var}(\mathbf{D}_{i,mis}|\mathbf{D}_{i,obs};\boldsymbol{\theta}) \\ = \; & \boldsymbol{\mu}_{i,1} + \boldsymbol{\Sigma}_{i2}\{\boldsymbol{\Sigma}_{i3}^{-1}(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,2})\}\{\boldsymbol{\mu}_{i,1} + \boldsymbol{\Sigma}_{i2}\boldsymbol{\Sigma}_{i3}^{-1}(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,2})\}^T + \boldsymbol{\Sigma}_{i1} - \boldsymbol{\Sigma}_{i2}\boldsymbol{\Sigma}_{i3}^{-1}\boldsymbol{\Sigma}_{i2}^T. \end{aligned}$$

Then, we can obtain $\mathbf{A}_{i1}$, $\mathbf{A}_{i2}$ and $\mathbf{A}_{i3}$ through

$$\begin{aligned} \mathbb{E}\{(\mathbf{X}_i^T, \mathbf{Y}_i^T)^T(\mathbf{X}_i^T, \mathbf{Y}_i^T)|\mathbf{D}_{i,obs};\boldsymbol{\theta}\} &= \begin{pmatrix} \mathbb{E}(\mathbf{X}_i\mathbf{X}_i^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}) & \mathbb{E}(\mathbf{X}_i\mathbf{Y}_i^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}) \\ \mathbb{E}(\mathbf{Y}_i\mathbf{X}_i^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}) & \mathbb{E}(\mathbf{Y}_i\mathbf{Y}_i^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}) \end{pmatrix} \\ = \begin{pmatrix} \mathbf{A}_{i3} & \mathbf{A}_{i2}^T \\ \mathbf{A}_{i2} & \mathbf{A}_{i1} \end{pmatrix} &= \mathbf{B}_i^T \begin{pmatrix} \mathbb{E}(\mathbf{D}_{i,mis}\mathbf{D}_{i,mis}^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}) & \mathbb{E}(\mathbf{D}_{i,mis}\mathbf{D}_{i,obs}^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}) \\ \mathbb{E}(\mathbf{D}_{i,obs}\mathbf{D}_{i,mis}^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}) & \mathbb{E}(\mathbf{D}_{i,obs}\mathbf{D}_{i,obs}^T|\mathbf{D}_{i,obs};\boldsymbol{\theta}) \end{pmatrix} \mathbf{B}_i \end{aligned}$$

The last equation holds because for a permutation matrix $\mathbf{B}_i$, we have $\mathbf{B}_i^{-1} = \mathbf{B}_i^T$. After obtaining $\mathbf{A}_{i1}$, $\mathbf{A}_{i2}$ and $\mathbf{A}_{i3}$, we can obtain $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}_3$ through a summation over i.

Proof of Example A.2.3

Let $\boldsymbol{\beta}_{i,obs}$ denote the submatrix of $\boldsymbol{\beta}$ with the rows corresponds to the observed responses $\mathbf{Y}_{i,obs}$. Let $\boldsymbol{\Sigma}_{i,obs}$ denote the submatrix of $\boldsymbol{\Sigma}$ with the elements corresponds to the covariance of $\mathbf{Y}_{i,obs}$. Let $\boldsymbol{\varepsilon}_{i,obs}$ denote the random error corresponds to $\mathbf{Y}_{i,obs}$. Hence, we have $\mathbf{Y}_{i,obs} = \boldsymbol{\beta}_{i,obs}X_i + \boldsymbol{\varepsilon}_{i,obs}$ where $\boldsymbol{\varepsilon}_{i,obs}$ independently follows $N(\mathbf{0}, \boldsymbol{\Sigma}_{i,obs})$.

First, we derive the distribution of $X_i|\mathbf{Y}_{i,obs}$ given $\boldsymbol{\theta} = \boldsymbol{\theta}_t$.

$$\begin{aligned}
& f(x_i|\mathbf{y}_{i,obs};\boldsymbol{\theta}_t) \\
\propto\ & f(x_i,\mathbf{y}_{i,obs};\boldsymbol{\theta}_t) \\
=\ & \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_{i,obs,t}^{\frac{1}{2}}|}\exp\left\{-\frac{1}{2}(\mathbf{y}_{i,obs}-x_i\boldsymbol{\beta}_{i,obs,t})\boldsymbol{\Sigma}_{i,obs,t}^{-1}(\mathbf{y}_{i,obs}-x_i\boldsymbol{\beta}_{i,obs,t})^T\right\}\pi^{x_i}(1-\pi)^{1-x_i} \\
\propto\ & \exp\left\{-\frac{1}{2}x_i\boldsymbol{\beta}_{i,obs,t}\boldsymbol{\Sigma}_{i,obs,t}^{-1}\boldsymbol{\beta}_{i,obs,t}^T x_i^T + \mathbf{y}_{i,obs}\boldsymbol{\Sigma}_{i,obs,t}^{-1}\boldsymbol{\beta}_{i,obs,t}^T x_i^T\right\}(\frac{\pi}{1-\pi})^{x_i} \\
=\ & \Big[\frac{\pi\exp\{\boldsymbol{\beta}_{i,obs,t}\boldsymbol{\Sigma}_{i,obs,t}^{-1}\mathbf{y}_{i,obs}^T - \boldsymbol{\beta}_{i,obs,t}\boldsymbol{\Sigma}_{i,obs,t}^{-1}\boldsymbol{\beta}_{i,obs,t}^T/2\}}{1-\pi}\Big]^{x_i}.
\end{aligned}$$

The last equation holds because for a Bernoulli variable, we have $x_i^2 = x_i$. Then, $X_i|(\mathbf{Y}_{i,obs} = \mathbf{y}_{i,obs})$ follows a Bernoulli distribution with parameter $\frac{\pi_t q_t}{1-\pi_t+\pi_t q_t}$, where
$q_t = \exp\{\boldsymbol{\beta}_{i,obs,t}\boldsymbol{\Sigma}_{i,obs,t}^{-1}\mathbf{y}_{i,obs}^T - \boldsymbol{\beta}_{i,obs,t}\boldsymbol{\Sigma}_{i,obs,t}^{-1}\boldsymbol{\beta}_{i,obs,t}^T/2\}$.

The likelihood function of $\mathbf{X}$ can be written as

$$l(\boldsymbol{\rho}|\mathbf{x}) = \sum_{i=1}^{n} x_i\log\pi + (n-\sum_{i=1}^{n}x_i)\log(1-\pi).$$

Hence,

$$\begin{aligned}
& \mathbb{E}\{l(\boldsymbol{\rho}|\mathbf{X})|\mathbf{D}_{i,obs};\boldsymbol{\theta}_t\} \\
=\ & \sum_{i=1}^{n}\mathbb{E}(X_i|\mathbf{D}_{i,obs};\boldsymbol{\theta}_t)\log\pi + \{n-\sum_{i=1}^{n}\mathbb{E}(X_i|\mathbf{D}_{i,obs};\boldsymbol{\theta}_t)\}\log(1-\pi).
\end{aligned}$$

For an individual i, if X_i is observed, $\mathbb{E}(X_i|\mathbf{D}_{i,obs}) = X_i$, and $\mathbb{E}(X_i|\mathbf{D}_{i,obs}) = \frac{\pi_t q_t}{1-\pi_t+\pi_t q_t}$ if otherwise. Denote $\tilde{\pi}_i = \left(\frac{\pi_t q_t}{1-\pi_t+\pi_t q_t}\right)^{1-R_{X_i}} X_i^{R_{X_i}}$, we have $\mathbb{E}\{l(\boldsymbol{\rho}|\mathbf{X})|\mathbf{D}_{i,obs};\boldsymbol{\theta}_t\} = \sum_{i=1}^{n}\tilde{\pi}_{i,t}\log\pi + (n-\sum_{i=1}^{n}\tilde{\pi}_{i,t})\log(1-\pi)$.

By taking derivative with regard to π, we get the update of parameter $\pi_{t+1} = \sum_{i=1}^{n}\tilde{\pi}_{i,t}/n$.

For simplicity, we again omit the subscript t in the following derivation. Next, we calculate the conditional covariance matrices $\mathbf{A}_1, \mathbf{A}_2, A_3$. For an individual i, if x_i is not missing, $\mathbf{A}_{i1}$, $\mathbf{A}_{i2}$ and A_{i3} can be computed trivially. Hence, we only need to demonstrate the case when x_i is missing. There exist a permutation matrix $\mathbf{B}_i$, such

that $(\mathbf{Y}_{i,mis}^T, \mathbf{Y}_{i,obs}^T)^T = \mathbf{B}_i\mathbf{Y}_i$. Then, $\text{Var}(\mathbf{y}_{i,mis}^T, \mathbf{y}_{i,obs}^T)^T = \mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_i^T = \begin{pmatrix} \boldsymbol{\Sigma}_{i1} & \boldsymbol{\Sigma}_{i2} \\ \boldsymbol{\Sigma}_{i2}^T & \boldsymbol{\Sigma}_{i3} \end{pmatrix}$, where $\boldsymbol{\Sigma}_{i1} = \text{Var}(\mathbf{Y}_{i,mis})$, $\boldsymbol{\Sigma}_{i2} = \text{Cov}(\mathbf{Y}_{i,mis}, \mathbf{Y}_{i,mis})$, and $\boldsymbol{\Sigma}_{i3} = \text{Var}(\mathbf{Y}_{i,obs})$.

Because $\mathbf{A}_{i1} = \mathbf{B}_i^T \begin{pmatrix} \mathbb{E}(\mathbf{Y}_{i,mis}^T\mathbf{Y}_{i,mis}|\mathbf{y}_{i,obs};\boldsymbol{\theta}) & \mathbb{E}(\mathbf{Y}_{i,mis}^T|\mathbf{y}_{i,obs};\boldsymbol{\theta})\mathbf{y}_{i,obs} \\ \mathbf{y}_{i,obs}^T\mathbb{E}(\mathbf{Y}_{i,mis}|\mathbf{y}_{i,obs};\boldsymbol{\theta}) & \mathbf{y}_{i,obs}^T\mathbf{y}_{i,obs} \end{pmatrix} \mathbf{B}_i$, we only need to compute $\mathbb{E}(\mathbf{Y}_{i,mis}^T\mathbf{Y}_{i,mis}|\mathbf{y}_{i,obs};\boldsymbol{\theta})$ and $\mathbb{E}(\mathbf{Y}_{i,mis}^T|\mathbf{y}_{i,obs};\boldsymbol{\theta})$. Since $\mathbb{E}(\mathbf{Y}_{i,mis}^T|\mathbf{y}_{i,obs};\boldsymbol{\theta}) = \{\boldsymbol{\beta}_{0,mis} + \tilde{\pi}_i\boldsymbol{\beta}_{i,mis} + (\mathbf{y}_{i,obs} - \boldsymbol{\beta}_{0,obs} - \tilde{\pi}_i\boldsymbol{\beta}_{i,obs})\boldsymbol{\Sigma}_{i3}^{-1}\boldsymbol{\Sigma}_{i2}^T\}^T$, and $\mathbb{E}(\mathbf{Y}_{i,mis}^T\mathbf{Y}_{i,mis}|\mathbf{y}_{i,obs};\boldsymbol{\theta}) = \boldsymbol{\Sigma}_{i1} - \boldsymbol{\Sigma}_{i2}\boldsymbol{\Sigma}_{i3}^{-1}\boldsymbol{\Sigma}_{i2}^T + \mathbb{E}(\mathbf{y}_{i,mis}^T|\mathbf{y}_{i,obs},\boldsymbol{\theta})\mathbb{E}(\mathbf{y}_{i,mis}|\mathbf{y}_{i,obs};\boldsymbol{\theta})$, $\mathbf{A}_{i1}$ can be obtained.

To calculate $\mathbf{A}_{i2}$, by the law of total expectation, we have

$$\begin{aligned}
\mathbf{A}_{i2} &= \mathbb{E}(\mathbf{Y}_iX_i|\mathbf{D}_{i,obs};\boldsymbol{\theta}) \\
&= \mathbb{E}\{\mathbb{E}(\mathbf{Y}_iX_i|X_i,\mathbf{y}_{i,obs};\boldsymbol{\theta})|\mathbf{y}_{i,obs};\boldsymbol{\theta}\} \\
&= \mathbf{B}_i^T\mathbb{E}[\mathbb{E}\{(\mathbf{Y}_{i,mis}^T,\mathbf{Y}_{i,obs}^T)^T|X_i,\mathbf{y}_{i,obs};\boldsymbol{\theta}\}X_i|\mathbf{y}_{i,obs};\boldsymbol{\theta}] \\
&= \mathbf{B}_i^T\mathbb{E}[\{\boldsymbol{\beta}_{i,mis}X_i + (\mathbf{y}_{i,obs}X_i - \boldsymbol{\Sigma}_{i2}\boldsymbol{\Sigma}_{i3}^{-1}\boldsymbol{\beta}_{i,obs}X_i), \mathbf{y}_{i,obs}X_i\}^T|\mathbf{y}_{i,obs};\boldsymbol{\theta}] \\
&= \mathbf{B}_i^T\{\boldsymbol{\beta}_{i,mis}\tilde{\pi}_i + (\mathbf{y}_{i,obs}\tilde{\pi}_i - \boldsymbol{\Sigma}_{i2}\boldsymbol{\Sigma}_{i3}^{-1}\boldsymbol{\beta}_{i,obs}\tilde{\pi}_i), \mathbf{y}_{i,obs}\tilde{\pi}_i\}^T.
\end{aligned}$$

Since $X_i|\mathbf{y}_{i,obs}$ follows Bernoulli distribution with parameter $\tilde{\pi}_i$, we have $A_{i3} = \tilde{\pi}_i$. After obtaining $\mathbf{A}_{i1}$, $\mathbf{A}_{i2}$ and $\mathbf{A}_{i3}$, we can obtain $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}_3$ through a summation over i.

A.3 Lemma and algorithms

Review of Lemma 4.3 in Cook et al. (2010)

Lemma A.3.1 *Let $\mathscr{B}$ denote the set of all positive semi-definite matrices in $\mathbb{R}^{r\times r}$ having the same column dimension k, $0 < k \leq r$, and let $\mathbf{P}$ be the projection onto the common column space. Let $\mathbf{U}$ be a matrix in $\mathbb{R}^{n\times r}$ and let $l(\mathbf{B}) = -n\text{det}_0(\mathbf{B}) - \text{tr}(\mathbf{U}\mathbf{B}^\dagger\mathbf{U}^T)$. Then, the optimizer of $l(\mathbf{B})$ over $\mathscr{B}$ is the matrix $n^{-1}\mathbf{P}\mathbf{U}^T\mathbf{U}\mathbf{P}$, and the maximum value of $l(\mathbf{B})$ is $nk\text{log}n - nk - n\text{det}_0(\mathbf{P}\mathbf{U}^T\mathbf{U}\mathbf{P})$.*

The 1-D algorithm

Cook and Zhang (2016) proposed the 1-D algorithm to calculate the envelope estimates. We review it as follows:

Algorithm 3: The 1-D algorithm

1. Initialization: $\mathbf{g}_0 = \mathbf{G}_0 = 0$;
2. For $k = 0, 1, ..., u-1$,
   (a) Let $\mathbf{G}_k = (\mathbf{g}_1, ..., \mathbf{g}_k)$ if $k \geq 1$ and let $(\mathbf{G}_k, \mathbf{G}_{0k})$ be an orthogonal basis for $\mathbb{R}^r$.
   (b) Define the stepwise objective function
$$D_k(\mathbf{w}) = \log(\mathbf{w}^T \mathbf{M}_k \mathbf{w}) + \log\{\mathbf{w}^T (\mathbf{M}_k + \mathbf{U}_k)^{-1} \mathbf{w}\},$$
where $\mathbf{M}_k = \mathbf{G}_{0k}^T (\mathbf{A}_{1,t} - \mathbf{A}_{2,t}\mathbf{A}_{3,t}^{-1}\mathbf{A}_{2,t}^T)\mathbf{G}_{0k}$, $\mathbf{U}_k = \mathbf{G}_{0k}^T \mathbf{A}_{2,t}\mathbf{A}_{3,t}^{-1}\mathbf{A}_{2,t}^T \mathbf{G}_{0k}$ and $\mathbf{w} \in \mathbb{R}^{r-k}$.
   (c) Solve $\mathbf{w}_{k+1} = \arg\min_w D_k(\mathbf{w})$ subject to a length constraint $\mathbf{w}^T\mathbf{w} = 1$.
   (d) Define $\mathbf{g}_{k+1} = \mathbf{G}_{0k}\mathbf{w}_{k+1}$ to be the unit length $(k+1)$th stepwise direction.

The EM envelope algorithm

We summarize the EM envelope algorithm as follows, where δ can be chosen depending on the accuracy to achieve.

Algorithm 4: The EM envelope algorithm

for *k = 1, 2, ..., u* **do**
  Initialization: $t = 0$, $\boldsymbol{\Sigma}_0 = \mathbf{I}_q$, $\boldsymbol{\beta}_0 = \mathbf{0}$, $\boldsymbol{\theta}_0 = (\boldsymbol{\Sigma}_{1,0}, \boldsymbol{\Sigma}_{2,0}, \boldsymbol{\eta}_0, \boldsymbol{\Gamma}_0, \boldsymbol{\rho}_0)$, $\boldsymbol{\rho}_0 = (\boldsymbol{\rho}_{0\boldsymbol{\mu}_x}, \boldsymbol{\rho}_{0\boldsymbol{\Sigma}_x})$, $\boldsymbol{\rho}_{0\boldsymbol{\mu}_x} = \mathbf{0}$, $\boldsymbol{\rho}_{0\boldsymbol{\Sigma}_x} = \mathbf{I}_p$, $\Delta_0 = \infty$.
  while $\Delta_t > \delta$ **do**
    1. Calculate $\mathbf{A}_{1,t} = \sum_{i=1}^n \mathbf{A}_{i1,t}$, $\mathbf{A}_{2,t} = \sum_{i=1}^n \mathbf{A}_{i2,t}$, $\mathbf{A}_{3,t} = \sum_{i=1}^n \mathbf{A}_{i3,t}$ based on $\boldsymbol{\theta}_t$;
    2. Using Algorithm 3 to calculate $\boldsymbol{\Gamma}_t$, then $\boldsymbol{\Sigma}_{1,t+1} = \mathbf{P}_{\boldsymbol{\Gamma}_t}(\mathbf{A}_{1,t} - \mathbf{A}_{2,t}\mathbf{A}_{3,t}^{-1}\mathbf{A}_{2,t}^T)\mathbf{P}_{\boldsymbol{\Gamma}_t}/n$;
    3. Update: $\boldsymbol{\rho}_{t+1} = \arg\max_{\boldsymbol{\rho}\in\boldsymbol{\Pi}} \mathbb{E}[\log\{f_x(\mathbf{x}_i|\boldsymbol{\rho})\}|\mathbf{D}_{obs}; \boldsymbol{\theta}_t]$, $\boldsymbol{\beta}_{t+1} = \mathbf{P}_{\boldsymbol{\Sigma}_{1,t+1}}\mathbf{A}_{2,t}\mathbf{A}_{3,t}^{-1}$, $\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_{1,t+1} + \mathbf{Q}_{\boldsymbol{\Gamma}_t}\mathbf{A}_{1,t}\mathbf{Q}_{\boldsymbol{\Gamma}_t}/n$;
    4. Set $\Delta_{t+1} = \|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\|_1$, $\boldsymbol{\theta}_{t+1} = (\boldsymbol{\Sigma}_{t+1}, \boldsymbol{\beta}_{t+1}, \boldsymbol{\rho}_{t+1})$, $t \leftarrow t+1$;
  end
  $\text{BIC}_{HQ,k} = -2Q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_t) + 2H(\boldsymbol{\theta}_t|\boldsymbol{\theta}_t) + pu\log n$, $\hat{\boldsymbol{\beta}}_k = \boldsymbol{\beta}_{t+1}$
end
Select k which minimize $\text{BIC}_{HQ,k}$. Corresponding $\boldsymbol{\beta}_k$ is the EM envelope estimator.

A.4 Additional tables and figure in Sections

Table A.1: Summary of MSE when $\mathbf{\Omega}_0 = 1000\mathbf{I}_q$

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
$\hat{\boldsymbol{\beta}}_{em\_env}$	1.64e-05	3.58e-05	4.44e-05	1.03e-03	5.70e-05	8.66e-02
$\hat{\boldsymbol{\beta}}_{cc\_env}$	3.80e-05	1.04e-04	2.00e-04	0.21	0.32	1.96
$\hat{\boldsymbol{\beta}}_{full\_env}$	3.90e-06	8.30e-06	1.02e-05	3.05e-02	1.23e-05	2.59
$\hat{\boldsymbol{\beta}}_{em\_std}$	2.37e-02	4.41e-02	5.34e-02	5.47e-02	6.38e-02	0.12
$\hat{\boldsymbol{\beta}}_{cc\_std}$	0.15	0.54	0.69	0.73	0.87	1.85
$\hat{\boldsymbol{\beta}}_{full\_std}$	1.99e-02	4.32e-02	5.23e-02	5.40e-02	6.23e-02	0.13

Table A.2: Summary of MSE when $\mathbf{\Omega}_0 = 10\boldsymbol{I}_q$

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
$\hat{\boldsymbol{\beta}}_{em\_env}$	4.54e-05	9.08e-05	1.06e-04	1.36e-04	1.25e-04	1.05e-03
$\hat{\boldsymbol{\beta}}_{cc\_env}$	2.16e-04	4.95e-04	6.16e-04	1.69e-03	9.42e-04	2.02e-02
$\hat{\boldsymbol{\beta}}_{full\_env}$	3.28e-05	7.32e-05	8.58e-05	9.36e-05	9.97e-05	1.10e-03
$\hat{\boldsymbol{\beta}}_{em\_std}$	2.17e-04	4.52e-04	5.42e-04	5.62e-04	6.49e-04	1.34e-03
$\hat{\boldsymbol{\beta}}_{cc\_std}$	1.49e-03	5.40e-03	6.81e-03	7.32e-03	8.80e-03	2.35e-02
$\hat{\boldsymbol{\beta}}_{full\_std}$	2.00e-04	4.33e-04	5.24e-04	5.40e-04	6.23e-04	1.28e-03

Table A.3: Summary of MSE under different distributions for $\boldsymbol{\varepsilon}_i$

error	method	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
t_5	$\hat{\boldsymbol{\beta}}_{em\_env}$	3.50e-04	8.94e-04	1.23e-03	1.31e-03	1.59e-03	4.44e-03
	$\hat{\boldsymbol{\beta}}_{em\_std}$	2.13e-03	5.84e-03	7.67e-03	8.24e-03	9.99e-03	3.83e-02
Uniform	$\hat{\boldsymbol{\beta}}_{em\_env}$	2.13e-05	6.40e-05	8.42e-05	8.90e-05	1.07e-04	3.21e-04
	$\hat{\boldsymbol{\beta}}_{em\_std}$	1.22e-04	3.21e-04	4.16e-04	4.37e-04	5.30e-04	1.27e-03
Laplace	$\hat{\boldsymbol{\beta}}_{em\_env}$	2.89e-04	1.10e-03	1.45e-03	1.54e-03	1.90e-03	4.24e-03
	$\hat{\boldsymbol{\beta}}_{em\_std}$	1.81e-03	7.33e-03	9.48e-03	1.02e-02	1.25e-02	2.63 e-02

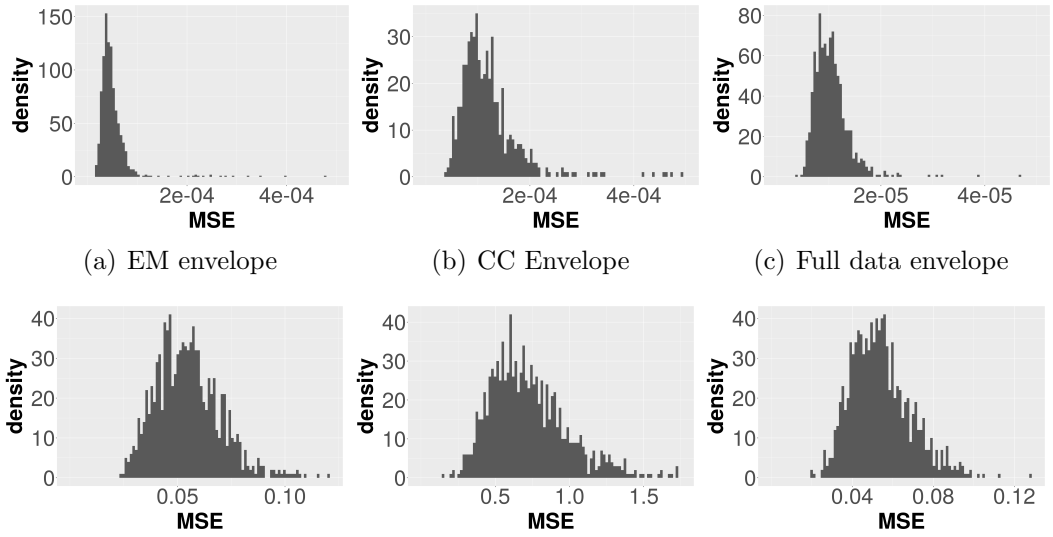
Table A.4: The point estimates, bootstrap standard errors, confidence intervals and p-values for the difference among patients with and without ESRD on biomarkers adjusted for the established biomarkers

	Our Method					Standard EM				
	$\hat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	2.5%	97.5%	p-value	$\hat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	2.5%	97.5%	p-value
log(Urine albumin)	-0.05	0.03	-0.12	3e-3	0.12	-0.09	0.05	-0.18	4e-3	0.06
Urine creatinine	-2.68	1.68	-5.97	0.55	0.11	-2.53	1.67	-5.79	0.70	0.13
log(HS_CRP)	-0.04	0.02	-0.07	-2e-3	0.05	-0.12	0.07	-0.28	0.02	0.10
log(BNP)	0.14	0.03	0.09	0.20	< 0.01	0.36	0.07	0.22	0.49	< 0.01
CXCL12	98.22	31.41	38.97	160.83	< 0.01	99.34	31.35	38.43	158.59	< 0.01
Scaled FETUIN_A	-0.85	0.64	-2.10	0.37	0.18	-0.85	0.63	-2.11	0.36	0.18
Fractalkine	0.05	8e-3	0.04	0.06	< 0.01	0.09	0.02	0.05	0.13	< 0.01
MPO	24.28	16.27	-7.13	59.23	0.14	22.32	16.81	-9.90	58.22	0.18
log(NGAL)	-0.01	0.03	-0.07	0.04	0.69	0.18	0.07	0.06	0.31	< 0.01
Fibrinogen	0.05	0.02	0.02	0.09	< 0.01	0.28	0.06	0.15	0.40	< 0.01
Troponini	4e-3	2e-3	3e-4	8e-3	0.06	5e-3	2e-3	1e-4	9e-3	0.04
log(Urine calcium)	-3e-3	0.02	-0.04	0.03	0.88	-0.03	0.06	-0.15	0.09	0.60
Urine sodium	-1.41	1.63	-4.58	1.89	0.39	-1.33	1.62	-4.49	1.86	0.41
Urine potassium	0.25	0.61	-0.96	1.46	0.68	0.18	0.60	-1.03	1.39	0.76
Urine phosphate	-0.36	0.93	-2.14	1.49	0.70	-0.28	0.92	-2.05	1.51	0.76
TNTHS	10.07	1.64	6.89	13.30	< 0.01	9.93	1.59	6.83	13.12	< 0.01
log(Aldosterone)	0.06	0.02	0.02	0.09	< 0.01	0.04	0.04	-0.04	0.13	0.31
C-peptide	-0.10	0.04	-0.17	-0.03	< 0.01	0.21	0.12	-0.02	0.44	0.08
Insulin	-2.12	1.25	-4.58	0.38	0.09	-2.08	1.25	-4.52	0.40	0.10
TOTAL PTH	27.29	4.81	18.43	37.26	< 0.01	27.16	4.78	18.31	36.96	< 0.01
CO_2	-0.04	0.05	-0.14	0.06	0.47	-0.24	0.18	-0.58	0.12	0.18

Table A.5: The point estimates, bootstrap standard errors, confidence intervals and p-values for the difference among patients with and without ESRD on biomarkers unadjusted for the established biomarkers

	Our Method					Standard EM				
	$\hat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	2.5%	97.5%	p-value	$\hat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	2.5%	97.5%	p-value
log(Urine albumin)	0.56	0.06	0.44	0.68	< 0.01	2.54	0.08	2.38	2.69	< 0.01
Urine creatinine	-11.98	1.33	-14.79	-9.30	< 0.01	-11.88	1.33	-14.69	-9.29	< 0.01
log(HS_CRP)	0.02	0.04	-0.04	0.11	0.54	-0.02	0.06	-0.12	0.10	0.76
log(BNP)	0.45	0.04	0.38	0.54	< 0.01	0.49	0.06	0.38	0.61	< 0.01
CXCL12	266.41	27.17	212.50	318.62	< 0.01	265.34	27.12	210.83	316.36	< 0.01
Scaled FETUIN_A	-0.69	0.51	-1.75	0.26	0.17	-0.72	0.51	-1.77	0.23	0.16
Fractalkine	0.16	0.01	0.14	0.18	< 0.01	0.22	0.02	0.19	0.26	< 0.01
MPO	43.04	16.99	11.20	78.69	0.01	43.07	16.95	11.28	78.69	0.01
log(NGAL)	0.30	0.06	0.14	0.38	< 0.01	0.83	0.06	0.73	0.95	< 0.01
Fibrinogen	0.29	0.04	0.23	0.39	< 0.01	0.76	0.05	0.65	0.88	< 0.01
Troponini	0.01	2e-3	3e-3	0.01	< 0.01	8e-3	3e-3	2e-3	0.01	< 0.01
log(Urine calcium)	-0.41	0.03	-0.47	-0.36	< 0.01	-0.58	0.045	-0.67	-0.48	< 0.01
Urine sodium	-7.51	1.33	-9.82	-4.82	< 0.01	-7.49	1.32	-9.78	-4.79	< 0.01
Urine potassium	-3.40	0.50	-4.40	-2.44	< 0.01	-3.33	0.4	-4.32	-2.37	< 0.01
Urine phosphate	-4.33	0.74	-5.77	-2.81	< 0.01	-4.34	0.73	-5.79	-2.87	< 0.01
TNTHS	20.22	1.64	17.19	23.58	< 0.01	20.12	1.63	17.12	23.48	< 0.01
log(Aldosterone)	0.08	0.02	0.04	0.13	< 0.01	0.14	0.03	0.08	0.21	< 0.01
C-peptide	0.37	0.06	0.24	0.49	< 0.01	0.64	0.10	0.45	0.84	< 0.01
Insulin	1.31	1.05	-0.74	3.37	0.21	1.27	1.05	-0.79	3.34	0.23
TOTAL PTH	54.48	4.68	46.19	64.22	< 0.01	54.42	4.69	46.11	64.22	< 0.01
CO_2	-0.99	0.19	-1.17	-0.80	< 0.01	-1.41	0.15	-1.69	-1.11	< 0.01
log(24-hour urine protein)	0.44	0.04	0.36	0.53	< 0.01	2.06	0.06	1.94	2.19	< 0.01
EGFR	-13.07	0.47	-13.98	-12.13	< 0.01	-12.95	0.47	-13.88	-12.00	< 0.01

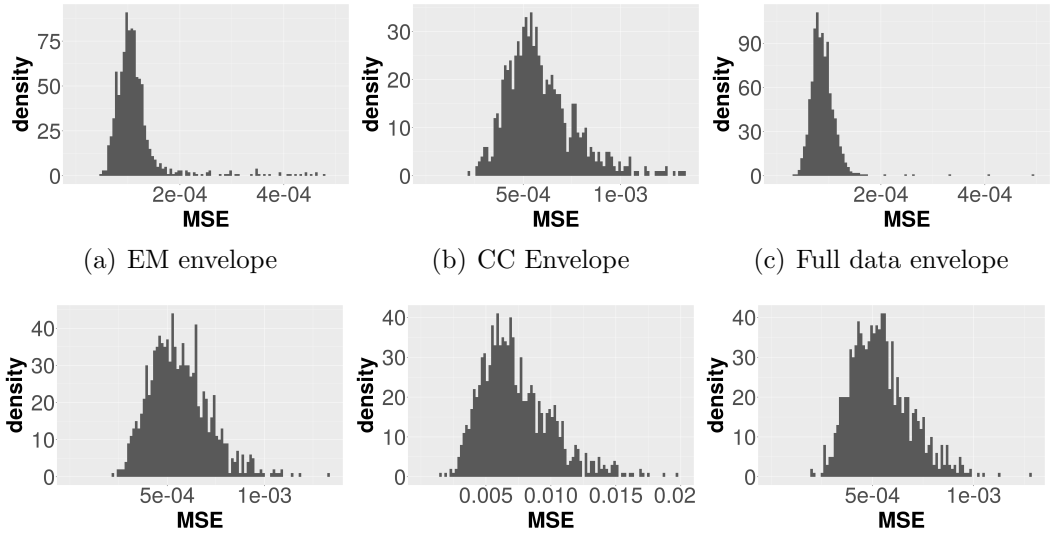
Figure A.1: Histograms of the MSEs of the EM envelope estimator, the complete case (CC) envelope estimator, the full data envelope estimator, the standard EM estimator, the standard complete case (CC) estimator and the full data MLE when $\mathbf{\Omega}_0 = 1000\mathbf{I}_q$.



(a) EM envelope (b) CC Envelope (c) Full data envelope

(d) Standard EM (e) Standard CC (f) Full data MLE

Figure A.2: Histograms of the MSEs of the EM envelope estimator, the complete case (CC) envelope estimator, the full data envelope estimator, the standard EM estimator, the standard complete case (CC) estimator and the full data MLE when $\mathbf{\Omega}_0 = 10\mathbf{I}_q$.



(a) EM envelope (b) CC Envelope (c) Full data envelope

(d) Standard EM (e) Standard CC (f) Full data MLE

Appendix B

Appendix for Chapter 3

B.1 Proof of Propositions

B.1.1 Proof of Proposition 3.2.1

Under model (3.3),

$$\widetilde{\mathbf{Y}}_i = \mathbf{1}_J \otimes \boldsymbol{\alpha} + (\mathbf{1}_J \otimes \boldsymbol{\beta})\mathbf{X}_i + \widetilde{\boldsymbol{\varepsilon}}_i, \tag{B.1}$$

Let

$$\widetilde{\boldsymbol{\Gamma}} = \frac{1}{\sqrt{J}} \begin{bmatrix} \mathbf{I}_r \\ \mathbf{I}_r \\ \vdots \\ \mathbf{I}_r \end{bmatrix}, \widetilde{\boldsymbol{\Gamma}}_0 = \frac{1}{\sqrt{J}} \begin{bmatrix} \mathbf{I}_r & \mathbf{I}_r & \cdots & \mathbf{I}_r \\ -\mathbf{I}_r & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_r & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & -\mathbf{I}_r \end{bmatrix}.$$

Notice that, $\widetilde{\boldsymbol{\Gamma}}_0^T \widetilde{\mathbf{Y}}_i = (\boldsymbol{\varepsilon}_{i1}^T - \boldsymbol{\varepsilon}_{i2}^T, \ldots, \boldsymbol{\varepsilon}_{i1}^T - \boldsymbol{\varepsilon}_{iJ}^T)^T$, and $\widetilde{\boldsymbol{\Gamma}}^T \widetilde{\mathbf{Y}}_i = \sqrt{J}(\boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \bar{\boldsymbol{\varepsilon}}_i)$, where $\bar{\boldsymbol{\varepsilon}}_i = \mathbf{b}_i + \sum_{j=1}^J \boldsymbol{\varepsilon}_{ij}/J$. We have $\widetilde{\boldsymbol{\Gamma}}_0^T \widetilde{\mathbf{Y}}_i \perp\!\!\!\perp \mathbf{X}_i$ and

$$\mathrm{Cov}(\widetilde{\boldsymbol{\Gamma}}_0^T \widetilde{\mathbf{Y}}_i, \widetilde{\boldsymbol{\Gamma}}^T \widetilde{\mathbf{Y}}_i) = \mathrm{Cov}(J\bar{\boldsymbol{\varepsilon}}_i^T, (\boldsymbol{\varepsilon}_{i1}^T - \boldsymbol{\varepsilon}_{i2}^T, \ldots, \boldsymbol{\varepsilon}_{i1}^T - \boldsymbol{\varepsilon}_{iJ}^T)^T) = \mathbf{0},$$

which indicates $\widetilde{\boldsymbol{\Gamma}}_0^T \widetilde{\mathbf{Y}}_i \perp\!\!\!\perp \widetilde{\boldsymbol{\Gamma}}^T \widetilde{\mathbf{Y}}_i | \mathbf{X}_i$. Hence, Conditions 3.1.1 and 3.1.2 are satisfied with $(\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Gamma}}_0)$. The envelope must be contained in $\mathrm{span}(\widetilde{\boldsymbol{\Gamma}})$, i.e., $\widetilde{u} \leq r$. Hence we proved Corollary 3.2.1. We can further find a semi-orthogonal matrix $\boldsymbol{\Phi}$ with maximum dimension such that $\boldsymbol{\Phi}_0^T \widetilde{\boldsymbol{\Gamma}}^T \mathbf{Y}_i \perp\!\!\!\perp (\boldsymbol{\Phi}^T \widetilde{\boldsymbol{\Gamma}}^T \mathbf{Y}_i, \mathbf{X}_i)$. Thus, by definition, $\widetilde{\boldsymbol{\Gamma}}\boldsymbol{\Phi} = \mathbf{1}_J \otimes \boldsymbol{\Phi}$ is the basis matrix for $\mathcal{E}_{\widetilde{\boldsymbol{\Sigma}}_\varepsilon}(\widetilde{\mathcal{B}})$.

B.1.2 Proof of Proposition 3.3.1

Under model (1), it is easy to verify that

$$\text{vec}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)|\mathbf{X}_i,\mathbf{Z}_i \sim N(\text{vec}(\boldsymbol{\Gamma}_0^T\boldsymbol{\beta}\mathbf{X}_i),\mathbf{M}_{22}),$$

$$\text{vec}(\boldsymbol{\Gamma}^T\mathbf{Y}_i)|\text{vec}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i),\mathbf{X}_i,\mathbf{Z}_i \sim N(\boldsymbol{\mu}^{**},\boldsymbol{\Sigma}^{**}),$$

where $\mathbf{M}_{11} = \mathbf{I}_{J_i}\otimes(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Gamma}) + (\mathbf{Z}_i^T\otimes\boldsymbol{\Gamma}^T)\boldsymbol{\Sigma}_b(\mathbf{Z}_i\otimes\boldsymbol{\Gamma})$, $\mathbf{M}_{12} = \mathbf{I}_{J_i}\otimes(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Gamma}_0) + (\mathbf{Z}_i^T\otimes\boldsymbol{\Gamma}^T)\boldsymbol{\Sigma}_b(\mathbf{Z}_i\otimes\boldsymbol{\Gamma}_0)$, $\mathbf{M}_{22} = \mathbf{I}_{J_i}\otimes(\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Gamma}_0) + (\mathbf{Z}_i^T\otimes\boldsymbol{\Gamma}_0^T)\boldsymbol{\Sigma}_b(\mathbf{Z}_i\otimes\boldsymbol{\Gamma}_0)$, $\boldsymbol{\mu}^{**} = \text{vec}(\boldsymbol{\Gamma}^T\boldsymbol{\beta}\mathbf{X}_i)+\mathbf{M}_{12}\mathbf{M}_{22}^{-1}\{\text{vec}(\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)-\text{vec}(\boldsymbol{\Gamma}_0^T\boldsymbol{\beta}\mathbf{X}_i)\}$, and $\boldsymbol{\Sigma}^{**} = \mathbf{M}_{11}-\mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{12}^T$.

Condition 1° holds if and only if the distribution of $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i|\mathbf{X}_i,\mathbf{Z}_i$ is free of $\mathbf{X}_i$ and $\mathbf{Z}_i$, that is, $\boldsymbol{\Gamma}_0^T\boldsymbol{\beta}=\mathbf{0}$, $\mathbf{Z}_i\otimes\boldsymbol{\Gamma}_0=\mathbf{0}$. Condition 2° holds if and only $\mathbf{M}_{12}=0$, that is, $\mathbf{I}_{J_i}\otimes(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Gamma}_0)+(\mathbf{Z}_i^T\otimes\boldsymbol{\Gamma}^T)\boldsymbol{\Sigma}_{\mathbf{b}}(\mathbf{Z}_i\otimes\boldsymbol{\Gamma}_0)=\mathbf{0}$.

B.1.3 Proof of Proposition 3.3.2

Let $\widehat{\boldsymbol{\theta}}_{obs\cdot em}$ denote the maximizer of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ and $\boldsymbol{\Sigma}_{\mathbf{b}}$ in the observed data likelihood in (3.5). Similarly, let $\widehat{\boldsymbol{\theta}}_{obs\cdot env}$ denote the maximizer of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ and $\boldsymbol{\Sigma}_{\mathbf{b}}$ in (3.5) under additional conditions (i)* and (ii)*. Let $\mathbf{V}_0 = \text{avar}(\widehat{\boldsymbol{\theta}}_{obs\cdot env})$ and $\mathbf{V} = \text{avar}(\widehat{\boldsymbol{\theta}}_{obs\cdot em})$ denote the asymptotic covariance matrices of the estimators obtained by directly maximizing (3.5) instead of using EM algorithm. Also, let $\boldsymbol{\phi} = (\boldsymbol{\eta},\boldsymbol{\Gamma},\boldsymbol{\Omega},\boldsymbol{\Omega}_0,\boldsymbol{\Sigma}_{\mathbf{b}})$ and $\boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}},\boldsymbol{\Sigma}_{\mathbf{b}})$ denote the parameter under the envelope model and the standard model. Let $\{\boldsymbol{\phi}_t\}$ and $\{\boldsymbol{\theta}_t\}$ denote the EM sequences, i.e., the parameters sequences we obtain from each EM iteration, of the envelope model and the standard model. By Corollary 1 of Wu (1983), the two EM sequences$\{\boldsymbol{\phi}_t\}$ and $\{\boldsymbol{\theta}_t\}$ converge to their unique maximizer of L_{obs}. Hence, in order to prove $\mathbf{V}_{mix\cdot env}\le\mathbf{V}_{mix\cdot em}$, it suffices to prove $\mathbf{V}_0\le\mathbf{V}$. We found function $\mathbf{h}$ such that $\mathbf{h}(\boldsymbol{\phi}) = (h_1(\boldsymbol{\phi}),h_2(\boldsymbol{\phi}),h_3(\boldsymbol{\phi}))$. Because of the over-parameterization of $\boldsymbol{\theta}$, the gradient matrix $\mathbf{G} = \dfrac{\partial\mathbf{h}(\boldsymbol{\phi})}{\partial\boldsymbol{\phi}^T}$ is not of full rank. By Proposition 4.1 in Shapiro (1986), we have

$$\mathbf{V}_0 = \mathbf{G}(\mathbf{G}^T\mathbf{V}^{-1}\mathbf{G})^\dagger\mathbf{G}^T.$$

$$\mathbf{V}-\mathbf{V}_0 = \mathbf{V}-\mathbf{G}(\mathbf{G}^T\mathbf{V}^{-1}\mathbf{G})^\dagger\mathbf{G}^T = \mathbf{V}^{\frac{1}{2}}\left[\mathbf{I}-\mathbf{V}^{-\frac{1}{2}}\mathbf{G}(\mathbf{G}^T\mathbf{V}^{-1}\mathbf{G})^\dagger\mathbf{G}^T\mathbf{V}^{-\frac{1}{2}}\right]\mathbf{V}^{\frac{1}{2}}.$$

Since $\mathbf{I}-\mathbf{V}^{-\frac{1}{2}}\mathbf{G}(\mathbf{G}^T\mathbf{V}^{-1}\mathbf{G})^\dagger\mathbf{G}^T\mathbf{V}^{-\frac{1}{2}}$ is the projection matrix onto the orthogonal complement of span$(\mathbf{V}^{-\frac{1}{2}}\mathbf{G})$, it is positive semi-definite. Hence, $\mathbf{V}_0\le\mathbf{V}$. In order

to find out the close form of $\mathbf{V}_0$. Because $\widehat{\boldsymbol{\theta}}_{obs\cdot em}$ is MLE, we can obtain $\mathbf{V}_0$ by inverting its Fisher information matrix.

The log-likelihood is

$$l(\boldsymbol{\theta}, \mathbf{Y}; \mathbf{X}, \mathbf{Z}) = C - \frac{1}{2}\sum_{i=1}^{n}\big[\log\det(\mathbf{I}_{J_i}\otimes\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} + \mathbf{A}_i\boldsymbol{\Sigma}_{\mathbf{b}}\mathbf{A}_i^T) - \{\mathrm{vec}(\mathbf{Y}_i) - \mathrm{vec}(\boldsymbol{\alpha}\otimes\mathbf{1}_{J_i}^T) - \mathrm{vec}(\boldsymbol{\beta}\mathbf{X}_i)\}^T$$
$$(\mathbf{I}_{J_i}\otimes\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} + \mathbf{A}_i\boldsymbol{\Sigma}_{\mathbf{b}}\mathbf{A}_i^T)^{-1}\{\mathrm{vec}(\mathbf{Y}_i) - \mathrm{vec}(\boldsymbol{\alpha}\otimes\mathbf{1}_{J_i}^T) - \mathrm{vec}(\boldsymbol{\beta}\mathbf{X}_i)\}\big]. \tag{B.2}$$

For calculating Fisher, we need the following result in Magus and Neudecker (1984):

Lemma B.1.1 *Let* $\mathbf{U}\in\mathbb{R}^{m\times p}$, $\mathbf{V}\in\mathbb{R}^{r\times s}$ *and* $\mathbf{X}\in\mathbb{R}^{n\times q}$, *then*

$$\frac{\partial vec(\mathbf{U}\otimes\mathbf{V})}{\partial vec^T(\mathbf{X})} = (\mathbf{I}_p\otimes\mathbf{G})\frac{\partial vec(\mathbf{U})}{\partial vec^T(\mathbf{X})} + (\mathbf{H}\otimes\mathbf{I}_r)\frac{\partial vec(\mathbf{V})}{\partial vec^T(\mathbf{X})},$$

where $\mathbf{G} = (\mathbf{K}_{sm}\otimes\mathbf{I}_r)(\mathbf{I}_m\otimes vec\mathbf{V})$, $\mathbf{H} = (\mathbf{I}_p\otimes\mathbf{K}_{sm})(vec(\mathbf{U})\otimes\mathbf{I}_s)$, *with* $\mathbf{K}_{sm}$ *being the commutation matrix in* $\mathbb{R}^{sm\times sm}$, *such that for any* $\mathbf{A}\in\mathbb{R}^{s\times m}$, $\mathbf{K}_{sm}vec(\mathbf{A}) = vec(\mathbf{A}^T)$.

Let l_i denote the log-likelihood for individual i and $\psi_i(\mathbf{Y}_i,\boldsymbol{\theta}) = \dfrac{\partial l_i}{\partial\boldsymbol{\theta}}$. Denote $\boldsymbol{\Sigma}_i = \mathbf{I}_{J_i}\otimes\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} + \mathbf{A}_i\boldsymbol{\Sigma}_{\mathbf{b}}\mathbf{A}_i^T\in\mathbb{R}^{J_ir\times J_ir}$, $\mathbf{D}_i = \mathrm{vec}(\mathbf{Y}_i) - \boldsymbol{\alpha}\otimes\mathbf{1}_{J_i} - \mathrm{vec}(\boldsymbol{\beta}\mathbf{X}_i)\in\mathbb{R}^{J_ir}$. We have $\mathbf{C}_r\mathbf{E}_r = \mathbf{I}_{r(r+1)/2}$. Also, denote $\mathbf{M}_{i1} = \dfrac{\partial\mathrm{vech}(\boldsymbol{\Sigma}_i)}{\partial\mathrm{vech}^T(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})}$, and $\mathbf{M}_{i2} = \dfrac{\partial\mathrm{vech}(\boldsymbol{\Sigma}_i)}{\partial\mathrm{vech}^T(\boldsymbol{\Sigma}_{\mathbf{b}})}$. By using Lemma B.1.1, we can also obtain the closed form for $\mathbf{M}_{i1}$:

$$\mathbf{M}_{i1} = \frac{\partial\mathrm{vech}(\boldsymbol{\Sigma}_i)}{\partial\mathrm{vech}^T(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})} = \mathbf{C}_{J_ir}\frac{\partial\mathrm{vec}(\boldsymbol{\Sigma}_i)}{\partial\mathrm{vec}^T(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})}\mathbf{E}_r = \mathbf{C}_{J_ir}\{(\mathbf{I}_{J_i}\otimes\mathbf{K}_{r,J_i})(\mathrm{vec}(\mathbf{I}_{J_i})\otimes\mathbf{I}_r)\otimes\mathbf{I}_r\}\mathbf{E}_r,$$

Also, the matrix $\mathbf{M}_{i2} = \mathbf{C}_{J_ir}(\mathbf{A}_i\otimes\mathbf{A}_i)\mathbf{E}_{qr}$.

Using the notation above, we have

$$l_i(\boldsymbol{\beta}, \mathbf{Y}_i; \mathbf{X}_i, \mathbf{Z}_i) = -\frac{1}{2}\log\det(\boldsymbol{\Sigma}_i) - \frac{1}{2}\mathbf{D}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{D}_i.$$

By matrix calculus, we have

$$\frac{\partial l_i}{\partial\mathrm{vec}^T(\boldsymbol{\beta})} = -\mathbf{D}_i^T\boldsymbol{\Sigma}_i^{-1}\frac{\partial\mathbf{D}_i}{\partial\mathrm{vec}^T(\boldsymbol{\beta})} = \mathbf{D}_i^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{X}_i^T\otimes\mathbf{I}_r),$$

$$\frac{\partial l_i}{\partial\mathrm{vech}^T(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})} = -\frac{1}{2}\mathrm{vec}^T(\boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_ir}\mathbf{M}_{i1} + \frac{1}{2}(\mathbf{D}_i^T\otimes\mathbf{D}_i^T)(\boldsymbol{\Sigma}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_ir}\mathbf{M}_{i1},$$

$$\frac{\partial l_i}{\partial \text{vech}^T(\mathbf{\Sigma_b})} = -\frac{1}{2}\text{vec}^T(\mathbf{\Sigma}_i^{-1})\mathbf{E}_{J_ir}\mathbf{M}_{i2} + \frac{1}{2}(\mathbf{D}_i^T \otimes \mathbf{D}_i^T)(\mathbf{\Sigma}_i^{-1} \otimes \mathbf{\Sigma}_i^{-1})\mathbf{E}_{J_ir}\mathbf{M}_{i2},$$

and

$$\psi_i(\mathbf{Y}_i, \boldsymbol{\theta}) = (\psi_{i1}^T, \psi_{i2}^T, \psi_{i3}^T)^T = (\frac{\partial l_i}{\partial \text{vec}^T(\boldsymbol{\beta})}, \frac{\partial l_i}{\partial \text{vech}^T(\mathbf{\Sigma_\varepsilon})}, \frac{\partial l_i}{\partial \text{vech}^T(\mathbf{\Sigma_b})})^T.$$

Then, we can calculate the expression for $\frac{\partial \psi_i}{\partial \boldsymbol{\theta}^T}$:

$$\frac{\partial \psi_{i1}}{\partial \text{vec}^T(\boldsymbol{\beta})} = -(\mathbf{X}_i \otimes \mathbf{I}_r)\mathbf{\Sigma}_i^{-1}(\mathbf{X}_i^T \otimes \mathbf{I}_r),$$

$$\begin{aligned}
\frac{\partial \psi_{i1}}{\partial \text{vech}^T(\mathbf{\Sigma_\varepsilon})} &= \{\mathbf{D}_i^T \otimes (\mathbf{X}_i \otimes \mathbf{I}_r)\}\frac{\partial \text{vec}(\mathbf{\Sigma}_i^{-1})}{\partial \text{vec}^T(\mathbf{\Sigma_\varepsilon})} \\
&= -\{\mathbf{D}_i^T \otimes (\mathbf{X}_i \otimes \mathbf{I}_r)\}(\mathbf{\Sigma}_i^{-1} \otimes \mathbf{\Sigma}_i^{-1})\mathbf{E}_{J_ir}\mathbf{M}_{i1}, \\
\frac{\partial \psi_{i1}}{\partial \text{vech}^T(\mathbf{\Sigma_b})} &= -\{\mathbf{D}_i^T \otimes (\mathbf{X}_i \otimes \mathbf{I}_r)\}(\mathbf{\Sigma}_i^{-1} \otimes \mathbf{\Sigma}_i^{-1})\mathbf{E}_{J_ir}\mathbf{M}_{i2}, \\
\frac{\partial \psi_{i2}}{\partial \text{vec}^T(\boldsymbol{\beta})} &= \frac{\partial}{\partial \text{vec}^T(\boldsymbol{\beta})}\left\{\frac{1}{2}\mathbf{M}_{i1}^T\mathbf{E}_{J_ir}^T(\mathbf{\Sigma}_i^{-1} \otimes \mathbf{\Sigma}_i^{-1})(\mathbf{D}_i \otimes \mathbf{D}_i)\right\} \\
&= -\frac{1}{2}\mathbf{M}_{i1}^T\mathbf{E}_{J_ir}^T(\mathbf{\Sigma}_i^{-1} \otimes \mathbf{\Sigma}_i^{-1})(\mathbf{I}_{rJ_i} \otimes \mathbf{D}_i + \mathbf{D}_i \otimes \mathbf{I}_{rJ_i})(\mathbf{X}_i^T \otimes \mathbf{I}_r), \\
\frac{\partial \psi_{i3}}{\partial \text{vec}^T(\boldsymbol{\beta})} &= -\frac{1}{2}\mathbf{M}_{i2}^T\mathbf{E}_{J_ir}^T(\mathbf{\Sigma}_i^{-1} \otimes \mathbf{\Sigma}_i^{-1})(\mathbf{I}_{rJ_i} \otimes \mathbf{D}_i + \mathbf{D}_i \otimes \mathbf{I}_{rJ_i})(\mathbf{X}_i^T \otimes \mathbf{I}_r).
\end{aligned}$$

In order to calculate $\frac{\partial \psi_{i2}}{\partial \text{vech}^T(\mathbf{\Sigma_\varepsilon})}$,

$$\begin{aligned}
\frac{\partial \psi_{i2}}{\partial \text{vech}^T(\mathbf{\Sigma_\varepsilon})} &= \frac{1}{2}\mathbf{M}_{i1}^T\mathbf{E}_{J_ir}^T(\mathbf{\Sigma}_i^{-1} \otimes \mathbf{\Sigma}_i^{-1})\mathbf{E}_{J_ir}\mathbf{M}_{i1} \\
&\quad + \frac{1}{2}\{(\mathbf{D}_i^T \otimes \mathbf{D}_i^T) \otimes (\mathbf{M}_{i1}^T\mathbf{E}_{J_ir}^T)\}\frac{\partial}{\partial \text{vech}^T(\mathbf{\Sigma_\varepsilon})}\text{vec}(\mathbf{\Sigma}_i^{-1} \otimes \mathbf{\Sigma}_i^{-1}).
\end{aligned}$$

By Lemma B.1.1, we have

$$\begin{aligned}
&\frac{\partial}{\partial \text{vech}^T(\mathbf{\Sigma_\varepsilon})}\text{vec}(\mathbf{\Sigma}_i^{-1} \otimes \mathbf{\Sigma}_i^{-1}) \\
=&[\mathbf{I}_{rJ_i} \otimes \{(\mathbf{K}_{rJ_i,rJ_i} \otimes \mathbf{I}_{rJ_i})(\mathbf{I}_{rJ_i} \otimes \text{vec}(\mathbf{\Sigma}_i^{-1}))\} + \{(\mathbf{I}_{rJ_i} \\
&\otimes \mathbf{K}_{rJ_i,rJ_i})(\text{vec}(\mathbf{\Sigma}_i^{-1}) \otimes \mathbf{I}_{rJ_i}) \otimes \mathbf{I}_{rJ_i}\}]\{-(\mathbf{\Sigma}_i^{-1} \otimes \mathbf{\Sigma}_i^{-1})\mathbf{E}_{rJ_i}\mathbf{M}_{i1}\}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\frac{\partial \psi_{i2}}{\partial \text{vech}^T(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})} \\
=&\frac{1}{2}\mathbf{M}_{i1}^T\mathbf{E}_{J_ir}^T(\boldsymbol{\Sigma}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_ir}\mathbf{M}_{i1} - \frac{1}{2}\{(\mathbf{D}_i^T\otimes\mathbf{D}_i^T)\otimes(\mathbf{M}_{i1}^T\mathbf{E}_{J_ir}^T)\}[\mathbf{I}_{rJ_i}\otimes\{(\mathbf{K}_{rJ_i,rJ_i}\otimes\mathbf{I}_{rJ_i})(\mathbf{I}_{rJ_i}\otimes \\
&\otimes\text{vec}(\boldsymbol{\Sigma}_i^{-1}))\} + \{(\mathbf{I}_{rJ_i}\otimes\mathbf{K}_{rJ_i,rJ_i})(\text{vec}(\boldsymbol{\Sigma}_i^{-1})\otimes\mathbf{I}_{rJ_i})\otimes\mathbf{I}_{rJ_i}\}]\{(\boldsymbol{\Sigma}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{rJ_i}\mathbf{M}_{i1}\},
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\frac{\partial \psi_{i2}}{\partial \text{vech}^T(\boldsymbol{\Sigma}_{\mathbf{b}})} \\
=&\frac{1}{2}\mathbf{M}_{i1}^T\mathbf{E}_{J_ir}^T(\boldsymbol{\Sigma}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_ir}\mathbf{M}_{i2} - \frac{1}{2}\{(\mathbf{D}_i^T\otimes\mathbf{D}_i^T)\otimes(\mathbf{M}_{i1}^T\mathbf{E}_{J_ir}^T)\}[\mathbf{I}_{rJ_i}\otimes\{(\mathbf{K}_{rJ_i,rJ_i}\otimes\mathbf{I}_{rJ_i})(\mathbf{I}_{rJ_i}\otimes \\
&\otimes\text{vec}(\boldsymbol{\Sigma}_i^{-1}))\} + \{(\mathbf{I}_{rJ_i}\otimes\mathbf{K}_{rJ_i,rJ_i})(\text{vec}(\boldsymbol{\Sigma}_i^{-1})\otimes\mathbf{I}_{rJ_i})\otimes\mathbf{I}_{rJ_i}\}]\{(\boldsymbol{\Sigma}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{rJ_i}\mathbf{M}_{i2}\},
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial \psi_{i3}}{\partial \text{vech}^T(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})} \\
=&\frac{1}{2}\mathbf{M}_{i2}^T\mathbf{E}_{J_ir}^T(\boldsymbol{\Sigma}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_ir}\mathbf{M}_{i1} - \frac{1}{2}\{(\mathbf{D}_i^T\otimes\mathbf{D}_i^T)\otimes(\mathbf{M}_{i2}^T\mathbf{E}_{J_ir}^T)\}[\mathbf{I}_{rJ_i}\otimes\{(\mathbf{K}_{rJ_i,rJ_i}\otimes\mathbf{I}_{rJ_i})(\mathbf{I}_{rJ_i}\otimes \\
&\otimes\text{vec}(\boldsymbol{\Sigma}_i^{-1}))\} + \{(\mathbf{I}_{rJ_i}\otimes\mathbf{K}_{rJ_i,rJ_i})(\text{vec}(\boldsymbol{\Sigma}_i^{-1})\otimes\mathbf{I}_{rJ_i})\otimes\mathbf{I}_{rJ_i}\}]\{(\boldsymbol{\Sigma}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{rJ_i}\mathbf{M}_{i1}\},
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial \psi_{i3}}{\partial \text{vech}^T(\boldsymbol{\Sigma}_{\mathbf{b}})} \\
=&\frac{1}{2}\mathbf{M}_{i2}^T\mathbf{E}_{J_ir}^T(\boldsymbol{\Sigma}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_ir}\mathbf{M}_{i2} - \frac{1}{2}\{(\mathbf{D}_i^T\otimes\mathbf{D}_i^T)\otimes(\mathbf{M}_{i2}^T\mathbf{E}_{J_ir}^T)\}[\mathbf{I}_{rJ_i}\otimes\{(\mathbf{K}_{rJ_i,rJ_i}\otimes\mathbf{I}_{rJ_i})(\mathbf{I}_{rJ_i}\otimes \\
&\otimes\text{vec}(\boldsymbol{\Sigma}_i^{-1}))\} + \{(\mathbf{I}_{rJ_i}\otimes\mathbf{K}_{rJ_i,rJ_i})(\text{vec}(\boldsymbol{\Sigma}_i^{-1})\otimes\mathbf{I}_{rJ_i})\otimes\mathbf{I}_{rJ_i}\}]\{(\boldsymbol{\Sigma}_i^{-1}\otimes\boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{rJ_i}\mathbf{M}_{i2}\}.
\end{aligned}$$

Hence,

$$\frac{\partial \psi_i}{\partial \boldsymbol{\theta}^T} = \begin{pmatrix} \frac{\partial \psi_{i1}}{\partial \text{vec}^T(\boldsymbol{\beta})} & \frac{\partial \psi_{i1}}{\partial \text{vech}^T(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})} & \frac{\partial \psi_{i1}}{\partial \text{vech}^T(\boldsymbol{\Sigma}_{\mathbf{b}})} \\ \frac{\partial \psi_{i2}}{\partial \text{vec}^T(\boldsymbol{\beta})} & \frac{\partial \psi_{i2}}{\partial \text{vech}^T(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})} & \frac{\partial \psi_{i2}}{\partial \text{vech}^T(\boldsymbol{\Sigma}_{\mathbf{b}})} \\ \frac{\partial \psi_{i3}}{\partial \text{vec}^T(\boldsymbol{\beta})} & \frac{\partial \psi_{i3}}{\partial \text{vech}^T(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})} & \frac{\partial \psi_{i3}}{\partial \text{vech}^T(\boldsymbol{\Sigma}_{\mathbf{b}})} \end{pmatrix}.$$

We can obtain the Fisher information of $\boldsymbol{\theta}$ by taking expectation of $\frac{\partial \psi_i}{\partial \boldsymbol{\theta}^T}$ with respective to $\mathbf{Y}_i$

$$I_i(\boldsymbol{\theta}) = \begin{pmatrix} I_{11}^i & I_{12}^i & I_{13}^i \\ I_{21}^i & I_{22}^i & I_{23}^i \\ I_{31}^i & I_{32}^i & I_{33}^i \end{pmatrix},$$

where

$$I_{11}^i = -\mathbb{E}\{(-\mathbf{X}_i \otimes \mathbf{I}_r)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{X}_i \otimes \mathbf{I}_r) = (\mathbf{X}_i \otimes \mathbf{I}_r)\boldsymbol{\Sigma}_i^{-1}(\mathbf{X}_i^T \otimes \mathbf{I}_r)\},$$

$$I_{12}^i = -\mathbb{E}\{-\{\mathbf{D}_i^T \otimes (\mathbf{X}_i \otimes \mathbf{I}_r)\}(\boldsymbol{\Sigma}_i^{-1} \otimes \boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_i r}\mathbf{M}_{i1}\} = 0,$$

$$I_{13}^i = -\mathbb{E}\{-\{\mathbf{D}_i^T \otimes (\mathbf{X}_i \otimes \mathbf{I}_r)\}(\boldsymbol{\Sigma}_i^{-1} \otimes \boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_i r}\mathbf{M}_{i2}\} = 0,$$

By symmetry of $I_i(\boldsymbol{\theta})$, $I_{21}^i = I_{31}^i = \mathbf{0}$.

$$I_{22}^i = -\frac{1}{2}\mathbf{M}_{i1}^T\mathbf{E}_{J_i r}^T(\boldsymbol{\Sigma}_i^{-1} \otimes \boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_i r}\mathbf{M}_{i1} + \frac{1}{2}\{\mathrm{vec}^T(\boldsymbol{\Sigma}_i) \otimes (\mathbf{M}_{i1}^T\mathbf{E}_{J_i r}^T)\}[\mathbf{I}_{rJ_i} \otimes \{(\mathbf{K}_{rJ_i,rJ_i} \otimes \mathbf{I}_{rJ_i})(\mathbf{I}_{rJ_i} \otimes \mathrm{vec}(\boldsymbol{\Sigma}_i^{-1}))\} + \{(\mathbf{I}_{rJ_i} \otimes \mathbf{K}_{rJ_i,rJ_i})(\mathrm{vec}(\boldsymbol{\Sigma}_i^{-1}) \otimes \mathbf{I}_{rJ_i}) \otimes \mathbf{I}_{rJ_i}\}]\{(\boldsymbol{\Sigma}_i^{-1} \otimes \boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{rJ_i}\mathbf{M}_{i1}\},$$

$$I_{23}^i = -\frac{1}{2}\mathbf{M}_{i1}^T\mathbf{E}_{J_i r}^T(\boldsymbol{\Sigma}_i^{-1} \otimes \boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_i r}\mathbf{M}_{i2} + \frac{1}{2}\{\mathrm{vec}^T(\boldsymbol{\Sigma}_i) \otimes (\mathbf{M}_{i1}^T\mathbf{E}_{J_i r}^T)\}[\mathbf{I}_{rJ_i} \otimes \{(\mathbf{K}_{rJ_i,rJ_i} \otimes \mathbf{I}_{rJ_i})(\mathbf{I}_{rJ_i} \otimes \mathrm{vec}(\boldsymbol{\Sigma}_i^{-1}))\} + \{(\mathbf{I}_{rJ_i} \otimes \mathbf{K}_{rJ_i,rJ_i})(\mathrm{vec}(\boldsymbol{\Sigma}_i^{-1}) \otimes \mathbf{I}_{rJ_i}) \otimes \mathbf{I}_{rJ_i}\}]\{(\boldsymbol{\Sigma}_i^{-1} \otimes \boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{rJ_i}\mathbf{M}_{i2}\},$$

$$I_{32}^i = -\frac{1}{2}\mathbf{M}_{i2}^T\mathbf{E}_{J_i r}^T(\boldsymbol{\Sigma}_i^{-1} \otimes \boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_i r}\mathbf{M}_{i1} + \frac{1}{2}\{\mathrm{vec}^T(\boldsymbol{\Sigma}_i) \otimes (\mathbf{M}_{i2}^T\mathbf{E}_{J_i r}^T)\}[\mathbf{I}_{rJ_i} \otimes \{(\mathbf{K}_{rJ_i,rJ_i} \otimes \mathbf{I}_{rJ_i})(\mathbf{I}_{rJ_i} \otimes \mathrm{vec}(\boldsymbol{\Sigma}_i^{-1}))\} + \{(\mathbf{I}_{rJ_i} \otimes \mathbf{K}_{rJ_i,rJ_i})(\mathrm{vec}(\boldsymbol{\Sigma}_i^{-1}) \otimes \mathbf{I}_{rJ_i}) \otimes \mathbf{I}_{rJ_i}\}]\{(\boldsymbol{\Sigma}_i^{-1} \otimes \boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{rJ_i}\mathbf{M}_{i1}\},$$

$$I_{33}^i = -\frac{1}{2}\mathbf{M}_{i2}^T\mathbf{E}_{J_i r}^T(\boldsymbol{\Sigma}_i^{-1} \otimes \boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_i r}\mathbf{M}_{i2} + \frac{1}{2}\{\mathrm{vec}^T(\boldsymbol{\Sigma}_i) \otimes (\mathbf{M}_{i2}^T\mathbf{E}_{J_i r}^T)\}[\mathbf{I}_{rJ_i} \otimes \{(\mathbf{K}_{rJ_i,rJ_i} \otimes \mathbf{I}_{rJ_i})(\mathbf{I}_{rJ_i} \otimes \mathrm{vec}(\boldsymbol{\Sigma}_i^{-1}))\} + \{(\mathbf{I}_{rJ_i} \otimes \mathbf{K}_{rJ_i,rJ_i})(\mathrm{vec}(\boldsymbol{\Sigma}_i^{-1}) \otimes \mathbf{I}_{rJ_i}) \otimes \mathbf{I}_{rJ_i}\}]\{(\boldsymbol{\Sigma}_i^{-1} \otimes \boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{rJ_i}\mathbf{M}_{i2}\},$$

this is because $\mathbb{E}(\mathbf{D}_i) = 0$ and $\mathbb{E}(\mathbf{D}_i \otimes \mathbf{D}_i) = \mathrm{vec}(\boldsymbol{\Sigma}_\mathbf{b})$.

The $I_i(\boldsymbol{\theta})$ we obtained is when assuming $\mathbf{X}_i$ and $\mathbf{Z}_i$ are fixed. Since $\widehat{\boldsymbol{\theta}}_{mix\cdot em}$ is the MLE with regularity conditions satisfied, we have

$$\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{mix\cdot em}) = \left\{\sum_{i=1}^n I_i(\boldsymbol{\theta})\right\}^{-1}.$$

Let $\boldsymbol{\phi} = (\mathrm{vec}^T(\boldsymbol{\eta}), \mathrm{vec}^T(\boldsymbol{\Gamma}), \mathrm{vech}^T(\boldsymbol{\Omega}), \mathrm{vech}^T(\boldsymbol{\Omega}_0), \mathrm{vech}^T(\boldsymbol{\Sigma}_\mathbf{b}))^T$, $\mathbf{h}(\boldsymbol{\phi}) = (\mathrm{vec}^T(\boldsymbol{\Gamma}\boldsymbol{\eta}), \mathrm{vech}^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T), \mathrm{vech}^T(\boldsymbol{\Sigma}_\mathbf{b}))^T$, and $\mathbf{G} = \dfrac{\partial \mathbf{h}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}^T}$. Then, we have

$$\mathbf{G} = \begin{pmatrix} \mathbf{I}_p \otimes \boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \mathbf{I}_r & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{C}_r(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \mathbf{I}_r - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) & \mathbf{C}_r(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_r(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{r-u} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_{qr(qr+1)/2} \end{pmatrix}.$$

Hence, $\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{mix\cdot env}) = \mathbf{V}_0 = \mathbf{G}(\mathbf{G}^T\mathbf{V}\mathbf{G})^\dagger\mathbf{G}^T$, where $\mathbf{V} = \sum_{i=1}^n I_i(\boldsymbol{\theta})$.

B.1.4 Proof of Proposition 3.3.3

Since the mixed effects envelope model is overparameterized, we will use Proposition 4.1 of Shapiro (1986) to prove Proposition 3.3.3. We will check their conditions. For convenience, we match Shapiro's notations in our context. Shapiro's $\mathbf{x}$ in our context is $\widehat{\boldsymbol{\theta}}_{mix\cdot em} = (\widehat{\boldsymbol{\beta}}_{mix\cdot em}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}\cdot mix\cdot em}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{b}\cdot mix\cdot em})$. We need to show the $\sqrt{n}$-consistency and asymptotical normality of $\mathbf{x}$. We assume the following regularity conditions: the error $\boldsymbol{\varepsilon}_{ij}$ and random effect $\mathbf{b}_i$ have finite $(4+\delta)$-th moment, $\sup_i \|\mathbf{X}_i\| < \infty$, $\sup_i \|\mathbf{Z}_i\| < \infty$, $\sup_i J_i < \infty$, $\inf_i \det(\mathbf{I}_{J_i} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} + \mathbf{A}_i \boldsymbol{\Sigma}_{\mathbf{b}} \mathbf{A}_i^T) > 0$, $\liminf_n \lambda_-[n^{-1}\mathrm{Var}(s_n(\boldsymbol{\theta}))] > 0$ and $\liminf_n \lambda_-[n^{-1}\mathbf{M}_n(\boldsymbol{\theta})] > 0$, where $\lambda_-[\mathbf{A}]$ denote the smallest eigenvalue of the matrix $\mathbf{A}$, $\mathbf{M}_n(\boldsymbol{\theta}) = -\mathbb{E}\Big\{\dfrac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\Big\}$, and l is the log-likelihood when the error is normally distributed.

The estimator $\widehat{\boldsymbol{\theta}}_{mix\cdot em}$ is obtained by maximizing the following misspecified log-likelihood:

$$l(\boldsymbol{\theta}, \mathbf{Y}; \mathbf{X}, \mathbf{Z}) = C - \frac{1}{2}\sum_{i=1}^{n} \big[\log\det(\mathbf{I}_{J_i} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} + \mathbf{A}_i \boldsymbol{\Sigma}_{\mathbf{b}} \mathbf{A}_i^T) - \{\mathrm{vec}(\mathbf{Y}_i) - \boldsymbol{\alpha} \otimes \mathbf{1}_{J_i} - \mathrm{vec}(\boldsymbol{\beta}\mathbf{X}_i)\}^T$$
$$(\mathbf{I}_{J_i} \otimes \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} + \mathbf{A}_i \boldsymbol{\Sigma}_{\mathbf{b}} \mathbf{A}_i^T)^{-1} \{\mathrm{vec}(\mathbf{Y}_i) - \boldsymbol{\alpha} \otimes \mathbf{1}_{J_i} - \mathrm{vec}(\boldsymbol{\beta}\mathbf{X}_i)\} \big].$$

Thus, $\widehat{\boldsymbol{\theta}}_{mix\cdot em}$ is the solution to the generalized estimating equation (GEE)

$$\frac{\partial l}{\partial \boldsymbol{\theta}^T} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \boldsymbol{\theta}^T} = \sum_{i=1}^{n} \psi_i^T(\mathbf{Y}_i, \boldsymbol{\theta}) = 0,$$

where l_i is the misspecified log-likelihood of each observation, and $\psi_i(\mathbf{Y}_i, \boldsymbol{\theta}) = \dfrac{\partial l_i}{\partial \boldsymbol{\theta}}$. Because $\mathbf{Y}_i$ has finite second moment, $\mathbb{E}\{\psi_i(\mathbf{Y}_i, \boldsymbol{\theta}_0)\} = 0$, where the subscript 0 indicates the true parameter value. We apply Proposition 5.5 and Theorem 5.14 in Shao (2003) to prove consistency and asymptotical normality of $\widehat{\boldsymbol{\theta}}_{mix\cdot em}$.

In order to use Proposition 5.5, we need to show the conditions in Lemma 5.3 in Shao (2003) holds for any compact subset of the parameter space. That is, for any $c > 0$ and sequence $\{\mathbf{y}_i\}_{i=1}^{\infty}$ satisfying $\|\mathbf{y}_i\| \le c$, the sequence of functions $\psi_i(\mathbf{y}_i, \boldsymbol{\theta})$ is equicontinuous on any compact set of the parameter space. It is easy to see that $\dfrac{\partial \psi_i}{\partial \boldsymbol{\theta}^T}$ (derived in the previous subsection) is uniformly bounded in any compact subset Θ of the parameter space when $\|\mathbf{y}_i\| \le c$ if $\sup_i \|\mathbf{X}_i\| < \infty$, $\sup_i \|\mathbf{Z}_i\| < \infty$, $\sup_i J_i < \infty$ and $\sup_i \|\boldsymbol{\Sigma}_i^{-1}\|_\infty < \infty$, where $\|\cdot\|_\infty$ indicates matrix infinity norm. Since $\sup_i \|\boldsymbol{\Sigma}_i^{-1}\|_\infty < \infty$ if and only if $\sup_i \|\mathbf{Z}_i\| < \infty$ and $\inf_i \det(\boldsymbol{\Sigma}_i) >$

0, the aforementioned conditions hold under the regularity conditions. Therefore, $\psi_i(\mathbf{y}_i, \boldsymbol{\theta})$ is equicontinuous on Θ. Moreover, since $\mathbf{Y}_i$ has finite $(4+\delta)$-th moment, $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\Theta} \|\psi_i(\mathbf{Y}_i, \boldsymbol{\theta})\|\}^2 < \infty$ and $\sup_i \mathbb{E}\|\mathbf{Y}_i\| < \infty$, the conditions in Lemma 5.3 in Shao (2003) holds.

According to Proposition 5.5 of Shao (2003), we also need to prove $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{\infty} \mathbb{E}\{\psi_i(\mathbf{Y}_i, \boldsymbol{\theta})\} = 0$ implies $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Let $\boldsymbol{\theta}_{10} = \boldsymbol{\beta}_0, \boldsymbol{\theta}_{20} = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}0}, \boldsymbol{\theta}_{30} = \boldsymbol{\Sigma}_{\mathbf{b}0}$ denote the true parameter value, and $\boldsymbol{\Sigma}_{i0} = \mathbf{I}_{J_i} \otimes \boldsymbol{\theta}_{20} + \mathbf{A}_i \boldsymbol{\theta}_{30} \mathbf{A}_i^T$. Taking expectation of $\psi_i(\mathbf{Y}_i, \boldsymbol{\theta})$, we have

$$\mathbb{E}\{\psi_{i1}(\mathbf{Y}_i, \boldsymbol{\theta})\} = \mathrm{vec}^T\{(\boldsymbol{\theta}_{10} - \boldsymbol{\beta})\mathbf{X}_i\}\boldsymbol{\Sigma}_i^{-1}(\mathbf{X}_i \otimes \mathbf{I}_r),$$

$$\mathbb{E}\{\psi_{i2}(\mathbf{Y}_i, \boldsymbol{\theta})\} = \frac{1}{2}\mathrm{vec}^T(\boldsymbol{\Sigma}_{i0}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_{i0}^{-1} - \boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_i r}\mathbf{M}_{i1},$$

$$\mathbb{E}\{\psi_{i3}(\mathbf{Y}_i, \boldsymbol{\theta})\} = \frac{1}{2}\mathrm{vec}^T(\boldsymbol{\Sigma}_{i0}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_{i0}^{-1} - \boldsymbol{\Sigma}_i^{-1})\mathbf{E}_{J_i r}\mathbf{M}_{i2}.$$

Also,

$$\mathrm{vec}^T(\boldsymbol{\Sigma}_{i0}^{-1}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_{i0}^{-1} - \boldsymbol{\Sigma}_i^{-1}) = 0 \quad \text{if and only if} \quad \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_{i0}.$$

Because $\mathbf{X}_i$ and $\mathbf{Z}_i$ can be arbitrary,

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_{i0} \quad \text{if and only if} \quad \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \boldsymbol{\theta}_{20} \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathbf{b}} = \boldsymbol{\theta}_{30}.$$

Therefore,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{\infty} \mathbb{E}\{\psi_{i1}(\mathbf{Y}_i, \boldsymbol{\theta})\} = 0 \quad \text{implies} \quad \boldsymbol{\beta} = \boldsymbol{\theta}_{10},$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{\infty} \mathbb{E}\{\psi_{i2}(\mathbf{Y}_i, \boldsymbol{\theta})\} = 0 \quad \text{implies} \quad \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \boldsymbol{\theta}_{20}, \boldsymbol{\Sigma}_{\mathbf{b}} = \boldsymbol{\theta}_{30}.$$

Hence, $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{\infty} \mathbb{E}\{\psi_i(\mathbf{Y}_i, \boldsymbol{\theta})\} = 0$ implies $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Since $\widehat{\boldsymbol{\theta}}_{mix\cdot em}$ is always $O_p(1)$, by Proposition 5.5 in (Shao, 2003), $\widehat{\boldsymbol{\theta}}_{mix\cdot em} \xrightarrow{p} \boldsymbol{\theta}_0$.

Then we prove asymptotic normality of $\widehat{\boldsymbol{\theta}}_{mix\cdot em}$ using Theorem 5.14 of (Shao, 2003). Since $\mathbf{Y}_i$ has finite $(4+\delta)$-th moment, $\sup_i \|\psi_i(\mathbf{Y}_i, \boldsymbol{\theta})\|^{2+\frac{\delta}{2}} < \infty$. Then, if conditions $\liminf_n \lambda_-[n^{-1}\mathrm{Var}(s_n(\boldsymbol{\theta}))] > 0$ and $\liminf_n \lambda_-[n^{-1}\mathbf{M}_n(\boldsymbol{\theta})] > 0$ holds, we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{mix\cdot em} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \widetilde{\mathbf{V}}),$$

where $\widetilde{\mathbf{V}} = \frac{1}{n}[\mathbf{M}_n(\boldsymbol{\theta}_0)]^{-1}\mathrm{Var}(s_n(\boldsymbol{\theta}_0))[\mathbf{M}_n(\boldsymbol{\theta}_0)]^{-1}$.

Shapiro's $\boldsymbol{\xi}$ in our context is $\boldsymbol{\xi} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}, \boldsymbol{\Sigma}_{\mathbf{b}})$. Following the same technique

in Su and Cook (2012) and Cook et al. (2015), we give the minimum discrepancy function as $f_{MDF} = l_{max} - l$, where l is the misspecified log-likelihood function (B.2), and l_{max} is obtained by substituting $\widehat{\boldsymbol{\theta}}_{mix\cdot em}$ for $\boldsymbol{\theta}$ in (B.2). Although f_{MDF} is written in terms of $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}_{mix\cdot em}$, there must be one-to-one functions f_1 from $\boldsymbol{\theta}$ to $\boldsymbol{\xi}$ and f_2 from $\widehat{\boldsymbol{\theta}}_{mix\cdot em}$ to $\mathbf{x}$ so that $\boldsymbol{\xi} = f_1(\boldsymbol{\theta})$ and $\mathbf{x} = f_2(\widehat{\boldsymbol{\theta}}_{mix\cdot em})$. As f_{MDF} is constructed under the normal likelihood, it satisfies the four conditions required by Shapiro (1986). Denote $\boldsymbol{\phi} = (\boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Sigma}_{\mathbf{b}})$, $\boldsymbol{\theta} = \mathbf{h}(\boldsymbol{\phi})$, $\mathbf{G} = \partial \mathbf{h}(\boldsymbol{\phi})/\partial \boldsymbol{\phi}^T$ and $\mathbf{J} = \frac{1}{2} \cdot \frac{\partial^2 f_{MDF}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$. Notice that $\mathbf{J}$ equals the Fisher information matrix for $\boldsymbol{\theta}$ when $\boldsymbol{\varepsilon}$ is normal.

Because $\widehat{\boldsymbol{\theta}}_{mix\cdot em}$ is obtained by minimizing f_{MDF}, by Proposition 4.1 of Shapiro (1986),

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{mix\cdot env} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \widetilde{\mathbf{V}}_{mix\cdot env}),$$

where $\widetilde{\mathbf{V}}_{mix\cdot env} = \mathbf{G}(\mathbf{G}^T\mathbf{J}\mathbf{G})^\dagger \mathbf{G}^T \mathbf{J} \widetilde{\mathbf{V}} \mathbf{J} \mathbf{G} (\mathbf{G}^T \mathbf{J} \mathbf{G})^\dagger \mathbf{G}^T$. If we define the inner product as $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle_{\mathbf{J}} = \mathbf{x}_1^T \mathbf{J} \mathbf{x}_2$, then the projection onto span($\mathbf{B}$) relative to $\mathbf{J}$ has the matrix representation $\mathbf{P}_{\mathbf{B}(\mathbf{J})} = \mathbf{B}(\mathbf{B}^T\mathbf{J}\mathbf{B})^\dagger \mathbf{B}^T \mathbf{J}$. Hence, $\widetilde{\mathbf{V}}_{mix\cdot env} = \mathbf{P}_{\mathbf{G}(\mathbf{J})} \widetilde{\mathbf{V}} \mathbf{P}_{\mathbf{G}(\mathbf{J})} \leq \widetilde{\mathbf{V}}$.

Graphical illustration of the fixed effects model

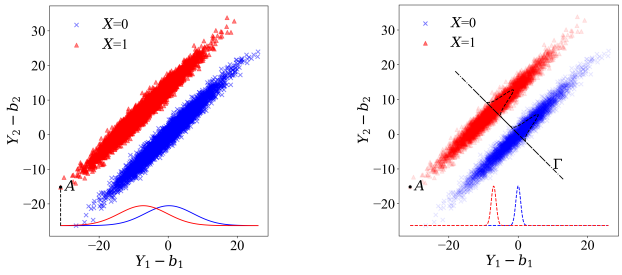
We present a graphical illustration of the fixed effects model. Using the same data as Section 3.3.2, except that we assume the random effects $\mathbf{b}_i$ are observed. In this case, we use $\mathbf{Y}_i - \mathbf{b}_i$ as response. That is, (1) holds for response $\mathbf{Y}_i$ and observations are independent for different i. After subtracting $\mathbf{b}_i$, (3) holds for response $\mathbf{Y}_i - \mathbf{b}_i$ and observations are independent for different i and j.

Figure B.1 demonstrates the intuition for the efficiency gain of the classic envelope method in standard multivariate regression with fixed effects only. In Figure B.1 (a), the ordinary least square estimator (OLS) is obtained by projecting all the data onto the $Y_1 - b_1$ axis, ignoring Y_2 completely. The density curves of the two group distributions of $Y_{ij1} - b_{i1}$ are given at the bottom in Figure B.1 and similar curves can be made for $Y_{ij2} - b_{i2}$. The two density curves are not well separated. The OLS of the group difference of $\mathbf{Y}_{ij} - \mathbf{b}_i$ is $(-7.67, 6.53)^T$ with standard error being $(0.60, 0.60)^T$ and mean square error (MSE) of the group difference being 0.69.

The idea of the envelope method is to reduce noise in the data by projecting each observation onto the direction that can best distinguish the groups. The two groups are well separated along the dashed black line. Also, they have almost

the same distribution in the direction that is orthogonal to the black solid line. Therefore, discarding that part of variation does not sacrifice the information of group difference, but instead, it makes the estimation more efficient. The density curves of the two groups under the envelope estimation are shown at the bottom of Figure B.1, and they have much smaller spreads than the OLS. The envelope estimate for the group difference of $\mathbf{Y}_{ij} - \mathbf{b}_i$ is $(-7.10, 7.10)^T$ with standard error $(0.04, 0.04)^T$ and MSE 0.0025.

Figure B.1: Graphical illustration of the OLS and the classic envelope estimator when the response is $\mathbf{Y} - \mathbf{b}$, i.e., with only fixed effect. The two groups are denoted by triangle and cross dots.



(a) OLS (b) Envelope

B.2 EM-algorithm for the mixed effects envelope model

Recall the standard EM-algorithm iterates through the following two steps:

(a) E-step. Suppose we have the parameter $\boldsymbol{\theta}_t$ from the $t-$th iteration, then we compute the function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t) = \int l(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{b}) \prod_{i=1}^{n} f(\mathbf{b}_i|\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i, \boldsymbol{\theta}_t) \mathrm{d}\mathbf{b}_i$, where l is the log-likelihood function, and f is the density function. (b) M-step. Find the maximizer of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$ as $\boldsymbol{\theta}_{t+1}$.

To calculate the function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$ in the E-step, we derive the formulas for both $f(\mathbf{b}_i|\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}_t)$ and $l(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{b})$. Note $f(\mathbf{b}_i|\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \boldsymbol{\theta}_t) = f(\mathbf{b}_i, \mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}_t)/f(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}_t)$, thus, $\mathbf{b}_i|\mathbf{X}_i, \mathbf{Y}_i, \mathbf{Z}_i, \boldsymbol{\theta}_t$ also follows a normal distribution $N(\boldsymbol{\mu}_{\mathbf{b}_i,t}, \boldsymbol{\Sigma}_{\mathbf{b}_i,t})$, where $\boldsymbol{\mu}_{\mathbf{b}_i,t} = \big\{\sum_{j=1}^{J_i} \mathbf{A}_{ij}^T \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t}^{-1} (\mathbf{Y}_{ij} - \boldsymbol{\alpha}_t - \boldsymbol{\beta}_t \mathbf{X}_{ij})\big\} (\boldsymbol{\Sigma}_{\mathbf{b},t}^{-1} + \sum_{j=1}^{J_i} \mathbf{A}_{ij}^T \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t} \mathbf{A}_{ij})^{-1}$, $\boldsymbol{\Sigma}_{\mathbf{b}_i,t} = (\boldsymbol{\Sigma}_{\mathbf{b},t}^{-1} + \sum_{j=1}^{J_i} \mathbf{A}_{ij}^T \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t} \mathbf{A}_{ij})^{-1}$. Derivation of $\boldsymbol{\mu}_{\mathbf{b}_i,t}$ and $\boldsymbol{\Sigma}_{\mathbf{b}_i,t}$ is given later in the Supplemen-

tary Material. Therefore,

$$
\begin{aligned}
&Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)\\
=&-\frac{rJ_{\cdot}}{2}\log(2\pi)-\frac{nqr}{2}\log(2\pi)\\
&-\frac{n}{2}\log|\boldsymbol{\Sigma}_{\mathbf{b}}|-\frac{J_{\cdot}}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}|-\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left\{\mathrm{vec}(\mathbf{b}_i)^T\boldsymbol{\Sigma}_{\mathbf{b}}^{-1}\mathrm{vec}(\mathbf{b}_i)|\mathbf{X},\mathbf{Y},\mathbf{Z};\boldsymbol{\mu}_{\mathbf{b}_i,t},\boldsymbol{\Sigma}_{\mathbf{b}_i,t}\right\}\\
&-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\mathbb{E}\left\{(\mathbf{Y}_{ij}-\boldsymbol{\alpha}-\boldsymbol{\beta}\mathbf{X}_{ij}-\mathbf{b}_i\mathbf{Z}_{ij})^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(\mathbf{Y}_{ij}-\boldsymbol{\alpha}-\boldsymbol{\beta}\mathbf{X}_{ij}-\mathbf{b}_i\mathbf{Z}_{ij})|\mathbf{X},\mathbf{Y},\mathbf{Z};\boldsymbol{\mu}_{\mathbf{b}_i,t},\boldsymbol{\Sigma}_{\mathbf{b}_i,t}\right\}\\
=&\sum_{i=1}^{n}\sum_{j=1}^{J_i}-\frac{1}{2}\left\{(\mathbf{Y}_{ij}-\boldsymbol{\alpha}-\boldsymbol{\beta}\mathbf{X}_{ij}-\mathbf{A}_{ij}\boldsymbol{\mu}_{\mathbf{b}_i,t})^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(\mathbf{Y}_{ij}-\boldsymbol{\alpha}-\boldsymbol{\beta}\mathbf{X}_{ij}-\mathbf{A}_{ij}\boldsymbol{\mu}_{\mathbf{b}_i,t})+\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}_{ij}\boldsymbol{\Sigma}_{\mathbf{b}_i,t}\mathbf{A}_{ij}^T)\right\}\\
&-\frac{1}{2}\sum_{i=1}^{n}\left\{\boldsymbol{\mu}_{\mathbf{b}_i,t}^T\boldsymbol{\Sigma}_{\mathbf{b}}^{-1}\boldsymbol{\mu}_{\mathbf{b}_i,t}+\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{b}}^{-1}\boldsymbol{\Sigma}_{\mathbf{b}_i,t})\right\}-\frac{J_{\cdot}}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}|-\frac{n}{2}\log|\boldsymbol{\Sigma}_{\mathbf{b}}|+C,
\end{aligned}
$$

where C is a constant.

Omitting the constant C, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)$ in the above equation can be decomposed as a summation of two parts, i.e., $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_t)=Q_1(\boldsymbol{\Sigma}_{\mathbf{b}}|\boldsymbol{\theta}_t)+Q_2(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}|\boldsymbol{\theta}_t)$, where

$$
Q_1(\boldsymbol{\Sigma}_{\mathbf{b}}|\boldsymbol{\theta}_t)=-\frac{n}{2}\log|\boldsymbol{\Sigma}_{\mathbf{b}}|+\sum_{i=1}^{n}-\frac{1}{2}\left\{\boldsymbol{\mu}_{\mathbf{b}_i,t}^T\boldsymbol{\Sigma}_{\mathbf{b}}^{-1}\boldsymbol{\mu}_{\mathbf{b}_i,t}+\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{b}}^{-1}\boldsymbol{\Sigma}_{\mathbf{b}_i,t})\right\},
$$

$$
\begin{aligned}
Q_2(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}|\boldsymbol{\theta}_t)=&\sum_{i=1}^{n}\sum_{j=1}^{J_i}-\frac{1}{2}\Big\{(\mathbf{Y}_{ij}-\boldsymbol{\alpha}-\boldsymbol{\beta}\mathbf{X}_{ij}-\mathbf{A}_{ij}\boldsymbol{\mu}_{\mathbf{b}_i,t})^T\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(\mathbf{Y}_{ij}-\boldsymbol{\alpha}-\boldsymbol{\beta}\mathbf{X}_{ij}-\mathbf{A}_{ij}\boldsymbol{\mu}_{\mathbf{b}_i,t})\\
&+\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}_{ij}\boldsymbol{\Sigma}_{\mathbf{b}_i,t}\mathbf{A}_{ij}^T)\Big\}-\frac{J_{\cdot}}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}|.
\end{aligned}
$$

Updates of parameters can be done separately for the two parts. Let $\boldsymbol{\mu}_{\mathbf{b},t}=(\boldsymbol{\mu}_{\mathbf{b}_1,t},\ldots,\boldsymbol{\mu}_{\mathbf{b}_n,t})$, then Q_1 can be written as

$$
Q_1(\boldsymbol{\Sigma}_{\mathbf{b}}|\boldsymbol{\theta}_t)=-\frac{n}{2}\log|\boldsymbol{\Sigma}_{\mathbf{b}}|-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathbf{b}}^{-1}(\boldsymbol{\mu}_{\mathbf{b},t}\boldsymbol{\mu}_{\mathbf{b},t}^T+\sum_{i=1}^{n}\boldsymbol{\Sigma}_{\mathbf{b}_i,t})\right\}.
$$

The update of $\boldsymbol{\Sigma}_{\mathbf{b}}$ is $\boldsymbol{\Sigma}_{\mathbf{b},t+1}=\sum_{i=1}^{n}\boldsymbol{\Sigma}_{\mathbf{b}_i,t}/n+\boldsymbol{\mu}_{\mathbf{b},t}\boldsymbol{\mu}_{\mathbf{b},t}^T/n$.

Now, we update $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ at the $t+1$ step. Under the envelope assumptions, we have $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}=\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2$, where $\boldsymbol{\Sigma}_1=\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T$, $\boldsymbol{\Sigma}_2=\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$ with $\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2=\mathbf{0}$, and $\boldsymbol{\beta}$ belongs to the subspace spanned by the column vectors in $\boldsymbol{\Sigma}_1$. So we have $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_2=\mathbf{0}$.

Moreover, we have $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} = \boldsymbol{\Sigma}_1^{\dagger} + \boldsymbol{\Sigma}_2^{\dagger}$, where the superscript '†' denotes generalized inverse.

When $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ are fixed, the parameter $\boldsymbol{\alpha}$ maximizing $Q_2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}|\boldsymbol{\theta}_t)$ is $\widehat{\boldsymbol{\alpha}} = \bar{\mathbf{Y}} - \boldsymbol{\beta}\bar{\mathbf{X}} - \bar{\boldsymbol{\mu}}_t$, where $\bar{\mathbf{Y}} = \sum_{ij} \mathbf{Y}_{ij}/J_{\cdot}$, $\bar{\mathbf{X}} = \sum_{ij} \mathbf{X}_{ij}/J_{\cdot}$, and $\bar{\boldsymbol{\mu}}_t = \sum_{ij} \mathbf{A}_{ij}\boldsymbol{\mu}_{\mathbf{b}_i,t}/J_{\cdot}$. According to this relationship, $\boldsymbol{\alpha}_{t+1}$ is a function of $\boldsymbol{\beta}_{t+1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t+1}$. Substitute this into the formula of Q_2, we obtain

$$Q_2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}|\boldsymbol{\theta}_t) = -\frac{J_{\cdot}}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}| - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\Big[\left\{\mathbf{Y}_{ij} - \boldsymbol{\beta}\mathbf{X}_{ij} - \mathbf{A}_{ij}\boldsymbol{\mu}_{\mathbf{b}_i,t} - (\bar{\mathbf{Y}} - \boldsymbol{\beta}\bar{\mathbf{X}} - \bar{\boldsymbol{\mu}}_t)\right\}^T \cdot \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\left\{\mathbf{Y}_{ij} - \boldsymbol{\beta}\mathbf{X}_{ij} - \mathbf{A}_{ij}\boldsymbol{\mu}_{\mathbf{b}_i,t} - (\bar{\mathbf{Y}} - \boldsymbol{\beta}\bar{\mathbf{X}} - \bar{\boldsymbol{\mu}}_t)\right\} + \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}_{ij}\boldsymbol{\Sigma}_{\mathbf{b}_i,t}\mathbf{A}_{ij}^T)\Big].$$

Let $\boldsymbol{\Psi}_{\mathbf{b},t} = \sum_{i=1}^{n}\sum_{j=1}^{J_i}\mathbf{A}_{ij}\boldsymbol{\Sigma}_{\mathbf{b}_i,t}\mathbf{A}_{ij}^T$. Also, let $\mathbf{Y}_c = \mathbf{Y} - \bar{\mathbf{Y}} \otimes \mathbf{1}_{1\times J_{\cdot}}$, $\mathbf{X}_c = \mathbf{X} - \bar{\mathbf{X}} \otimes \mathbf{1}_{1\times J_{\cdot}}$, $\boldsymbol{\mu}_{c,t} = (\mathbf{A}_{11}\boldsymbol{\mu}_{\mathbf{b}_1,t}, \ldots, \mathbf{A}_{nJ_n}\boldsymbol{\mu}_{\mathbf{b}_n,t}) - \bar{\boldsymbol{\mu}}_t \otimes \mathbf{1}_{1\times J_{\cdot}}$. Then we have

$$\begin{aligned}
&Q_2(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}|\boldsymbol{\theta}_t)\\
&= -\frac{1}{2}J_{\cdot}\log|\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2| - \frac{1}{2}\mathrm{tr}\left\{(\mathbf{Y}_c - \boldsymbol{\beta}\mathbf{X}_c - \boldsymbol{\mu}_{c,t})^T(\boldsymbol{\Sigma}_1^{\dagger} + \boldsymbol{\Sigma}_2^{\dagger})(\mathbf{Y}_c - \boldsymbol{\beta}\mathbf{X}_c - \boldsymbol{\mu}_{c,t}) + (\boldsymbol{\Sigma}_1^{\dagger} + \boldsymbol{\Sigma}_2^{\dagger})\boldsymbol{\Psi}_{\mathbf{b},t}\right\}\\
&= -\frac{1}{2}J_{\cdot}\log\det{}_0\boldsymbol{\Sigma}_1 - \frac{1}{2}\mathrm{tr}\left\{(\mathbf{Y}_c - \boldsymbol{\beta}\mathbf{X}_c - \boldsymbol{\mu}_{c,t})^T\boldsymbol{\Sigma}_1^{\dagger}(\mathbf{Y}_c - \boldsymbol{\beta}\mathbf{X}_c - \boldsymbol{\mu}_{c,t})\right\} - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_1^{\dagger}\boldsymbol{\Psi}_{\mathbf{b},t})\\
&\quad - \frac{1}{2}J_{\cdot}\log\det{}_0\boldsymbol{\Sigma}_2 - \frac{1}{2}\mathrm{tr}\left\{(\mathbf{Y}_c - \boldsymbol{\mu}_{c,t})^T\boldsymbol{\Sigma}_2^{\dagger}(\mathbf{Y}_c - \boldsymbol{\mu}_{c,t})\right\} - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_2^{\dagger}\boldsymbol{\Psi}_{\mathbf{b},t}).
\end{aligned}$$

When $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ are fixed, $\widehat{\boldsymbol{\beta}}$ that maximizes Q_2 is

$$\widehat{\boldsymbol{\beta}}_{t+1} = P_{\boldsymbol{\Sigma}_1}\mathbf{U}_{c,t}\mathbf{X}_c^T(\mathbf{X}_c\mathbf{X}_c^T)^{-1}, \tag{B.3}$$

where $P_{\boldsymbol{\Sigma}_1}$ denotes the projection matrix on the space spanned by the column vectors of $\boldsymbol{\Sigma}_1$, i.e., $P_{\boldsymbol{\Sigma}_1} = \boldsymbol{\Sigma}_1({\boldsymbol{\Sigma}_1}^T\boldsymbol{\Sigma}_1)^{-1}{\boldsymbol{\Sigma}_1}^T$, and $\mathbf{U}_{c,t} = \mathbf{Y}_c - \boldsymbol{\mu}_{c,t}$. Also, denote $\mathbf{Q}_{\boldsymbol{\Sigma}_1} = \mathbf{I}_r - P_{\boldsymbol{\Sigma}_1}$,

Then, we split Q_2 into the following two parts:

$$\begin{aligned}
Q_{2,1}(\boldsymbol{\Sigma}_1|\boldsymbol{\theta}_t) &= -\frac{J_{\cdot}}{2}\log\det{}_0\boldsymbol{\Sigma}_1 - \frac{1}{2}\mathrm{tr}(\mathbf{Q}_{\mathbf{X}_c}\mathbf{U}_{c,t}^T\boldsymbol{\Sigma}_1^{\dagger}\mathbf{U}_{c,t}\mathbf{Q}_{\mathbf{X}_c}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_1^{\dagger}\boldsymbol{\Psi}_{\mathbf{b},t}),\\
Q_{2,2}(\boldsymbol{\Sigma}_2|\boldsymbol{\theta}_t) &= -\frac{J_{\cdot}}{2}\log\det{}_0\boldsymbol{\Sigma}_2 - \frac{1}{2}\mathrm{tr}(\mathbf{U}_{c,t}^T\boldsymbol{\Sigma}_2^{\dagger}\mathbf{U}_{c,t}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_2^{\dagger}\boldsymbol{\Psi}_{\mathbf{b},t}),
\end{aligned}$$

where $\mathbf{Q}_{\mathbf{X}_c} = \mathbf{I} - P_{\mathbf{X}_c}$, and $\det_0(\mathbf{A})$ is defined as the product of its non-zero eigenvalues. Suppose $\boldsymbol{\Gamma}$ is given, then the maximizers of $Q_{2,1}$ and $Q_{2,2}$ respectively are

$\widehat{\mathbf{\Sigma}}_{1,t+1} = P_{\mathbf{\Gamma}}(\mathbf{U}_{c,t}\mathbf{Q}_{\mathbf{X}_c^T}\mathbf{U}_{c,t}^T + \mathbf{\Psi}_{\mathbf{b},t})P_{\mathbf{\Gamma}}/J_\bullet$, $\widehat{\mathbf{\Sigma}}_{2,t+1} = \mathbf{Q}_{\mathbf{\Gamma}}(\mathbf{U}_{c,t}\mathbf{U}_{c,t}^T + \mathbf{\Psi}_{\mathbf{b},t})\mathbf{Q}_{\mathbf{\Gamma}}/J_\bullet$. The maximized functions are $Q_{2,1} = C_1 - J_\bullet \log\det_0\left\{P_{\mathbf{\Gamma}}(\mathbf{U}_{c,t}\mathbf{Q}_{\mathbf{X}_c^T}\mathbf{U}_{c,t}^T + \mathbf{\Psi}_{\mathbf{b},t})P_{\mathbf{\Gamma}}\right\}/2$, $Q_{2,2} = C_2 - J_\bullet \log\det_0\left\{\mathbf{Q}_{\mathbf{\Gamma}}(\mathbf{U}_{c,t}\mathbf{U}_{c,t}^T + \mathbf{\Psi}_{\mathbf{b},t})\mathbf{Q}_{\mathbf{\Gamma}}\right\}/2$. Hence, $\widehat{\mathbf{\Sigma}}_{\varepsilon,t+1} = \widehat{\mathbf{\Sigma}}_{1,t+1} + \widehat{\mathbf{\Sigma}}_{2,t+1}$.

The final step is to find the semi-orthogonal matrix $\mathbf{\Gamma}$ to maximize the function Q_2, which is equivalent to minimizing the function

$$F(\text{span}(\mathbf{\Gamma})) = \log\det\left\{P_{\mathbf{\Gamma}}(\mathbf{U}_{c,t}\mathbf{Q}_{\mathbf{X}_c^T}\mathbf{U}_{c,t}^T + \mathbf{\Psi}_{\mathbf{b},t})P_{\mathbf{\Gamma}} + \mathbf{Q}_{\mathbf{\Gamma}}(\mathbf{U}_{c,t}\mathbf{U}_{c,t}^T + \mathbf{\Psi}_{\mathbf{b},t})\mathbf{Q}_{\mathbf{\Gamma}}\right\}.$$

We only need to identify the span of the column space of $\mathbf{\Gamma}$ from minimizing the above objective function. We use the 1D algorithm (Cook and Zhang, 2016) to obtain $\widehat{\mathbf{\Gamma}}$, where $\text{span}(\widehat{\mathbf{\Gamma}})$ is a $\sqrt{n}$-consistent estimator of $\text{span}(\mathbf{\Gamma})$, rather than MLE (more details on 1D algorithm is given later in the Supplementary Material). In our simulation studies in Section 3.4, our 1D algorithm is feasible and fast converging.

B.3 Technical details for the EM updates

We derive $\boldsymbol{\mu}_{\mathbf{b}_i,t}$ and $\mathbf{\Sigma}_{\mathbf{b}_i,t}$ in this section. They can be determined from $f(\mathbf{b}_i, \mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}_t)$.

$$\begin{aligned}
&f(\mathbf{b}_i, \mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\theta}_t)\\
=&(2\pi)^{-\frac{qr}{2}}|\mathbf{\Sigma}_{\mathbf{b},t}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\text{vec}(\mathbf{b}_i)^T\mathbf{\Sigma}_{\mathbf{b},t}^{-1}\text{vec}(\mathbf{b}_i)\right\}\\
&\left[\prod_{j=1}^{J_i}(2\pi)^{-\frac{r}{2}}|\mathbf{\Sigma}_{\boldsymbol{\varepsilon},t}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{Y}_{ij} - \boldsymbol{\alpha}_t - \boldsymbol{\beta}_t\mathbf{X}_{ij} - \mathbf{b}_i\mathbf{Z}_{ij})^T\mathbf{\Sigma}_{\boldsymbol{\varepsilon},t}^{-1}(\mathbf{Y}_{ij} - \boldsymbol{\alpha}_t - \boldsymbol{\beta}_t\mathbf{X}_{ij} - \mathbf{b}_i\mathbf{Z}_{ij})\right\}\right]\\
=&(2\pi)^{-\frac{qr}{2}}|\mathbf{\Sigma}_{\mathbf{b},t}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\text{vec}(\mathbf{b}_i)^T\mathbf{\Sigma}_{\mathbf{b},t}^{-1}\text{vec}(\mathbf{b}_i)\right\}\\
&\left[\prod_{j=1}^{J_i}(2\pi)^{-\frac{r}{2}}|\mathbf{\Sigma}_{\boldsymbol{\varepsilon},t}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{Y}_{ij} - \boldsymbol{\alpha}_t - \boldsymbol{\beta}_t\mathbf{X}_{ij} - \mathbf{A}_{ij}\text{vec}(\mathbf{b}_i))^T\mathbf{\Sigma}_{\boldsymbol{\varepsilon},t}^{-1}(\mathbf{Y}_{ij} - \boldsymbol{\alpha}_t - \boldsymbol{\beta}_t\mathbf{X}_{ij} - \mathbf{A}_{ij}\text{vec}(\mathbf{b}_i)\right.\right.\\
\propto&|\mathbf{\Sigma}_{\mathbf{b},t}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\text{vec}(\mathbf{b}_i) - \boldsymbol{\mu}_{\mathbf{b}_i,t})^T\mathbf{\Sigma}_{\mathbf{b}_i,t}^{-1}(\text{vec}(\mathbf{b}_i) - \boldsymbol{\mu}_{\mathbf{b}_i,t})\right\}.
\end{aligned}$$

Hence,

$$\boldsymbol{\mu}_{\mathbf{b}_i,t} = \left(\boldsymbol{\Sigma}_{\mathbf{b},t}^{-1} + \sum_{j=1}^{J_i} \mathbf{A}_{ij}^T \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t} \mathbf{A}_{ij} \right)^{-1} \sum\nolimits_{j=1}^{J_i} \left\{ \mathbf{A}_{ij}^T \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t}^{-1} (\mathbf{Y}_{ij} - \boldsymbol{\alpha}_t - \boldsymbol{\beta}_t \mathbf{X}_{ij}) \right\},$$

$$\boldsymbol{\Sigma}_{\mathbf{b}_i,t} = \left(\boldsymbol{\Sigma}_{\mathbf{b},t}^{-1} + \sum_{j=1}^{J_i} \mathbf{A}_{ij}^T \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t} \mathbf{A}_{ij} \right)^{-1}.$$

B.4 The mixed effects envelope algorithm

We combine the 1D algorithm with EM algorithm to obtain an estimator of the mixed effects model under conditions (i)* and (ii)* as follows, where δ can be chosen depending on the accuracy to achieve.

Algorithm 5: The mixed effects envelope algorithm

for *k = 1, 2, ..., u* **do**

Initialization: $t = 0$, $\boldsymbol{\Sigma}_{\mathbf{b},0} = \mathbf{I}_{qr}$, $\boldsymbol{\Sigma}_{\varepsilon,0} = \mathbf{I}_r$, $\boldsymbol{\alpha}_0 = \mathbf{0}$, $\boldsymbol{\beta}_0 = \mathbf{0}$, $\boldsymbol{\theta}_0 = (\boldsymbol{\Sigma}_{\mathbf{b},0}, \boldsymbol{\Sigma}_{\varepsilon,0}, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, $\Delta_0 = \infty$.

while $\Delta_t > \delta$ **do**

1. Set $\boldsymbol{\Sigma}_{\mathbf{b}_i,t} = \left(\boldsymbol{\Sigma}_{\mathbf{b},t}^{-1} + \sum_{j=1}^{J_i} \mathbf{A}_{ij}^T \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t}^{-1} \mathbf{A}_{ij}\right)^{-1}$, where $\mathbf{A}_{ij} = \mathbf{I}_r \otimes \mathbf{Z}_{ij}^T$ and $\boldsymbol{\mu}_{\mathbf{b}_i,t} =$ $\left(\boldsymbol{\Sigma}_{\mathbf{b},t}^{-1} + \sum_{j=1}^{J_i} \mathbf{A}_{ij}^T \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t}^{-1} \mathbf{A}_{ij}\right)^{-1} \sum_{j=1}^{J_i} \left\{\mathbf{A}_{ij}^T \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t}^{-1} (\mathbf{Y}_{ij} - \boldsymbol{\alpha}_t - \boldsymbol{\beta}_t \mathbf{X}_{ij})\right\}$.
The update of $\boldsymbol{\Sigma}_{\mathbf{b}}$ is $\boldsymbol{\Sigma}_{\mathbf{b},t+1} = \left(\sum_{i=1}^n \boldsymbol{\Sigma}_{\mathbf{b}_i,t} + \boldsymbol{\mu}_{\mathbf{b},t} \boldsymbol{\mu}_{\mathbf{b},t}^T\right)/N$, where $\boldsymbol{\mu}_{\mathbf{b},t} = (\boldsymbol{\mu}_{\mathbf{b}_1,t}, \ldots, \boldsymbol{\mu}_{\mathbf{b}_n,t})$.
2. The update of $\boldsymbol{\alpha}$ is $\boldsymbol{\alpha}_{t+1} = \bar{\mathbf{Y}} - \boldsymbol{\beta}_t \bar{\mathbf{X}} - \bar{\boldsymbol{\mu}}_t$, where $\bar{\mathbf{Y}} = \sum_{ij} \mathbf{Y}_{ij}/J_{\cdot}$, $\bar{\mathbf{X}} = \sum_{ij} \mathbf{X}_{ij}/J_{\cdot}$, and $\bar{\boldsymbol{\mu}}_t = \sum_{ij} \mathbf{A}_{ij} \boldsymbol{\mu}_{\mathbf{b}_i,t}/J_{\cdot}$.
3. Using the 1D Algorithm to get $\boldsymbol{\Gamma}_{t+1}$. Then, $\boldsymbol{\beta}_{t+1} = \mathbf{P}_{\boldsymbol{\Gamma}_{t+1}} \mathbf{U}_{c,t} \mathbf{X}_c^T (\mathbf{X}_c \mathbf{X}_c^T)^{-1}$, where $\mathbf{Y}_c = \mathbf{Y} - \bar{\mathbf{Y}} \otimes \mathbf{1}_{1\times NJ}$, $\mathbf{X}_c = \mathbf{X} - \bar{\mathbf{X}} \otimes \mathbf{1}_{1\times NJ}$, $\boldsymbol{\mu}_{c,t} = (\mathbf{A}_{11}\boldsymbol{\mu}_{\mathbf{b}_1,t}, \ldots, \mathbf{A}_{nJ_n}\boldsymbol{\mu}_{\mathbf{b}_n,t}) - \bar{\boldsymbol{\mu}}_t \otimes \mathbf{1}_{1\times NJ}$ and $\mathbf{U}_{c,t} = \mathbf{Y}_c - \boldsymbol{\mu}_{c,t}$.
4. $\boldsymbol{\Sigma}_{\varepsilon,t+1} = \boldsymbol{\Sigma}_{1,t+1} + \boldsymbol{\Sigma}_{2,t+1}$, where $\boldsymbol{\Sigma}_{1,t+1} = P_{\boldsymbol{\Gamma}_{t+1}} (\mathbf{U}_{c,t} \mathbf{Q}_{\mathbf{X}_c^T} \mathbf{U}_{c,t}^T + \boldsymbol{\Psi}_{\mathbf{b},t}) P_{\boldsymbol{\Gamma}_{t+1}}/J_{\cdot}$, $\boldsymbol{\Sigma}_{2,t+1} = \mathbf{Q}_{\boldsymbol{\Gamma}_{t+1}} (\mathbf{U}_{c,t} \mathbf{U}_{c,t}^T + \boldsymbol{\Psi}_{\mathbf{b},t}) \mathbf{Q}_{\boldsymbol{\Gamma}_{t+1}}/J_{\cdot}$, $\boldsymbol{\Psi}_{\mathbf{b},t} = \sum_{i=1}^n \sum_{j=1}^{J_i} \mathbf{A}_{ij} \boldsymbol{\Sigma}_{\mathbf{b}_i,t} \mathbf{A}_{ij}^T$.
5. Set $\Delta_{t+1} = \|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\|_1 / \|\boldsymbol{\beta}_{t+1}\|_1$, $\boldsymbol{\theta}_{t+1} = (\boldsymbol{\Sigma}_{t+1}, \boldsymbol{\beta}_{t+1}, \boldsymbol{\rho}_{t+1})$, $t \leftarrow t + 1$;

end

$\mathrm{BIC}_k = -2l(\widehat{\boldsymbol{\theta}}_k; \mathbf{Y}|\mathbf{X}) + pu \log n$, where $\widehat{\boldsymbol{\theta}}_k$ and $\widehat{\boldsymbol{\beta}}_k$ are the estimators when the iteration stops.

end

Select k which minimize BIC_k, and $\widehat{\boldsymbol{\beta}}_k$ is the mixed effects envelope estimator.

B.5 Additional tables

Table B.1: The point estimates, bootstrap standard errors and p-values for the regression parameter with respect to those patients attended all the measurements

	Our Method			Standard EM		
Corresponding to Treatment	$\widehat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	p-value	$\widehat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	p-value
Treatment Satisfaction	0.46	0.44	0.30	0.69	0.59	0.24
Depression Scale	-0.057	0.20	0.77	0.088	0.24	0.71
Physical Score	-2.13×10^{-3}	6.25×10^{-3}	0.73	-1.81×10^{-3}	0.025	0.94
Mental Score	0.011	0.033	0.73	0.011	0.061	0.86
Interference Score	4.57×10^{-4}	1.87×10^{-3}	0.81	2.83×10^{-3}	4.48×10^{-3}	0.52
Symptom & Distress Score	-0.99	1.89	0.60	-0.77	2.30	0.74
SBP	-0.32	0.39	0.41	0.074	0.69	0.91
DBP	-0.27	0.39	0.49	-0.22	0.46	0.64
Heart Rate	0.40	0.53	0.45	0.34	0.53	0.53
Corresponding to Age	$\widehat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	p-value	$\widehat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	p-value
Treatment Satisfaction	0.24	0.10	0.02	0.23	0.046	< 0.01
Depression Scale	-0.039	0.019	0.04	-0.071	0.013	< 0.01
Physical Score	-2.66×10^{-3}	9.95×10^{-4}	< 0.01	-6.78×10^{-3}	1.87×10^{-3}	< 0.01
Mental Score	5.40×10^{-3}	3.65×10^{-3}	0.14	0.012	4.14×10^{-3}	< 0.01
Interference Score	6.93×10^{-5}	2.12×10^{-4}	0.74	1.50×10^{-4}	2.71×10^{-4}	0.58
Symptom & Distress Score	-0.22	0.24	0.36	-0.29	0.15	0.06
SBP	0.071	0.01	0.45	0.041	0.051	0.41
DBP	-0.60	0.059	< 0.01	-0.61	0.035	< 0.01
Heart Rate	-0.34	0.037	< 0.01	-0.34	0.036	< 0.01

Table B.2: The point estimates, bootstrap standard errors and p-values for the regression parameter with respect to all the patients attended all four measurements

	Our Method			Standard EM		
Corresponding to Treatment	$\widehat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	p-value	$\widehat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	p-value
Treatment Satisfaction	-0.019	0.20	0.92	0.67	0.53	0.21
Depression Scale	7.44×10^{-3}	0.080	0.93	0.13	0.18	0.48
Physical Score	6.16×10^{-4}	1.41×10^{-3}	0.66	3.59×10^{-5}	0.019	0.99
Mental Score	-1.36×10^{-3}	0.017	0.94	-0.031	0.042	0.46
Interference Score	-3.77×10^{-5}	9.39×10^{-4}	0.97	-6.49×10^{-4}	2.72×10^{-3}	0.81
Symptom & Distress Score	0.14	1.50	0.93	1.23	1.69	0.47
SBP	8.19×10^{-3}	0.088	0.93	-0.45	0.63	0.47
DBP	4.56×10^{-3}	0.050	0.93	-0.60	0.36	0.10
Heart Rate	3.06×10^{-3}	0.034	0.93	0.13	0.40	0.74
Corresponding to Age	$\widehat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	p-value	$\widehat{\boldsymbol{\beta}}$	$\widehat{\text{SE}}$	p-value
Treatment Satisfaction	-1.67	0.34	< 0.01	0.23	0.51	< 0.01
Depression Scale	-0.65	0.12	< 0.01	-0.071	0.16	< 0.01
Physical Score	-0.061	0.011	0.40	-6.78×10^{-3}	0.018	< 0.01
Mental Score	-0.17	0.023	< 0.01	0.012	0.050	< 0.01
Interference Score	-0.015	1.21×10^{-3}	< 0.01	1.50×10^{-4}	3.27×10^{-3}	< 0.01
Symptom & Distress Score	1.22	2.16	< 0.01	-0.29	1.88	< 0.01
SBP	0.72	0.18	< 0.01	0.041	0.50	0.59
DBP	0.40	0.097	< 0.01	-0.61	0.39	0.85
Heart Rate	0.27	0.082	< 0.01	-0.34	0.45	< 0.01

Appendix C

Appendix for Chapter 4

C.1 Proof of Lemmas and Theorems

From equation (4.4), the nuisance tangent space can be written as

$$\begin{aligned}\Lambda &= \{s_1(\mathbf{\Gamma}^T\mathbf{Y} \mid \mathbf{X}) + s_2(\mathbf{B}^T\mathbf{\Gamma}_0^T\mathbf{Y} \mid \mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) + s_3(\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}) + s_4(\mathbf{X})\} \\ &= \Lambda_1 \cup \Lambda_2 \cup \Lambda_3 \cup \Lambda_4,\end{aligned}$$

where s_1, s_2, s_3 and s_4 are measurable functions satisfying

$$\int s_1(\mathbf{\Gamma}^T\mathbf{y} \mid \mathbf{x})\eta_1(\mathbf{\Gamma}^T\mathbf{y} \mid \mathbf{x})d\mathbf{y} = \mathbf{0}, \quad \text{for any } \mathbf{x},$$

$$\int s_2(\mathbf{B}^T\mathbf{\Gamma}_0^T\mathbf{y} \mid \mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{y}, \mathbf{x})\eta_2(\mathbf{B}^T\mathbf{\Gamma}_0^T\mathbf{y} \mid \mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{y}, \mathbf{x})d\mathbf{y} = \mathbf{0}, \quad \text{for any } (\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{y}, \mathbf{x}),$$

$$\int s_3(\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{y})d\mathbf{y} = \mathbf{0}, \qquad \int s_4(\mathbf{x})d\mathbf{x} = \mathbf{0},$$

and Λ_i is the Hilbert space spanned by s_i for $i = 1, \ldots, 4$. Therefore, the structure of Λ_i has the following representation

$$\begin{aligned}\Lambda_1 &= \{\mathbf{f}(\mathbf{\Gamma}^T\mathbf{Y}, \mathbf{X}) : \forall \mathbf{f} \text{ such that } \mathbb{E}(\mathbf{f} \mid \mathbf{X}) = \mathbf{0}\}, \\ \Lambda_2 &= \{\mathbf{f}(\mathbf{B}^T\mathbf{\Gamma}_0^T\mathbf{Y}, \mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) : \forall \mathbf{f} \text{ such that } \mathbb{E}(\mathbf{f} \mid \mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) = \mathbf{0}\}, \\ \Lambda_3 &= \{\mathbf{f}(\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}) : \forall \mathbf{f} \text{ such that } \mathbb{E}(\mathbf{f}) = \mathbf{0}\}, \\ \Lambda_4 &= \{\mathbf{f}(\mathbf{X}) : \forall \mathbf{f} \text{ such that } \mathbb{E}(\mathbf{f}) = \mathbf{0}\},\end{aligned}$$

where $\mathbf{f}$ is a square integrable function.

The following lemmas provide an orthogonal decomposition of the nuisance tan-

gent space and the form of the orthogonal nuisance tangent space.

Lemma C.1.1 *Under Condition (4.3.1) and (4.3.2), the nuisance tangent space can be written as*

$$\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3 \oplus \Lambda_4.$$

Proof of Lemma C.1.1. We prove that $\Lambda_1 \perp \Lambda_2 \perp \Lambda_3 \perp \Lambda_4$. Obviously, $\Lambda_4 \perp (\Lambda_1 \cup \Lambda_2 \cup \Lambda_3)$. We only prove that Λ_1, Λ_2 and Λ_3 are orthogonal to each other. For any $f_1 \in \Lambda_1$ and $f_2 \in \Lambda_2$,

$$\begin{aligned}
&\mathbb{E}\{f_1^T(\boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) f_2(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X})\} \\
=&\mathbb{E}[\mathbb{E}\{f_1^T(\boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) f_2(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) \mid \mathbf{X}\}] \\
=&\mathbb{E}[\mathbb{E}\{f_1^T(\boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) \mid \mathbf{X}\}\mathbb{E}\{f_2(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) \mid \mathbf{X}\}] = \mathbf{0}.
\end{aligned}$$

Hence, $\Lambda_1 \perp \Lambda_2$. The second equation to the third equation is because $\boldsymbol{\Gamma}^T\mathbf{Y} \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{X}$. The third equation equals to $\mathbf{0}$ because $\mathbb{E}(f_1 \mid \mathbf{X}) = \mathbf{0}$. For any $f_1 \in \Lambda_1$ and $f_3 \in \Lambda_3$,

$$\begin{aligned}
&\mathbb{E}\{f_1^T(\boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) f_3(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})\} \\
=&\mathbb{E}[\mathbb{E}\{f_1^T(\boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) f_3(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) \mid \mathbf{X}\}] \\
=&\mathbb{E}[\mathbb{E}\{f_1^T(\boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) \mid \mathbf{X}\}\mathbb{E}\{f_3(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) \mid \mathbf{X}\} = \mathbf{0}.
\end{aligned}$$

Thus, $\Lambda_1 \perp \Lambda_3$. The second to the third equation is again because of Condition 4.3.1. For any $f_2 \in \Lambda_2$ and$f_3 \in \Lambda_3$,

$$\begin{aligned}
&\mathbb{E}\{f_2^T(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) f_3(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})\} \\
=&\mathbb{E}[\mathbb{E}\{f_2^T(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) f_3(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\}] \\
=&\mathbb{E}[\mathbb{E}\{f_2^T(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\} f_3(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})] = \mathbf{0}
\end{aligned}$$

The last equation equals to 0 because $\mathbb{E}(f_2 \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) = 0$. Hence, Λ_i, $i = 1, \ldots, 4$ are orthogonal to each other. Therefore,

$$\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3 \oplus \Lambda_4.$$

■

Lemma C.1.2 *The orthogonal nuisance tangent space* $\Lambda^\perp = \Lambda_1^\perp \cap \Lambda_2^\perp \cap \Lambda_3^\perp \cap \Lambda_4^\perp$

where

$$\Lambda_1^{\perp} = \{\mathbf{f}(\mathbf{Y},\mathbf{X}) : \mathbb{E}(\mathbf{f} \mid \mathbf{\Gamma}^T\mathbf{Y},\mathbf{X}) \textit{ is a function of } \mathbf{X}\},$$
$$\Lambda_2^{\perp} = \{\mathbf{f}(\mathbf{Y},\mathbf{X}) : \mathbb{E}(\mathbf{f} \mid \mathbf{B}^T\mathbf{\Gamma}_0^T\mathbf{Y}, \mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y},\mathbf{X}) \textit{ is a function of } \mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y},\mathbf{X}\},$$
$$\Lambda_3^{\perp} = \{\mathbf{f}(\mathbf{Y},\mathbf{X}) : \mathbb{E}(\mathbf{f} \mid \mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}) = \mathbf{0}\},$$
$$\Lambda_4^{\perp} = \{\mathbf{f}(\mathbf{Y},\mathbf{X}) : \mathbb{E}(\mathbf{f} \mid \mathbf{X}) = \mathbf{0}\}.$$

Proof of Lemma C.1.2. From the structure of Λ, we have Our conjecture for the orthogonal complements for Λ_i, $i = 1, \ldots, 4$ are

$$\Lambda_{1,\text{conj}}^{\perp} = \{f(\mathbf{Y},\mathbf{X}) : \mathbb{E}(f \mid \mathbf{\Gamma}^T\mathbf{Y},\mathbf{X}) \text{ is a function of } \mathbf{X}\},$$
$$\Lambda_{2,\text{conj}}^{\perp} = \{f(\mathbf{Y},\mathbf{X}) : \mathbb{E}(f \mid \mathbf{B}^T\mathbf{\Gamma}_0^T\mathbf{Y}, \mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y},\mathbf{X}) \text{ is a function of } \mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y},\mathbf{X}\},$$
$$\Lambda_{3,\text{conj}}^{\perp} = \{f(\mathbf{Y},\mathbf{X}) : \mathbb{E}(f \mid \mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y}) = \mathbf{0}\}$$
$$\Lambda_{4,\text{conj}}^{\perp} = \{f(\mathbf{Y},\mathbf{X}) : \mathbb{E}(f \mid \mathbf{X}) = \mathbf{0}\}.$$

We only prove our conjecture is true for $\Lambda_1^{\perp}$.

For any $f_1^{\perp} \in \Lambda_{1,\text{conj}}^{\perp}$ and $f_1 \in \Lambda_1$,

$$\begin{aligned}\mathbb{E}\{f_1^T f_1^{\perp}\} &= \mathbb{E}[\mathbb{E}\{f_1^T f_1^{\perp} \mid \mathbf{\Gamma}^T\mathbf{Y},\mathbf{X}\}] \\ &= \mathbb{E}[f_1^T \mathbb{E}\{f_1^{\perp} \mid \mathbf{\Gamma}^T\mathbf{Y},\mathbf{X}\}] \\ &= \mathbb{E}\{f_1^T a(\mathbf{X})\} \\ &= \mathbb{E}\{\mathbb{E}(f_1^T \mid \mathbf{X}) a(\mathbf{X})\} = 0.\end{aligned}$$

Hence, we have $\Lambda_{1,\text{conj}}^{\perp} \perp \Lambda_1$. Then, we show that for any $f \in \Lambda_1^{\perp}$, f satisfies $\mathbb{E}(f \mid \mathbf{\Gamma}^T\mathbf{Y},\mathbf{X}) = a(\mathbf{X})$ for some function $a(\mathbf{X})$.

Firstly, define $g = \mathbb{E}(f \mid \mathbf{\Gamma}^T\mathbf{Y},\mathbf{X}) - \mathbb{E}(f \mid \mathbf{X})$. Clearly, $g \in \Lambda_1$. Thus,

$$\begin{aligned}0 = \mathbb{E}(g^T f) &= \mathbb{E}\{g^T \mathbb{E}(f \mid \mathbf{\Gamma}^T\mathbf{Y},\mathbf{X})\} \\ &= \mathbb{E}(g^T g) + \mathbb{E}\{g^T \mathbb{E}(f \mid \mathbf{X})\} = \mathbb{E}(g^T g).\end{aligned}$$

Therefore, $g = \mathbf{0}$, which implies $\mathbb{E}(f \mid \mathbf{\Gamma}^T\mathbf{Y},\mathbf{X}) = \mathbb{E}(f \mid \mathbf{X}) = a(\mathbf{X})$. We proved that $\Lambda_{1,\text{conj}} = \Lambda_1$.

Validity of the other three follows the same technique. ■

Lemma C.1.3 *The functions* $\mathbf{f}_1(\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y},\mathbf{X})$ *and* $\mathbf{f}_2(\mathbf{\Gamma}^T\mathbf{Y},\mathbf{B}_0^T\mathbf{\Gamma}_0^T\mathbf{Y})$ *defined above belongs to the orthogonal nuisance tangent space* $\Lambda^{\perp}$.

Proof of Lemma C.1.3. Recall that

$$f_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) = \{g_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}g_1\}\{h_1(\mathbf{X}) - \mathbb{E}h_1\} = \mathbf{0},$$

$$f_2(\boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) = \{g_2(\boldsymbol{\Gamma}^T\mathbf{Y}) - \mathbb{E}g_2\}\{h_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}h_2\} = \mathbf{0}.$$

We want to show that $f_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) \in \Lambda^\perp$ and $f_2(\boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) \in \Lambda^\perp$.

$$\begin{aligned}
\mathbb{E}(f_1 \mid \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) &= \mathbb{E}[\{g_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}g_1\}\{h_1(\mathbf{X}) - \mathbb{E}h_1\} \mid \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}] \\
&= \mathbb{E}\{g_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}g_1 \mid \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}\}\{h_1(\mathbf{X}) - \mathbb{E}h_1\} \\
&= \mathbb{E}\{g_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}g_1\}\{h_1(\mathbf{X}) - \mathbb{E}h_1\} \\
&= \mathbf{0}.
\end{aligned}$$

Hence, $f_1 \in \Lambda_1^\perp$. The second to the third equation is because $\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y} \perp\!\!\!\perp (\boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X})$.

$$\begin{aligned}
\mathbb{E}(f_1 \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) &= \mathbb{E}[\{g_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}g_1\}\{h_1(\mathbf{X}) - \mathbb{E}h_1\} \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}] \\
&= \{g_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}g_1\}\{h_1(\mathbf{X}) - \mathbb{E}h_1\}
\end{aligned}$$

is a function of $\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}$ and $\mathbf{X}$. Therefore, $f_1 \in \Lambda_2^\perp$. Notice that we also have

$$\begin{aligned}
\mathbb{E}(f_1 \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) &= \mathbb{E}[\{g_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}g_1\}\{h_1(\mathbf{X}) - \mathbb{E}h_1\} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}] \\
&= \{g_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}g_1\}\mathbb{E}\{h_1(\mathbf{X}) - \mathbb{E}h_1 \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}\} \\
&= \{g_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}g_1\}\mathbb{E}\{h_1(\mathbf{X}) - \mathbb{E}h_1\} \\
&= \mathbf{0},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}(f_1 \mid \mathbf{X}) &= \mathbb{E}[\{g_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}g_1\}\{h_1(\mathbf{X}) - \mathbb{E}h_1\} \mid \mathbf{X}] \\
&= \mathbb{E}[\{g_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}g_1 \mid \mathbf{X}\}\{h_1(\mathbf{X}) - \mathbb{E}h_1\} \\
&= \mathbb{E}\{g_1(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}g_1\}\{h_1(\mathbf{X}) - \mathbb{E}h_1\} \\
&= \mathbf{0}.
\end{aligned}$$

Therefore, $f_1 \in \Lambda_1^\perp \cap \Lambda_2^\perp \cap \Lambda_3^\perp \cap \Lambda_4^\perp = \Lambda^\perp$. Then we prove that $f_2(\boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) \in$

$\Lambda^{\perp}$.

$$\begin{aligned}\mathbb{E}(f_2 \mid \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) &= \mathbb{E}[\{g_2(\boldsymbol{\Gamma}^T\mathbf{Y}) - \mathbb{E}g_2\}\{h_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}h_2\} \mid \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}] \\ &= \{g_2(\boldsymbol{\Gamma}^T\mathbf{Y}) - \mathbb{E}g_2\}\mathbb{E}\{h_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}h_2 \mid \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}\} \\ &= \{g_2(\boldsymbol{\Gamma}^T\mathbf{Y}) - \mathbb{E}g_2\}\mathbb{E}\{h_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}h_2\} \\ &= \mathbf{0}.\end{aligned}$$

$$\begin{aligned}\mathbb{E}(f_2 \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) &= \mathbb{E}[\{g_2(\boldsymbol{\Gamma}^T\mathbf{Y}) - \mathbb{E}g_2\}\{h_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}h_2\} \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}] \\ &= \mathbb{E}\{g_2(\boldsymbol{\Gamma}^T\mathbf{Y}) - \mathbb{E}g_2 \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\}\{h_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}h_2\} \\ &= \mathbb{E}\{g_2(\boldsymbol{\Gamma}^T\mathbf{Y}) - \mathbb{E}g_2 \mid \mathbf{X}\}\{h_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}h_2\}\end{aligned}$$

is a function of $(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X})$.

$$\begin{aligned}\mathbb{E}(f_2 \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) &= \mathbb{E}[\{g_2(\boldsymbol{\Gamma}^T\mathbf{Y}) - \mathbb{E}g_2\}\{h_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}h_2\} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}] \\ &= \mathbb{E}\{g_2(\boldsymbol{\Gamma}^T\mathbf{Y}) - \mathbb{E}g_2 \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}\}\{h_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}h_2\} \\ &= \mathbb{E}\{g_2(\boldsymbol{\Gamma}^T\mathbf{Y}) - \mathbb{E}g_2\}\{h_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}h_2\} \\ &= \mathbf{0}.\end{aligned}$$

$$\begin{aligned}\mathbb{E}(f_2 \mid \mathbf{X}) &= \mathbb{E}[\{g_2(\boldsymbol{\Gamma}^T\mathbf{Y}) - \mathbb{E}g_2\}\{h_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}h_2\} \mid \mathbf{X}] \\ &= \mathbb{E}\{g_2(\boldsymbol{\Gamma}^T\mathbf{Y}) - \mathbb{E}g_2 \mid \mathbf{X}\}\mathbb{E}\{h_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}h_2 \mid \mathbf{X}\} \\ &= \mathbb{E}\{g_2(\boldsymbol{\Gamma}^T\mathbf{Y}) - \mathbb{E}g_2 \mid \mathbf{X}\}\mathbb{E}\{h_2(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}h_2\} \\ &= \mathbf{0}.\end{aligned}$$

The first to the second equation is because $\boldsymbol{\Gamma}^T\mathbf{Y} \perp\!\!\!\perp \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{X}$. Hence, we also have $f_2 \in \Lambda^{\perp}$. ■

Lemma C.1.4 *The tangent space generated by the score vector with respect to the parameter of interest* $\boldsymbol{\theta}$ *is* $\mathscr{T}_{\boldsymbol{\theta}} = \{\mathbf{M}S_{\boldsymbol{\theta}}$ *for all* $\mathbf{M} \in \mathbb{R}^{q\times q}\}$. *The score vector* $S_{\boldsymbol{\theta}} = (S_{\boldsymbol{\gamma}}^T, S_{\mathbf{b}}^T)^T$,

$$\begin{aligned}S_{\boldsymbol{\gamma}} =& vec^T\left\{\mathbf{Y}\frac{\partial \log \eta_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T}\right\}\frac{\partial vec(\boldsymbol{\Gamma})}{\partial \boldsymbol{\gamma}^T} + vec^T\left\{\mathbf{Y}\frac{\partial \log \eta_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\mathbf{B}_0^T\right\}\frac{\partial vec(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T} \\ &+ vec^T\left\{\mathbf{Y}\frac{\partial \log \eta_2}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\mathbf{B}^T + \mathbf{Y}\frac{\partial \log \eta_2}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\mathbf{B}_0^T\right\}\frac{\partial vec(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T}\end{aligned}$$

and

$$S_{\mathbf{b}} = vec^T\left\{\boldsymbol{\Gamma}_0^T\mathbf{Y}\frac{\partial \log \eta_2}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\right\}\frac{\partial vec(\mathbf{B})}{\partial \mathbf{b}^T} + vec^T\left\{\boldsymbol{\Gamma}_0^T\mathbf{Y}\frac{\partial \log \eta_2}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\right\}\frac{\partial vec(\mathbf{B}_0)}{\partial \mathbf{b}^T}$$
$$+ vec^T\left\{\boldsymbol{\Gamma}_0^T\mathbf{Y}\frac{\partial \log \eta_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\right\}\frac{\partial vec(\mathbf{B}_0)}{\partial \mathbf{b}^T}.$$

Proof of Lemma C.1.4. Define $l(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Y})$ as the log-likelihood of $(\mathbf{X}, \mathbf{Y})$, i.e.,

$$l(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Y}) = \log \eta_1(\boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) + \log \eta_2(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) + \log \eta_3(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) + \log \eta_4(\mathbf{X}).$$

Then, the score with respect to $\boldsymbol{\gamma}$ is as follows:

$$S_\gamma = \frac{\partial l(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Y})}{\partial \boldsymbol{\gamma}^T} = \frac{\partial \log \eta_1}{\partial \boldsymbol{\gamma}^T} + \frac{\partial \log \eta_2}{\partial \boldsymbol{\gamma}^T} + \frac{\partial \log \eta_3}{\partial \boldsymbol{\gamma}^T}.$$

Firstly,

$$\frac{\partial \log \eta_1}{\partial \boldsymbol{\gamma}^T} = \frac{\partial \log \eta_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T}\frac{\partial \boldsymbol{\Gamma}^T\mathbf{Y}}{\partial \mathrm{vec}(\boldsymbol{\Gamma})^T}\frac{\partial \mathrm{vec}(\boldsymbol{\Gamma})}{\partial \boldsymbol{\gamma}^T}.$$

It's easy to verify that

$$\frac{\partial \boldsymbol{\Gamma}^T\mathbf{Y}}{\partial \mathrm{vec}(\boldsymbol{\Gamma})^T} = \mathbf{I}_u \otimes \mathbf{Y}^T.$$

Hence,

$$\frac{\partial \log \eta_1}{\partial \boldsymbol{\gamma}^T} = \frac{\partial \log \eta_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T}(\mathbf{I}_u \otimes \mathbf{Y}^T)\frac{\partial \mathrm{vec}(\boldsymbol{\Gamma})}{\partial \boldsymbol{\gamma}^T} = \mathrm{vec}^T\left\{\mathbf{Y}\frac{\partial \log \eta_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T}\right\}\frac{\partial \mathrm{vec}(\boldsymbol{\Gamma})}{\partial \boldsymbol{\gamma}^T}.$$

Secondly,

$$\frac{\partial \log \eta_2}{\partial \boldsymbol{\gamma}^T} = \frac{\partial \log \eta_2}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\frac{\partial \mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}}{\partial \boldsymbol{\gamma}^T} + \frac{\partial \log \eta_2}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\frac{\partial \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}}{\partial \boldsymbol{\gamma}^T}.$$

We know that $\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y} = \mathrm{vec}(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) = \mathrm{vec}(\mathbf{Y}^T\boldsymbol{\Gamma}_0\mathbf{B}) = (\mathbf{B}^T \otimes \mathbf{Y}^T)\mathrm{vec}(\boldsymbol{\Gamma}_0)$. Thus,

$$\frac{\partial \mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}}{\partial \boldsymbol{\gamma}^T} = (\mathbf{B}^T \otimes \mathbf{Y}^T)\frac{\partial \mathrm{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T}.$$

Therefore,

$$\begin{aligned}\frac{\partial \log \eta_2}{\partial \boldsymbol{\gamma}^T} =& \left\{ \frac{\partial \log \eta_2}{\partial (\mathbf{B}^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} (\mathbf{B}^T \otimes \mathbf{Y}^T) + \frac{\partial \log \eta_2}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} (\mathbf{B}_0^T \otimes \mathbf{Y}^T) \right\} \frac{\partial \text{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T} \\ =& \text{vec}^T \left\{ \mathbf{Y} \frac{\partial \log \eta_2}{\partial (\mathbf{B}^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \mathbf{B}^T + \mathbf{Y} \frac{\partial \log \eta_2}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \mathbf{B}_0^T \right\} \frac{\partial \text{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T}\end{aligned}$$

Similarly, we can show that

$$\frac{\partial \log \eta_3}{\partial \boldsymbol{\gamma}^T} = \frac{\partial \log \eta_3}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} (\mathbf{B}_0^T \otimes \mathbf{Y}^T) \frac{\partial \text{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T} = \text{vec}^T \left\{ \mathbf{Y} \frac{\partial \log \eta_3}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \mathbf{B}_0^T \right\} \frac{\partial \text{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T}.$$

Consequently, the score for $\boldsymbol{\gamma}$ is

$$\begin{aligned}S_\gamma =& \text{vec}^T \left\{ \mathbf{Y} \frac{\partial \log \eta_1}{\partial (\boldsymbol{\Gamma}^T \mathbf{Y})^T} \right\} \frac{\partial \text{vec}(\boldsymbol{\Gamma})}{\partial \boldsymbol{\gamma}^T} + \text{vec}^T \left\{ \mathbf{Y} \frac{\partial \log \eta_3}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \mathbf{B}_0^T \right\} \frac{\partial \text{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T} \\ &+ \text{vec}^T \left\{ \mathbf{Y} \frac{\partial \log \eta_2}{\partial (\mathbf{B}^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \mathbf{B}^T + \mathbf{Y} \frac{\partial \log \eta_2}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \mathbf{B}_0^T \right\} \frac{\partial \text{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T}\end{aligned}$$

Next, we calculate the score with respect to $\mathbf{b}$.

$$\frac{\partial l(\boldsymbol{\gamma} \mid \mathbf{X}, \mathbf{Y})}{\partial \mathbf{b}^T} = \frac{\partial \log \eta_2}{\partial \mathbf{b}^T} + \frac{\partial \log \eta_3}{\partial \mathbf{b}^T}.$$

By the chain rule, we have

$$\frac{\partial \log \eta_2}{\partial \mathbf{b}^T} = \frac{\partial \log \eta_2}{\partial (\mathbf{B}^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \frac{\partial \mathbf{B}^T \boldsymbol{\Gamma}_0^T \mathbf{Y}}{\partial \mathbf{b}^T} + \frac{\partial \log \eta_2}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \frac{\partial \mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y}}{\partial \mathbf{b}^T}.$$

Recall that $\mathbf{B} \in \mathbb{R}^{(r-u)\times d}$ and $\mathbf{B}_0 \in \mathbb{R}^{(r-u)\times(r-u-d)}$, where d is the dimension for $\mathcal{S}_2$. By the same "vec" trick, we have

$$\mathbf{B}^T \boldsymbol{\Gamma}_0^T \mathbf{Y} = \text{vec}(\mathbf{B}^T \boldsymbol{\Gamma}_0^T \mathbf{Y}) = \text{vec}(\mathbf{Y}^T \boldsymbol{\Gamma}_0 \mathbf{B} \mathbf{I}_d) = (\mathbf{I}_d \otimes \mathbf{Y}^T \boldsymbol{\Gamma}_0) \text{vec}(\mathbf{B}).$$

Hence,

$$\begin{aligned}\frac{\partial \log \eta_2}{\partial \mathbf{b}^T} =& \frac{\partial \log \eta_2}{\partial (\mathbf{B}^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} (\mathbf{I}_d \otimes \mathbf{Y}^T \boldsymbol{\Gamma}_0) \frac{\partial \text{vec}(\mathbf{B})}{\partial \mathbf{b}^T} + \frac{\partial \log \eta_2}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} (\mathbf{I}_{r-u-d} \otimes \mathbf{Y}^T \boldsymbol{\Gamma}_0) \frac{\partial \text{vec}(\mathbf{B}_0)}{\partial \mathbf{b}^T} \\ =& \text{vec}^T \left\{ \boldsymbol{\Gamma}_0^T \mathbf{Y} \frac{\partial \log \eta_2}{\partial (\mathbf{B}^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \right\} \frac{\partial \text{vec}(\mathbf{B})}{\partial \mathbf{b}^T} + \text{vec}^T \left\{ \boldsymbol{\Gamma}_0^T \mathbf{Y} \frac{\partial \log \eta_2}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \right\} \frac{\partial \text{vec}(\mathbf{B}_0)}{\partial \mathbf{b}^T}.\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{\partial \log \eta_3}{\partial \mathbf{b}^T} =& \frac{\partial \log \eta_3}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} (\mathbf{I}_{r-u-d} \otimes \mathbf{Y}^T \boldsymbol{\Gamma}_0) \frac{\partial \text{vec}(\mathbf{B}_0)}{\partial \mathbf{b}^T} \\ =& \text{vec}^T \left\{ \boldsymbol{\Gamma}_0^T \mathbf{Y} \frac{\partial \log \eta_3}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \right\} \frac{\partial \text{vec}(\mathbf{B}_0)}{\partial \mathbf{b}^T}.\end{aligned}$$

Therefore,

$$\begin{aligned}S_{\mathbf{b}} =& \text{vec}^T \left\{ \boldsymbol{\Gamma}_0^T \mathbf{Y} \frac{\partial \log \eta_2}{\partial (\mathbf{B}^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \right\} \frac{\partial \text{vec}(\mathbf{B})}{\partial \mathbf{b}^T} + \text{vec}^T \left\{ \boldsymbol{\Gamma}_0^T \mathbf{Y} \frac{\partial \log \eta_2}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \right\} \frac{\partial \text{vec}(\mathbf{B}_0)}{\partial \mathbf{b}^T} \\ &+ \text{vec}^T \left\{ \boldsymbol{\Gamma}_0^T \mathbf{Y} \frac{\partial \log \eta_3}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})^T} \right\} \frac{\partial \text{vec}(\mathbf{B}_0)}{\partial \mathbf{b}^T}.\end{aligned}$$

The score vector can be formulated as $S_{\boldsymbol{\theta}} = (S_{\boldsymbol{\gamma}}, S_{\mathbf{b}})$. ■

Regularity Conditions and Proof of Theorem 4.3.1. Let $\boldsymbol{\Theta}$ denote the parameter space that contains $\boldsymbol{\theta}$, and let $\boldsymbol{\theta}_0$ denote the true parameter value. Firstly, we state the regularity conditions needed for Theorem 4.3.1:

(S1) $\mathbb{E}\{\mathbf{f}(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})\} \neq \mathbf{0}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Also, $\boldsymbol{\theta}_0$ is an interior point in $\boldsymbol{\Theta}$.

(S2) The parameter space $\boldsymbol{\Theta}$ is a compact set.

(S3) $\mathbb{E}\{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\mathbf{f}(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})\|_2\} < \infty$.

(S4) $\mathbf{C}_1$ is continuous on some neighborhood $\boldsymbol{\Theta}_\epsilon$ of $\boldsymbol{\theta}_0$, where $\mathbf{C}_1 = \mathbb{E}\{\partial \text{vec}(\mathbf{f}) / \partial \boldsymbol{\theta}^{\mathrm{T}}\}$.

(S5) The matrix $\mathbf{C}_1$ is of full column rank.

(S6) $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_\epsilon} \|n^{-1} \sum_{i=1}^N \partial \text{vec}(\mathbf{f}) / \partial \boldsymbol{\theta}^T - \mathbf{C}_1\| \xrightarrow{p} \mathbf{0}$.

Regularity conditions (S1)–(S5) are standard conditions for the GMM estimators. Based on these equations, by Theorem 3.2 in Hall (2004), we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}\big\{\mathbf{0}, (\mathbf{C}_1^T \mathbf{C}_1)^{-1} \mathbf{C}_1^T \mathbf{D}_1 \mathbf{C}_1 (\mathbf{C}_1^T \mathbf{C}_1)^{-1}\big\},$$

where $\mathbf{D}_1 = \text{Var}\{\mathbf{f}(\mathbf{Y}_i, \mathbf{X}_i; \boldsymbol{\theta})\}$. ■

Derivation of the efficient score S_{eff}. Since $S_{\boldsymbol{\theta}}$ satisfies $\mathbb{E}(S_{\boldsymbol{\theta}}) = 0$, we only need to consider the projection of any mean zero function $h(\mathbf{Y}, \mathbf{X})$. Let $\Pi(h \mid \Lambda^\perp)$ denote the projection of h onto $\Lambda^\perp$. Then, by Lemma C.1.1,

$$\Pi(h \mid \Lambda^\perp) = h - \Pi(h \mid \Lambda) = h - \Pi(h \mid \Lambda_1) - \Pi(h \mid \Lambda_2) - \Pi(h \mid \Lambda_3) - \Pi(h \mid \Lambda_4).$$

Notice that $\Pi(h \mid \Lambda_1) = \mathbb{E}(h \mid \boldsymbol{\Gamma}^T \mathbf{Y}, \mathbf{X}) - \mathbb{E}(h \mid \mathbf{X})$, $\Pi(h \mid \Lambda_2) = \mathbb{E}(h \mid \boldsymbol{\Gamma}_0^T \mathbf{Y}, \mathbf{X}) - \mathbb{E}(h \mid \mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y}, \mathbf{X})$, $\Pi(h \mid \Lambda_3) = \mathbb{E}(h \mid \mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{Y})$ and $\Pi(h \mid \Lambda_4) = \mathbb{E}(h \mid \mathbf{X})$.

Therefore,

$$S_{eff} = S_{\boldsymbol{\theta}} - \mathbb{E}(S_{\boldsymbol{\theta}} \mid \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) - \mathbb{E}(S_{\boldsymbol{\theta}} \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) + \mathbb{E}(S_{\boldsymbol{\theta}} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) - \mathbb{E}(S_{\boldsymbol{\theta}} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}).$$

Notice that

$$S_{\boldsymbol{\theta}} = \frac{\partial \log \eta_1}{\partial \boldsymbol{\theta}} + \frac{\partial \log \eta_2}{\partial \boldsymbol{\theta}} + \frac{\partial \log \eta_3}{\partial \boldsymbol{\theta}}.$$

Firstly, since $\eta_1(\cdot)$ is the conditional distribution of $\boldsymbol{\Gamma}^T\mathbf{Y} \mid \mathbf{X}$, and the model assumption $\boldsymbol{\Gamma}^T\mathbf{Y} \perp \boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{X}$ holds, we must have

$$\mathbb{E}\left(\frac{\partial \log \eta_1}{\partial \boldsymbol{\theta}} \;\middle|\; \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\right) = \mathbf{0}.$$

As a consequence,

$$\mathbb{E}\left(\frac{\partial \log \eta_1}{\partial \boldsymbol{\theta}} \;\middle|\; \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\right) = \mathbf{0},$$

and

$$\mathbb{E}\left(\frac{\partial \log \eta_1}{\partial \boldsymbol{\theta}} \;\middle|\; \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}\right) = \mathbf{0}.$$

Similarly, for $\eta_2(\cdot)$, we have

$$\mathbb{E}\left(\frac{\partial \log \eta_2}{\partial \boldsymbol{\theta}} \;\middle|\; \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}\right) = \mathbb{E}\left(\frac{\partial \log \eta_2}{\partial \boldsymbol{\theta}} \;\middle|\; \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\right) = \mathbb{E}\left(\frac{\partial \log \eta_2}{\partial \boldsymbol{\theta}} \;\middle|\; \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}\right) = \mathbf{0}.$$

Also, because $\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y} \perp\!\!\!\perp (\boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X})$,

$$\mathbb{E}\left(\frac{\partial \log \eta_3}{\partial \boldsymbol{\theta}} \;\middle|\; \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}\right) = \mathbf{0}.$$

Hence,

$$\begin{aligned} S_{eff,\boldsymbol{\gamma}} = {} & \frac{\partial \log \eta_1}{\partial \boldsymbol{\gamma}^T} + \frac{\partial \log \eta_2}{\partial \boldsymbol{\gamma}^T} + \frac{\partial \log \eta_3}{\partial \boldsymbol{\gamma}^T} - \mathbb{E}\left(\frac{\partial \log \eta_1}{\partial \boldsymbol{\gamma}} \;\middle|\; \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}\right) - \mathbb{E}\left(\frac{\partial \log \eta_2}{\partial \boldsymbol{\gamma}} \;\middle|\; \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\right) \\ & - \mathbb{E}\left(\frac{\partial \log \eta_3}{\partial \boldsymbol{\gamma}} \;\middle|\; \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\right) + \mathbb{E}\left(\frac{\partial \log \eta_3}{\partial \boldsymbol{\gamma}} \;\middle|\; \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\right) - \mathbb{E}\left(\frac{\partial \log \eta_3}{\partial \boldsymbol{\gamma}} \;\middle|\; \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}\right). \end{aligned}$$

Notably,

$$
\begin{aligned}
&\frac{\partial \log \eta_1}{\partial \boldsymbol{\gamma}^T} - \mathbb{E}\left(\frac{\partial \log \eta_1}{\partial \boldsymbol{\gamma}} \;\middle|\; \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}\right) \\
=&\mathrm{vec}^T\left[\{\mathbf{Y} - \mathbb{E}(\mathbf{Y} \mid \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X})\}\frac{\partial \log \eta_1}{\partial (\boldsymbol{\Gamma}^T\mathbf{Y})^T}\right]\frac{\partial \mathrm{vec}(\boldsymbol{\Gamma})}{\partial \boldsymbol{\gamma}^T}, \\
&\frac{\partial \log \eta_2}{\partial \boldsymbol{\gamma}^T} - \mathbb{E}\left(\frac{\partial \log \eta_2}{\partial \boldsymbol{\gamma}} \;\middle|\; \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\right) \\
=&\mathrm{vec}^T\left[\left\{\mathbf{Y} - \mathbb{E}(\mathbf{Y} \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X})\right\}\left\{\frac{\partial \log \eta_2}{\partial (\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\mathbf{B}^T + \frac{\partial \log \eta_2}{\partial (\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\mathbf{B}_0^T\right\}\right]\frac{\partial \mathrm{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T}, \\
&\frac{\partial \log \eta_3}{\partial \boldsymbol{\gamma}^T} - \mathbb{E}\left(\frac{\partial \log \eta_3}{\partial \boldsymbol{\gamma}} \;\middle|\; \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\right) + \mathbb{E}\left(\frac{\partial \log \eta_3}{\partial \boldsymbol{\gamma}} \;\middle|\; \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\right) - \mathbb{E}\left(\frac{\partial \log \eta_3}{\partial \boldsymbol{\gamma}} \;\middle|\; \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}\right) \\
=&\mathrm{vec}^T\left\{\left\{\mathbf{Y} - \mathbb{E}(\mathbf{Y} \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) + \mathbb{E}(\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) - \mathbb{E}(\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})\right\}\frac{\partial \log \eta_3}{\partial (\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\mathbf{B}_0^T\right\}\frac{\partial \mathrm{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T}
\end{aligned}
$$

In addition,

$$
\begin{aligned}
\mathbf{Y} - \mathbb{E}(\mathbf{Y} \mid \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) &= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} + \mathbf{Q}_{\boldsymbol{\Gamma}} - \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) - \mathbb{E}(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) \\
&= \mathbf{Q}_{\boldsymbol{\Gamma}} - \mathbb{E}(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X}) \\
&= \mathbf{Q}_{\boldsymbol{\Gamma}} - \mathbb{E}(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \mathbf{X}),
\end{aligned}
$$

where the last equation is because $\boldsymbol{\Gamma}^T\mathbf{Y} \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{Y} \mid \mathbf{X}$.

$$
\begin{aligned}
\mathbf{Y} - \mathbb{E}(\mathbf{Y} \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) &= \mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} + \mathbf{Q}_{\boldsymbol{\Gamma}} - \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) - \mathbb{E}(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) \\
&= \mathbf{P}_{\boldsymbol{\Gamma}} - \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) \\
&= \mathbf{P}_{\boldsymbol{\Gamma}} - \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \mathbf{X}).
\end{aligned}
$$

$$
\begin{aligned}
&\mathbf{Y} - \mathbb{E}(\mathbf{Y} \mid \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) + \mathbb{E}(\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) - \mathbb{E}(\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) \\
=&\mathbf{P}_{\boldsymbol{\Gamma}} - \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \mathbf{X}) + \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) + \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) + \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}_0}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) \\
&- \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) - \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}_0}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) \\
=&\mathbf{P}_{\boldsymbol{\Gamma}} + \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) - \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}).
\end{aligned}
$$

The last equation is because

$$\begin{aligned}
&\mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) = \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \mathbf{X}),\\
&\mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}_0}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) - \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}_0}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) = \mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}_0}\mathbf{Y} - \mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}_0}\mathbf{Y} = \mathbf{0},\\
&\mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}) = \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y}) = \mathbf{P}_{\boldsymbol{\Gamma}}\mathbb{E}(\mathbf{Y}) = \mathbf{0}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
S_{\text{eff},\boldsymbol{\gamma}} = \text{vec}^T\left\{\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\Delta}_1\frac{\partial \log \eta_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T}\right\}\frac{\partial \text{vec}(\boldsymbol{\Gamma})}{\partial \boldsymbol{\gamma}^T} + \text{vec}^T\left\{(\mathbf{P}_{\boldsymbol{\Gamma}}\mathbf{Y} + \boldsymbol{\Delta}_2)\frac{\partial \log \eta_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\mathbf{B}_0^T\right\}\frac{\partial \text{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T}\\
+ \text{vec}^T\left[\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Delta}_1\left\{\frac{\partial \log \eta_2}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\mathbf{B}^T + \frac{\partial \log \eta_2}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\mathbf{B}_0^T\right\}\right]\frac{\partial \text{vec}(\boldsymbol{\Gamma}_0)}{\partial \boldsymbol{\gamma}^T},
\end{aligned}$$

where $\boldsymbol{\Delta}_1 = \mathbf{Y} - \mathbb{E}(\mathbf{Y} \mid \mathbf{X})$, and $\boldsymbol{\Delta}_2 = \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}) - \mathbb{E}(\mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}}\mathbf{Y} \mid \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})$. Similarly,

$$\begin{aligned}
S_{\text{eff},\mathbf{b}} &= S_{\mathbf{b}} - \mathbb{E}\left(\frac{\partial \log \eta_2}{\partial \mathbf{b}} \,\middle|\, \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\right) - \mathbb{E}\left(\frac{\partial \log \eta_3}{\partial \mathbf{b}} \,\middle|\, \boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\right)\\
&\quad + \mathbb{E}\left(\frac{\partial \log \eta_3}{\partial \mathbf{b}} \,\middle|\, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}, \mathbf{X}\right) - \mathbb{E}\left(\frac{\partial \log \eta_3}{\partial \mathbf{b}} \,\middle|\, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}\right)\\
&= \text{vec}^T\left\{\boldsymbol{\Gamma}_0^T\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Delta}_1\frac{\partial \log \eta_2}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\right\}\frac{\partial \text{vec}(\mathbf{B})}{\partial \mathbf{b}^T} + \text{vec}^T\left\{\boldsymbol{\Gamma}_0^T\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Delta}_1\frac{\partial \log \eta_2}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\right\}\frac{\partial \text{vec}(\mathbf{B}_0)}{\partial \mathbf{b}^T}\\
&\quad + \text{vec}^T\left\{\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_2\frac{\partial \log \eta_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\right\}\frac{\partial \text{vec}(\mathbf{B}_0)}{\partial \mathbf{b}^T}.
\end{aligned}$$

Since $\boldsymbol{\Gamma}_0^T\mathbf{P}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}_0^T\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = 0$, $S_{\text{eff},\mathbf{b}}$ can be further simplified as

$$S_{\text{eff},\mathbf{b}} = \text{vec}^T\left\{\boldsymbol{\Gamma}_0^T\boldsymbol{\Delta}_2\frac{\partial \log \eta_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y})^T}\right\}\frac{\partial \text{vec}(\mathbf{B}_0)}{\partial \mathbf{b}^T}$$

■

Before proving the following theorem, we first prove a lemma.

Lemma C.1.5 *Under regularity conditions (A1)–(A5), we have*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\boldsymbol{\Delta}_{i1}\left\{\frac{\partial\log\hat{\eta}_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y}_i)^T}-\frac{\partial\log\eta_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y}_i)^T}\right\}\right]=O_p(h^2+n^{1/2}h^4+n^{-1/2}h^{-(u+p+1)}\log n),$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\boldsymbol{\Delta}_{i1}\left\{\frac{\partial\log\hat{\eta}_2}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}-\frac{\partial\log\eta_2}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right\}\right]=O_p(h^2+n^{1/2}h^4+n^{-1/2}h^{-(r+p-u+1)}\log n),$$

and $$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)\left\{\frac{\partial\log\hat{\eta}_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}-\frac{\partial\log\eta_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right\}\right]=O_p(h^2+n^{-1/2}h^4+n^{-1/2}h^{-(r-u-d+1)}\log$$

Under regularity conditions (B1)–(B4) and (A4)–(A5), we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}vec^T\left[\boldsymbol{\Gamma}_0^T\left\{\hat{\boldsymbol{\Delta}}_{i2}(\boldsymbol{\theta}_0)-\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)\right\}\left\{\frac{\partial\log\eta_3^*}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}-\mathbb{E}\left(\frac{\partial\log\eta_3^*}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right)\right\}\right]\\ =&O_p(h^2+n^{-1/2}h^{p+r-u-d}\log n+n^{1/2}h^4),\end{aligned}$$

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{P}_{\boldsymbol{\Gamma}}\{\hat{\boldsymbol{\Delta}}_{i1}-\boldsymbol{\Delta}_{i1}\}\left\{\frac{\partial\log\eta_2^*}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}-\mathbb{E}\left(\frac{\partial\log\eta_2^*}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\;\middle|\;\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i,\mathbf{X}_i\right)\right\}\\ =&O_p(h^2+n^{-1/2}h^p\log n+n^{1/2}h^4),\end{aligned}$$

where

$$\begin{aligned}\hat{\boldsymbol{\Delta}}_{i1}&=\mathbf{Y}_i-\hat{\mathrm{E}}(\mathbf{Y}_i\mid\mathbf{X}_i)\\ \hat{\boldsymbol{\Delta}}_{i2}(\boldsymbol{\theta}_0)&=\mathbf{P}_{\boldsymbol{\Gamma}_0\mathbf{B}}\{\hat{\mathrm{E}}(\mathbf{Y}_i\mid\mathbf{B}_0\boldsymbol{\Gamma}_0^T\mathbf{Y}_i,\mathbf{X}_i)-\hat{\mathrm{E}}(\mathbf{Y}_i\mid\mathbf{B}_0\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)\}.\end{aligned}$$

Proof of Lemma C.1.5. Since the equalities and their proofs are similar, we only show the proof of the first one. Recall the kernel density estimation of $\partial\eta_1/\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T$ has the form

$$\frac{\partial\hat{\eta}_1(\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X})}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T}=\frac{\sum_{i=1}^{n}K_h'(\boldsymbol{\Gamma}^T\mathbf{Y}-\boldsymbol{\Gamma}^T\mathbf{Y}_i)K_h(\mathbf{X}-\mathbf{X}_i)}{\sum_{i=1}^{n}K_h(\mathbf{X}-\mathbf{X}_i)}.$$

Hence,

$$\frac{\partial\log\hat{\eta}_1(\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X})}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T}=\frac{\sum_{i=1}^{n}K_h'(\boldsymbol{\Gamma}^T\mathbf{Y}-\boldsymbol{\Gamma}^T\mathbf{Y}_i)K_h(\mathbf{X}-\mathbf{X}_i)}{\sum_{i=1}^{n}K_h(\boldsymbol{\Gamma}^T\mathbf{Y}-\boldsymbol{\Gamma}^T\mathbf{Y}_i)K_h(\mathbf{X}-\mathbf{X}_i)}.$$

Let $\hat{f}(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)=n^{-1}\sum_{j=1}^{n}K_h(\boldsymbol{\Gamma}^T\mathbf{Y}_i-\boldsymbol{\Gamma}^T\mathbf{Y}_j)K_h(\mathbf{X}_i-\mathbf{X}_j)$ and $\hat{\mathbf{r}}_1(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)=n^{-1}\sum_{j=1}^{n}K_h'(\boldsymbol{\Gamma}^T\mathbf{Y}_i-\boldsymbol{\Gamma}^T\mathbf{Y}_j)K_h(\mathbf{X}_i-\mathbf{X}_j)$. Also, let $\mathbf{r}_1(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)=\partial\eta_1(\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X})/\partial(\boldsymbol{\Gamma}^T\mathbf{Y})^T$.

Hence, we have

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Delta}_{i1}\left\{\frac{\partial\log\hat{\eta}_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y}_i)^T}-\frac{\partial\log\eta_1}{\partial(\boldsymbol{\Gamma}^T\mathbf{Y}_i)^T}\right\} \\
= & \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Delta}_{i1}\left[\frac{\hat{\mathbf{r}}_1(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)}{\hat{f}(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)}-\frac{\mathbf{r}_1(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)}{f(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)}\right] \\
= & \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Delta}_{i1}\left[\frac{\hat{\mathbf{r}}_1(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)-\mathbf{r}_1(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)}{f(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)}\right] \\
& -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Delta}_{i1}\left[\frac{\mathbf{r}_1(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\{\hat{f}(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)-f(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\}}{f^2(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)}\right]+o_p(1) \\
& -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Delta}_{i1}\left[\frac{\{\hat{\mathbf{r}}_1(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)-\mathbf{r}_1(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\}\{\hat{f}(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)-f(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\}}{f(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\hat{f}(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)}\right] \\
& +\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Delta}_{i1}\left[\frac{\mathbf{r}_1(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\{\hat{f}(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)-f(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\}^2}{f^2(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\hat{f}(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)}\right].
\end{aligned}
$$

The second equation is because $\|\hat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}\|\le Cn^{-1/2}$. By the uniform convergence of nonparametric regression Li and Racine (2007), we have

$$
\sup_{\mathbf{X},\mathbf{Y}}|\hat{f}(\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X})-f(\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X})|=O_p(\sqrt{n^{-1}h^{-(p+u)}\log n}+h^2)
$$

and

$$
\sup_{\mathbf{X},\mathbf{Y}}|\hat{\mathbf{r}}_1(\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X})-\mathbf{r}_1(\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X})|=O_p(\sqrt{n^{-1}h^{-(p+u+2)}\log n}+h^2).
$$

Therefore, the third quantity can be bounded by

$$
\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Delta}_{i1}\left[\frac{\{\hat{\mathbf{r}}_1(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)-\mathbf{r}_1(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\}\{\hat{f}(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)-f(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\}}{f(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\hat{f}(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)}\right]\right| \\
\le & \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{\Delta}_{i1}|\left[\frac{\sup_{\mathbf{X},\mathbf{Y}}|\hat{\mathbf{r}}_1(\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X})-\mathbf{r}_1(\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X})|\sup_{\mathbf{X},\mathbf{Y}}|\hat{f}(\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X})-f(\boldsymbol{\Gamma}^T\mathbf{Y},\mathbf{X})|}{|f(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\hat{f}(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)|}\right] \\
= & \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{\Delta}_{i1}|\left[\frac{O_p(n^{-1}h^{-(p+u+1)}\log n+h^4)}{|f(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\hat{f}(\boldsymbol{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)|}\right] \\
= & O_p(n^{-1}h^{-(p+u+1)}\log n+h^4)\cdot\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{\Delta}_{i1}| \\
= & O_p(n^{-1}h^{-(p+u+1)}\log n+h^4).
\end{aligned}
$$

The second to the third equation is because

$$|f(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\hat{f}(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)|^{-1} = |f(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\{f(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)+o_p(1)\}|^{-1} = O_p(1).$$

Using the exact same technique, the fourth quantity in the above decomposition is of order $O_p(n^{-1}h^{-(p+u)}\log n + h^4)$.

Notice that the first two quantities also share the same structure, in the sequel we only deal with the first quantity.

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Delta}_{i1}\left[\frac{\hat{\mathbf{r}}_1(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)-\mathbf{r}_1(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)}{f(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)}\right]\\
= \;& O_p(1)\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Delta}_{i1}\big[\hat{\mathbf{r}}_1(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)-\mathbb{E}\{\hat{\mathbf{r}}_1(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\mid\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i\}\big]\\
&+O_p(1)\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Delta}_{i1}\big[\mathbb{E}\{\hat{\mathbf{r}}_1(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\mid\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i\}-\mathbf{r}_1(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\big].
\end{aligned}$$

We can write

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Delta}_{i1}\hat{\mathbf{r}}_1(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i) = \frac{1}{n(n-1)}\sum_{i\neq j}K_h'(\mathbf{\Gamma}^T\mathbf{Y}_i-\mathbf{\Gamma}^T\mathbf{Y}_j)K_h(\mathbf{X}_i-\mathbf{X}_j)\{\mathbf{\Delta}_{i1}+\mathbf{\Delta}_{j1}\}+O_p(n^{-1})$$

as a second order U-statistic with kernel function $K_h'(\mathbf{\Gamma}^T\mathbf{Y}_i-\mathbf{\Gamma}^T\mathbf{Y}_j)K_h(\mathbf{X}_i-\mathbf{X}_j)\{\mathbf{\Delta}_{i1}+\mathbf{\Delta}_{j1}\}$. Hence, by Lemma 5.2.1.A of Serfling (2009), the degenerated U-statistic has the rate

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Delta}_{i1}\big[\hat{\mathbf{r}}_1(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)-\mathbb{E}\{\hat{\mathbf{r}}_1(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\mid\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i\}\big] = O_p(n^{-1}h^{-(p+u+1)}).$$

Notice that $\mathbb{E}\{\hat{\mathbf{r}}_1(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\mid\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i\}-\mathbf{r}_1(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)$ is the bias term in non-parametric regression, hence

$$\sup_{\mathbf{X},\mathbf{Y}}|\mathbb{E}\{\hat{\mathbf{r}}_1(\mathbf{\Gamma}^T\mathbf{Y},\mathbf{X})\mid\mathbf{\Gamma}^T\mathbf{Y},\mathbf{X}\}-\mathbf{r}_1(\mathbf{\Gamma}^T\mathbf{Y},\mathbf{X})| = O_p(h^2).$$

Therefore

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Delta}_{i1}\big[\mathbb{E}\{\hat{\mathbf{r}}_1(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\mid\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i\}-\mathbf{r}_1(\mathbf{\Gamma}^T\mathbf{Y}_i,\mathbf{X}_i)\big] = O_p(n^{-1/2}h^2).$$

Combine the above results, we have the desired order $O_p(n^{-1/2}h^2+h^4+n^{-1}h^{-(p+u+1)}\log n)$.

■

Proof of Theorem 4.4.1. Similar to the proof of Theorem 3, we use Lemma 6 and 7 to prove the consistency and asymptotic normality. We only check condition (iii) in Lemma 7. All other conditions can be proved using the same way as the proof of Theorem 2. Consider the functions $\hat{Q}_n(\boldsymbol{\theta})$ and $Q_0(\boldsymbol{\theta})$ as

$$Q_0(\boldsymbol{\theta}) = \frac{1}{2}\left\{\frac{1}{n}\sum_{i=1}^{n} S_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})\right\}^2$$

$$\hat{Q}_n(\boldsymbol{\theta}) = \frac{1}{2}\left\{\frac{1}{n}\sum_{i=1}^{n} \hat{S}_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \hat{\eta}_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta})\right\}^2.$$

We want to show that $\sqrt{n}\partial\hat{Q}_n(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$ is asymptotically normal.

$$\sqrt{n}\frac{\partial\hat{Q}_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{S}_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \hat{\eta}_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta}_0) \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\boldsymbol{\theta}} \hat{S}_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \hat{\eta}_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta}_0),$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\boldsymbol{\theta}} \hat{S}_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \hat{\eta}_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta}_0) \xrightarrow{p} \mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\theta}} S_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})\right\}.$$

By regularity condition (B1), the order of differentiation and integration can be exchanged. Hence, we have

$$\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\theta}} S_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})\right\} = -\mathbb{E}\{S_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta})^{\otimes 2}\}.$$

By Slutsky's Theorem, we only need to show $n^{-1/2}\sum_{i=1}^{n} \hat{S}_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \hat{\eta}_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta}_0)$ converges to a normal distribution. Also, because

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \mathbb{E}\big\{S_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta_1, \eta_2, \eta_3; \boldsymbol{\theta}_0)^{\otimes 2}\big\}\right],$$

we only need to show

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{S}_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \hat{\eta}_{1,2,3}, \hat{\boldsymbol{\Delta}}_1, \hat{\boldsymbol{\Delta}}_2; \boldsymbol{\theta}_0) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S_{\text{eff}}(\mathbf{Y}_i, \mathbf{X}_i, \eta_{1,2,3}, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2; \boldsymbol{\theta}_0) \xrightarrow{p} 0.$$

Since $S^*_{\text{eff}} = (S^*_{\text{eff},\boldsymbol{\gamma}}, S^*_{\text{eff},\mathbf{b}})$, and the proof for each component are similar, we only prove the convergence in $S^*_{\text{eff},\mathbf{b}}$.

Let $\boldsymbol{\Gamma}$ and $\mathbf{B}$ denote the orthogonal basis derived from $\boldsymbol{\theta}_0$. Then,

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{S}_{\text{eff},\mathbf{b}}(\mathbf{Y}_i,\mathbf{X}_i,\hat{\eta}_{1,2,3},\hat{\boldsymbol{\Delta}}_1,\hat{\boldsymbol{\Delta}}_2;\boldsymbol{\theta}_0)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}S_{\text{eff},\mathbf{b}}(\mathbf{Y}_i,\mathbf{X}_i,\eta_{1,2,3},\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2;\boldsymbol{\theta}_0)\\
=&n^{-1/2}\sum_{i=1}^{n}\left[\{\hat{\boldsymbol{\Delta}}_{i2}(\boldsymbol{\theta}_0)-\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)\}\frac{\partial\log\eta_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right]\\
&+n^{-1/2}\sum_{i=1}^{n}\left[\{\hat{\boldsymbol{\Delta}}_{i2}(\boldsymbol{\theta}_0)-\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)\}\left\{\frac{\partial\log\hat{\eta}_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}-\frac{\partial\log\eta_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right\}\right]\\
&+n^{-1/2}\sum_{i=1}^{n}\left[\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)\left\{\frac{\partial\log\hat{\eta}_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}-\frac{\partial\log\eta_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right\}\right].
\end{aligned}
$$

By Lemma 5, the first term is $O_p(h^2+n^{1/2}h^4+n^{-1/2}h^{-(r+p-u-d)}\log n)$ and the third term is $O_p(h^2+n^{1/2}h^4+n^{-1/2}h^{-(r-u-d+1)}\log n)$. By the uniform convergence theorem in nonparametric regression (Theorem 2.6 in Li and Racine (2007) and Theorem 6 in Hansen (2008)), we have

$$
\sup_{\mathbf{X}_i,\mathbf{Y}_i}|\hat{\boldsymbol{\Delta}}_{i2}(\boldsymbol{\theta}_0)-\boldsymbol{\Delta}_{i2}(\boldsymbol{\theta}_0)|=O_p\left\{\left(\frac{\log n}{nh^{(r+p-u-d)}}\right)^{1/2}+h^2\right\},
$$

and

$$
\sup_{\mathbf{X}_i,\mathbf{Y}_i}\left|\frac{\partial\log\hat{\eta}_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}-\frac{\partial\log\eta_3}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^T\mathbf{Y}_i)^T}\right|=O_p\left\{\left(\frac{\log n}{nh^{r-u-d+2}}\right)^{1/2}+h^2\right\}.
$$

Hence, the second term can be bounded by

$O_p\{n^{-1/2}h^{r-u-d+1+p/2}\log n+n^{1/2}h^4\}=o_p(1)$. Therefore,

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{S}_{\text{eff},\mathbf{b}}(\mathbf{Y}_i,\mathbf{X}_i,\hat{\eta}_{1,2,3},\hat{\boldsymbol{\Delta}}_1,\hat{\boldsymbol{\Delta}}_2;\boldsymbol{\theta}_0)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}S_{\text{eff},\mathbf{b}}(\mathbf{Y}_i,\mathbf{X}_i,\eta_{1,2,3},\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2;\boldsymbol{\theta}_0)\xrightarrow{p}0.
$$

Hence, by Slutsky's Theorem,

$$
\sqrt{n}\frac{\partial\hat{Q}_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\xrightarrow{d}\mathcal{N}(0,\mathcal{V}_{\boldsymbol{\theta}}^{-3})
$$

where $\mathcal{V}_{\boldsymbol{\theta}}=\mathbb{E}\{S_{\text{eff}}(\mathbf{Y}_i,\mathbf{X}_i,\eta_{1,2,3},\boldsymbol{\Delta}_1,\boldsymbol{\Delta}_2;\boldsymbol{\theta})^{\otimes 2}\}^{-1}$.

Similarly, we can prove that

$$\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\hat{Q}_n(\boldsymbol{\theta}_0) \xrightarrow{p} \mathcal{V}_{\boldsymbol{\theta}}^{-2} = \mathbf{H}.$$

Therefore, by Lemma 7, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{V}_{\boldsymbol{\theta}}),$$

which achieves the semiparametric efficiency bound.

■

Lemma C.1.6 *(Theorem 2.1 in Newey and McFadden (1994)) Suppose there is a function $Q_0(\boldsymbol{\theta})$ such that (i) $Q_0(\boldsymbol{\theta})$ is uniquely minimized at $\boldsymbol{\theta}_0$, (ii) $\boldsymbol{\Theta}$ is compact, (iii) $Q_0(\boldsymbol{\theta})$ is continuous, (iv) $\hat{Q}_n(\boldsymbol{\theta})$ converges uniformly in probability to $Q_0(\boldsymbol{\theta})$, then $\hat{Q}_n(\boldsymbol{\theta})$, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.*

Lemma C.1.7 *(Theorem 3.1 in Newey and McFadden (1994)) Suppose that $\hat{\boldsymbol{\theta}}$ is a minimizer of $\hat{Q}_n(\boldsymbol{\theta})$, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ and (i) $\boldsymbol{\theta}_0 \in$ interior$(\boldsymbol{\Theta})$, (ii) $\hat{Q}_n(\boldsymbol{\theta})$ is twice continuously differentiable in a neighborhood $\mathcal{N}_\epsilon$ of $\boldsymbol{\theta}_0$, (iii) $\sqrt{n}\nabla_{\boldsymbol{\theta}}\hat{Q}_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, (iv) there is $H(\boldsymbol{\theta})$ that is continuous at $\boldsymbol{\theta}_0$ and $\sup_{\boldsymbol{\theta}\in\mathcal{N}_\epsilon} \|\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\hat{Q}_n(\boldsymbol{\theta}) - H(\boldsymbol{\theta})\| \xrightarrow{p} \mathbf{0}$, (v) $\mathbf{H} = H(\boldsymbol{\theta}_0)$ is nonsingular. Then, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{H}^{-1}\boldsymbol{\Sigma}\mathbf{H}^{-1})$.*

C.2 Additional Materials

C.2.1 MLE of the regression parameter under the inner envelope model

In this part, we present the MLE of the regression parameter $\boldsymbol{\beta}$ under the inner envelope model. Here, we assume the inner envelope spaces $\hat{\boldsymbol{\Gamma}}$, $\hat{\boldsymbol{\Gamma}}_0$, $\hat{\mathbf{B}}$ and $\hat{\mathbf{B}}_0$ are already calculated.

Let $\mathbf{S}_{\text{fit}}$ and $\mathbf{S}_{\text{res}}$ denote the sample covariance matrices of the fitted and residual vectors from the OLS fit of $\mathbf{Y}$ on $\mathbf{X}$. Let $\tilde{\lambda}_i(\mathbf{G}_0)$ denote the ordered, descending eigenvalues of $(\mathbf{G}_0^T\mathbf{S}_{\text{res}}\mathbf{G}_0)^{-1/2}(\mathbf{G}_0^T\mathbf{S}_{\text{fit}}\mathbf{G}_0)(\mathbf{G}_0^T\mathbf{S}_{\text{res}}\mathbf{G}_0)^{-1/2}$, where $\mathbf{G}_0 \in \mathbb{R}^{r\times(r-u)}$ is a semi-orthogonal matrix. Also, we denote the matrices of ordered eigenvectors and eigenvalues as $\tilde{\mathbf{V}}(\mathbf{G}_0)$ and $\tilde{\boldsymbol{\Lambda}}(\mathbf{G}_0) = \text{diag}\{\tilde{\lambda}_1(\mathbf{G}_0), \ldots, \tilde{\lambda}_{r-u}(\mathbf{G}_0)\}$, and let $\tilde{\mathbf{K}}(\mathbf{G}_0) = \text{diag}\{0, \ldots, 0, \tilde{\lambda}_{p-u+1}(\mathbf{G}_0), \ldots, \tilde{\lambda}_{r-u}(\mathbf{G}_0)\}$. Let $\mathbf{F}$ denote the $n \times p$ matrix with ith

row $\mathbf{X}_i^T$, let $\mathbf{U}$ be the $n \times r$ matrix with ith row $(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$, and let $\hat{\boldsymbol{\beta}}_{\text{ols}}$ denote the MLE of $\boldsymbol{\beta}$ under the standard model (4.1). Then,

$$
\begin{aligned}
&\hat{\boldsymbol{\zeta}}_1^T = \hat{\boldsymbol{\Gamma}}^T \hat{\boldsymbol{\beta}}_{\text{ols}}, \\
&\hat{\boldsymbol{\Omega}}_1 = (\mathbf{U}\hat{\boldsymbol{\Gamma}} - \mathbf{F}\hat{\boldsymbol{\zeta}}_1)^T(\mathbf{U}\hat{\boldsymbol{\Gamma}} - \mathbf{F}\hat{\boldsymbol{\zeta}}_1)/n, \\
&\hat{\boldsymbol{\Omega}}_0 = \hat{\boldsymbol{\Gamma}}_0 \mathbf{S}_{\text{res}} \hat{\boldsymbol{\Gamma}}_0 + (\hat{\boldsymbol{\Gamma}}_0 \mathbf{S}_{\text{res}} \hat{\boldsymbol{\Gamma}}_0)^{1/2} \tilde{\mathbf{V}}(\hat{\boldsymbol{\Gamma}}_0) \tilde{\mathbf{K}}(\hat{\boldsymbol{\Gamma}}_0) \tilde{\mathbf{V}}(\hat{\boldsymbol{\Gamma}}_0) (\hat{\boldsymbol{\Gamma}}_0 \mathbf{S}_{\text{res}} \hat{\boldsymbol{\Gamma}}_0)^{1/2} \\
&\text{span}(\hat{\mathbf{B}}) = \hat{\boldsymbol{\Omega}}_0 \mathcal{S}_{p-d}(\hat{\boldsymbol{\Omega}}_0, \hat{\boldsymbol{\Gamma}}_0^T) \\
&\hat{\boldsymbol{\zeta}}_2^T = (\hat{\mathbf{B}}^T \hat{\boldsymbol{\Omega}}_0^{-1} \hat{\mathbf{B}})^{-1} \hat{\mathbf{B}}^T \hat{\boldsymbol{\Omega}}_0^{-1} \hat{\boldsymbol{\Gamma}}_0^T \hat{\boldsymbol{\beta}}_{\text{ols}}, \\
&\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Gamma}} \hat{\boldsymbol{\zeta}}_1^T + \hat{\boldsymbol{\Gamma}}_0 \hat{\boldsymbol{\Omega}}_0 \hat{\boldsymbol{\zeta}}_2^T, \\
&\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Gamma}} \hat{\boldsymbol{\Omega}}_1 \hat{\boldsymbol{\Gamma}}^T + \hat{\boldsymbol{\Gamma}}_0 \hat{\boldsymbol{\Omega}}_0 \hat{\boldsymbol{\Gamma}}_0^T,
\end{aligned}
$$

where $\mathcal{S}_k(\mathbf{A}, \mathbf{B})$ is the span of $\mathbf{A}^{-1/2}$ times the first k eigenvectors of $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$. Detailed derivations are carried out by Su and Cook (2012).

C.2.2 Nonparametric density estimation and nonparametric regression

The derivative of the log densities are estimated by

$$
\frac{\partial \log \hat{\eta}_1}{\partial (\boldsymbol{\Gamma}^T \mathbf{y})^T} = \frac{\sum_{i=1}^n K_h'(\boldsymbol{\Gamma}^\intercal \mathbf{Y}_i - \boldsymbol{\Gamma}^T \mathbf{y}) K_h(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^n K_h(\boldsymbol{\Gamma}^\intercal \mathbf{Y}_i - \boldsymbol{\Gamma}^T \mathbf{y}) K_h(\mathbf{X}_i - \mathbf{x})}, \tag{C.1}
$$

$$
\frac{\partial \log \hat{\eta}_2}{\partial (\mathbf{B}^T \boldsymbol{\Gamma}_0^T \mathbf{y})^T} = \frac{\sum_{i=1}^n K_h'\{\mathbf{B}^T \boldsymbol{\Gamma}_0^\intercal (\mathbf{Y}_i - \mathbf{y})\} K_h\{\mathbf{B}_0^T \boldsymbol{\Gamma}_0^\intercal (\mathbf{Y}_i - \mathbf{y})\} K_h(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^n K_h\{\mathbf{B}^T \boldsymbol{\Gamma}_0^\intercal (\mathbf{Y}_i - \mathbf{y})\} K_h\{\mathbf{B}_0^T \boldsymbol{\Gamma}_0^\intercal (\mathbf{Y}_i - \mathbf{y})\} K_h(\mathbf{X}_i - \mathbf{x})}, \tag{C.2}
$$

$$
\frac{\partial \log \hat{\eta}_3}{\partial (\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T \mathbf{y})^T} = \frac{\sum_{i=1}^n K_h'\{\mathbf{B}_0^T \boldsymbol{\Gamma}_0^\intercal (\mathbf{Y}_i - \mathbf{y})\}}{\sum_{i=1}^n K_h\{\mathbf{B}_0^T \boldsymbol{\Gamma}_0^\intercal (\mathbf{Y}_i - \mathbf{y})\}}. \tag{C.3}
$$

The nonparametric regression $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ are estimated by

$$
\begin{aligned}
\hat{\boldsymbol{\Delta}}_1 =& \mathbf{y} - \frac{\sum_{i=1}^N \mathbf{Y}_i K_h(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^N K_h(\mathbf{X}_i - \mathbf{x})}, \\
\hat{\boldsymbol{\Delta}}_2 =& \frac{\sum_{i=1}^N \mathbf{P}_{\boldsymbol{\Gamma}_0 \mathbf{B}} \mathbf{Y}_i K_h\{\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T (\mathbf{Y}_i - \mathbf{y})\} K_h(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^N K_h\{\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T (\mathbf{Y}_i - \mathbf{y})\} K_h(\mathbf{X}_i - \mathbf{x})} \\
& - \frac{\sum_{i=1}^N \mathbf{P}_{\boldsymbol{\Gamma}_0 \mathbf{B}} \mathbf{Y}_i K_h\{\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T (\mathbf{Y}_i - \mathbf{y})\}}{\sum_{i=1}^N K_h\{\mathbf{B}_0^T \boldsymbol{\Gamma}_0^T (\mathbf{Y}_i - \mathbf{y})\}}.
\end{aligned} \tag{C.4}
$$

The nonparametric regression for the conditional expectations of η_i^* are estimated by

$$\hat{\mathbb{E}}\left(\frac{\partial \log \eta_1^*}{\partial(\boldsymbol{\Gamma}^\intercal \mathbf{Y}_i)^T} \,\middle|\, \mathbf{x}\right) = \frac{\sum_{i=1}^N \partial \log \eta_1^* / \partial(\boldsymbol{\Gamma}^\intercal \mathbf{Y}_i)^T K_h(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^N K_h(\mathbf{X}_i - \mathbf{x})}, \tag{C.5}$$

$$\hat{\mathbb{E}}\left(\frac{\partial \log \eta_2^*}{\partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^\intercal \mathbf{Y}_i)^T} \,\middle|\, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^\intercal \mathbf{y}, \mathbf{x}\right) = \frac{\sum_{i=1}^N \partial \log \eta_2^* / \partial(\mathbf{B}^T\boldsymbol{\Gamma}_0^\intercal \mathbf{Y}_i)^T K_h\{\mathbf{B}_0^T\boldsymbol{\Gamma}_0^\intercal(\mathbf{y}_i - \mathbf{y})\} K_h(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^N K_h\{\mathbf{B}_0^T\boldsymbol{\Gamma}_0^\intercal(\mathbf{y}_i - \mathbf{y})\} K_h(\mathbf{X}_i - \mathbf{x})}, \tag{C.6}$$

$$\hat{\mathbb{E}}\left(\frac{\partial \log \eta_2^*}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^\intercal \mathbf{Y}_i)^T} \,\middle|\, \mathbf{B}_0^T\boldsymbol{\Gamma}_0^\intercal \mathbf{y}, \mathbf{x}\right) = \frac{\sum_{i=1}^N \partial \log \eta_2^* / \partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^\intercal \mathbf{Y}_i)^T K_h\{\mathbf{B}_0^T\boldsymbol{\Gamma}_0^\intercal(\mathbf{y}_i - \mathbf{y})\} K_h(\mathbf{X}_i - \mathbf{x})}{\sum_{i=1}^N K_h\{\mathbf{B}_0^T\boldsymbol{\Gamma}_0^\intercal(\mathbf{y}_i - \mathbf{y})\} K_h(\mathbf{X}_i - \mathbf{x})}, \tag{C.7}$$

$$\hat{\mathbb{E}}\left(\frac{\partial \log \eta_3^*}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^\intercal \mathbf{Y}_i)^T}\right) = \frac{1}{n}\sum_{i=1}^N \frac{\partial \log \eta_3^*}{\partial(\mathbf{B}_0^T\boldsymbol{\Gamma}_0^\intercal \mathbf{Y}_i)^T}. \tag{C.8}$$

C.2.3 Simulation: linear model with additive, normal errors

For the out-of-sample predictive RMSE results in Section 6.1, the oracle, InnEnv, local, global, GMM, OLS, Env and PLS methods have prediction RMSE equal to 1.82, 1.82, 1.83, 1.83, 1.83, 1.91, 1.97, 1.96 respectively.

C.2.4 Real data: synthetic dataset from iris data

The estimated inner envelope spaces obtained from the globally efficient algorithm is

$$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} -0.01 \\ -0.02 \\ 0.75 \\ -0.46 \\ 0.35 \\ 0.33 \end{pmatrix}, \quad \hat{\boldsymbol{\Gamma}}_0\hat{\mathbf{B}} = \begin{pmatrix} 0 \\ 0 \\ -0.50 \\ 0.08 \\ 0.66 \\ 0.55 \end{pmatrix}, \quad \hat{\boldsymbol{\Gamma}}_0\hat{\mathbf{B}}_0 = \begin{pmatrix} -0.98 & 0.08 & -0.14 & 0 \\ 0 & -0.86 & -0.50 & 0.01 \\ 0 & 0 & -0.04 & -0.43 \\ 0 & 0 & 0 & -0.88 \\ 0.11 & 0.32 & -0.56 & -0.12 \\ -0.13 & -0.38 & 0.64 & -0.12 \end{pmatrix}.$$

Because $\mathcal{S}_0$ and $\mathcal{S}_3$ are of different dimension, we consider the spaces $\mathbf{P}_{\mathcal{S}_3}\mathcal{S}_0$ and $\mathcal{S}_3^\perp$. The space $\mathbf{P}_{\mathcal{S}_3}\mathcal{S}_0$ should be close to $\mathcal{S}_0$, and $\mathcal{S}_3^\perp$ should be perpendicular to $\mathcal{S}_0$. It turns out that $\text{dist}(\mathbf{P}_{\mathcal{S}_3}\mathcal{S}_0, \mathcal{S}_0) = 0.029$ and $\text{dist}(\mathcal{S}_3^\perp, \mathcal{S}_0) = 1.999$. Since the upper bound for $\text{dist}(\mathcal{S}_3^\perp, \mathcal{S}_0)$ is 2, it indicates that $\mathcal{S}_0$ is contained in $\mathcal{S}_3$. That is, our method successfully identified the artificial noise space $\mathcal{S}_0$.

C.3 Tables and Figures

n		InnEnv	Local	Global	GMM
100	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	0.111	0.134	0.166	0.224
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	0.089	0.121	0.152	0.201
300	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	0.085	0.101	0.125	0.161
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	0.062	0.078	0.102	0.119
500	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	0.070	0.089	0.114	0.141
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	0.050	0.064	0.087	0.101
750	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	0.055	0.071	0.095	0.112
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	0.041	0.057	0.072	0.084
1000	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	0.045	0.060	0.084	0.099
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	0.035	0.044	0.065	0.073

Table C.1: Mean distance of the estimated space and the true space (linear case)

Size	Oracle	InnEnv	Local	Global	GMM	OLS	Env	PLS
100	0.136	0.162	0.190	0.212	0.244	0.669	0.705	0.751
300	0.077	0.088	0.097	0.115	0.135	0.378	0.403	0.460
500	0.060	0.065	0.071	0.084	0.102	0.287	0.386	0.317
750	0.049	0.054	0.063	0.074	0.085	0.240	0.346	0.278
1000	0.042	0.049	0.056	0.068	0.077	0.213	0.310	0.236

Table C.2: Mean and standard errors of $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F^2$ in Scenario 1 with different sample sizes.

Size		InnEnv	Local	Global	GMM
100	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	1.175	0.542	0.360	0.745
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	1.028	0.382	0.274	0.507
300	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	1.192	0.429	0.277	0.589
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	1.052	0.259	0.186	0.344
500	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	1.213	0.360	0.236	0.502
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	1.068	0.209	0.142	0.277
750	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	1.201	0.297	0.192	0.412
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	1.089	0.177	0.117	0.234
1000	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	1.189	0.265	0.168	0.368
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	1.090	0.154	0.101	0.203

Table C.3: Mean distance of the estimated space and the true space (nonlinear case)

Size		InnEnv	Local	Global	GMM
100	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	0.992	0.385	0.320	0.425
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	0.998	0.498	0.401	0.552
300	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	0.984	0.225	0.192	0.268
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	0.920	0.288	0.238	0.340
500	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	0.917	0.175	0.147	0.208
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	0.903	0.223	0.182	0.263
750	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	0.980	0.140	0.117	0.169
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	0.913	0.182	0.150	0.215
1000	$dist(\mathcal{S}_1, \hat{\mathcal{S}}_1)$	0.932	0.132	0.105	0.147
	$dist(\mathcal{S}_3, \hat{\mathcal{S}}_3)$	0.919	0.157	0.128	0.186

Table C.4: Mean distance of the estimated space and the true space (exponential case)

Table C.5: The point estimates, bootstrap standard errors and p-values for the regression parameter for the *iris* dataset

	Our Method			Standard		
Corresponding to X_1	$\hat{\boldsymbol{\beta}}$	$\hat{\text{SE}}$	p-value	$\hat{\boldsymbol{\beta}}$	$\hat{\text{SE}}$	p-value
Noise_1	0.08	0.10	0.40	0.07	0.12	0.54
Noise_2	0.05	0.13	0.68	0.07	0.14	0.60
Sepal length	-1.01	0.06	¡0.01	-1.01	0.06	¡0.01
Sepal width	0.85	0.12	¡0.01	0.85	0.12	¡0.01
Pedal length	-1.30	0.02	¡0.01	-1.30	0.02	¡0.01
Pedal width	-1.25	0.02	¡0.01	-1.25	0.02	¡0.01
Corresponding to X_2	$\hat{\boldsymbol{\beta}}$	$\hat{\text{SE}}$	p-value	$\hat{\boldsymbol{\beta}}$	$\hat{\text{SE}}$	p-value
Noise_1	0.03	0.03	0.27	0.07	0.13	0.59
Noise_2	-0.01	0.03	0.73	-0.02	0.16	0.91
Sepal length	0.10	0.07	0.12	0.14	0.10	0.17
Sepal width	-0.65	0.10	¡0.01	-0.65	0.10	¡0.01
Pedal length	0.28	0.04	¡0.01	0.29	0.004	¡0.01
Pedal width	0.22	0.04	¡0.01	0.17	0.04	¡0.01

p-values	setosa:versicolor	setosa:virginica	versicolor:virginica
$\mathcal{S}_{31}$	0.396	0.717	0.272
$\mathcal{S}_{32}$	0.869	0.717	0.549
$\mathcal{S}_{33}$	0.869	0.396	0.869
$\mathcal{S}_{34}$	0.869	0.998	0.717

Table C.6: p-values for testing whether the distributions of each species is different on $\mathcal{S}_3$.

Appendix D

Appendix for Chapter 5

D.1 Extension to Invalid IVs

The exclusion restriction assumption, i.e., Assumption 5.1.3, requires a valid IV not to have a direct effect on the outcome. This can be violated. For example, in genetic studies, the pleiotropic effect happens when a single gene have multiple effects. Lewbel (2012)'s estimation strategies exploit IVs with a heteroskedastic covariance restriction in endogenous regressor models. Lewbel's approach is extended by Tchetgen Tchetgen et al. (2017) to MR G-Estimation under No Interaction with Unmeasured Selection" (MR GENIUS). The extension of our method to invalid IVs can be built on the identification and estimation results of Lewbel (2012) and Tchetgen Tchetgen et al. (2017). We adapt notations introduced by Tchetgen Tchetgen et al. (2017).

Consider the following model:

$$\begin{aligned} E(X_i|Z_i, U_i) &= \alpha_Z(Z_i, U_i) + \alpha_U(U_i), \\ E(Y_i|X_i, Z_i, U_i) &= \beta X_i + \beta_Z(Z_i, U_i) + \beta_U(U_i), \end{aligned} \tag{D.1}$$

where $E\{\beta_U(U_i)\} = 0$, $E\{\alpha_U(U_i)\} = 0$, $\alpha_Z(0, U_i) = 0$ and $\beta_Z(0, U_i) = 0$. We further assume that $E(X_i|Z_i) = E\{\alpha_Z(Z_i, U_i)|Z_i\} = \delta Z_i$. Comparing Model (D.1) with Model (5.1), when $\alpha_Z(Z, U) = \delta Z$, $\alpha_U(U) = \varepsilon_X$ and $\beta_U(U) = \varepsilon_Y$, the second equation in Model (D.1) has an additional term $\beta_Z(Z_i, U_i)$ which allows a direct impact of IV on the outcome, with potential effect modification by unmeasured confounders U_i. Tchetgen Tchetgen et al. (2017) showed that the causal effect β can be identified under the following two assumptions.

Assumption D.1.1 (orthogonal conditions) *The following equations hold with probability 1,*

$$cov\{\beta_Z(Z_i, U_i), \alpha_Z(Z_i, U_i)|Z_i\} = cov\{\beta_Z(Z_i, U_i), \alpha_U(U_i)|Z_i\} = 0,$$
$$cov\{\alpha_Z(Z_i, U_i), \beta_U(U_i)|Z_i\} = 0.$$

As illustrated in Tchetgen Tchetgen et al. (2017), Assumption D.1.1 does not imply orthogonality of $\beta_U(U_i)$ and $\alpha_U(U_i)$ and therefore the degree of unmeasured confounding is not restricted by these orthogonality conditions. However, Assumption D.1.1 restricts the degree of common effect modifiers in the outcome and exposure models. As a special case, Assumption D.1.1 is satisfied if $\beta_Z(Z_i, U_i) = \beta_Z(Z_i)$ and $\alpha_Z(Z_i, U_i) = \alpha_Z(Z_i)$. More examples of models that hold under Assumption D.1.1 can be found in Tchetgen Tchetgen et al. (2017). However, Assumption D.1.1 alone is not enough for identification, we also need to make the following assumption.

Assumption D.1.2 (heteroskedastic condition) $rank[cov\{Z_i, var(X_i|Z_i)\}] = p$, *which is equivalent to* $cov\{Z_i, var(X_i|Z_i)\}$ *is full column rank.*

Assumption D.1.2 requires the variance of X_i to depend on Z_i. When Z_i is binary, Assumption D.1.2 is equivalent to $var(X_i|Z_i = 1) \neq var(X_i|Z_i = 0)$. As commented in Tchetgen Tchetgen et al. (2017), this assumption is empirically testable, and will typically hold for binary X_i, other than at some exceptional laws. The key to achieve identification of β is to do a transformation of the responses and predictors so that the direct effect term $\beta_Z(Z_i, U_i)$ is canceled. Specifically, multiply $X_i - E(X_i|Z_i)$ on both sides of Model (D.1), we have $E(Y_{new,i}|X_{new,i}, Z_i, U_i) = \Psi X_{new,i} + \varepsilon_{new,i}$, where $Y_{new,i} = \text{vec}[\{X_i - E(X_i|Z_i)\}Y_i^T]$ and $X_{new,i} = \text{vech}[\{X_i - E(X_i|Z_i)\}X_i^T]$, $\varepsilon_{new,i} = \text{vec}[E\{\{X_i - E(X_i|Z_i)\}\beta_Z(Z_i, U_i)^T + \{X_i - E(X_i|Z_i)\}\beta_U(U_i)^T|X_{new,i}, Z_i, U_i\}]$ and $\Psi = (\beta \otimes I_p)E_p$, where vector-half operator vech: $\mathbb{S}^{a\times a} \to \mathbb{R}^{a(a+1)/2}$ stacks only the unique part of each column lines on or above the diagonal of a symmetric matrix into a vector, and the expansion matrix $E_a \in \mathbb{R}^{a^2\times a(a+1)/2}$ satisfies $\text{vec}(A) = E_r\text{vech}(A)$ (see Henderson and Searle (1979) for further background). The identification of β is based on the following representation.

Lemma D.1.1 *Under Model D.1, assume that Assumptions 5.1.1,5.1.2 and D.1.1 hold,*

$$E(Z_i Y_{new,i}^T) = E(Z_i {X_{new,i}}^T)\Psi^T.$$

Because $E(X|Z)$ can be identified, Ψ and β can also be identified. Under the condition of Lemma D.1.1, if we further make Assumption D.1.2, then $E(ZX_{new,i}^T)^T M E(ZX_{new,i}^T)$ is nonsingular for any positive definite matrix $M \in \mathbb{R}^{k\times k}$. Thus, Ψ can be solved as

$$\Psi = E(Z_i Y_{new,i}^T)^T \Sigma_Z^{-1} E(ZX_{new,i}^T)\{E(ZX_{new,i}^T)^T \Sigma_Z^{-1} E(ZX_{new,i}^T)\}^{-1},$$

where $M = \Sigma_Z^{-1}$. To construct an estimator of β based on Lemma D.1.1, we first estimate X_{new} and Y_{new} by estimating $E(X|Z)$ by OLS, which we denoted as $\widehat{X}_{new}$ and $\widehat{Y}_{new}$. Then, $\widehat{\beta}$ can be obtained by

$$\widehat{\beta}_{2SLS}^* = \widehat{\Psi}\text{vech}(I_p) = \hat{Y}_{new}^T P_Z \widehat{X}_{new}(\widehat{X}_{new}^T P_Z \widehat{X}_{new})^{-1}\text{vech}(I_p), \tag{D.2}$$

where $\widehat{Y}_{new}^T = (\widehat{Y}_{new,1}, \ldots, \widehat{Y}_{new,n})$, $\widehat{X}_{new}^T = (\widehat{X}_{new,1}, \ldots, \widehat{X}_{new,n})$, $\widehat{Y}_{new,i} = \text{vec}[\{X_i - \widehat{\delta}Z_i\}Y_i^T]$, $\widehat{X}_{new,i} = \text{vech}[\{X_i - \widehat{\delta}Z_i\}X_i^T]$ and $\widehat{\delta} = X^T Z(Z^T Z)^{-1}$.
When $r = p = 1$, the estimator $\widehat{\beta}_{2SLS}^*$ given in (D.2) reduces to the estimator proposed by Lewbel (2012), which is similar to 2SLS, except that X and Y are replaced by $\{X - \widehat{E}(X|Z)\}X$ and $\{X - \widehat{E}(X|Z)\}Y$. We denote the estimator in (D.2) as $\widehat{\beta}_{2SLS}^*$ because it is similar to the standard 2SLS estimator but is calculated using the transformed data.

Motivated by the fact that $\widehat{\Psi}$ is estimated similar to 2SLS, we can impose the first and second stage envelope conditions similarly as in Section 5.2.1 and 5.2.2. In this article, we assume that $X_{new,i} = \text{vech}[\{X_i - E(X_i|Z_i)\}X_i^T] = \delta Z_i + \varepsilon_{X_{new,i}}$, then λ can be defined as $\lambda = \beta\delta$. We apply the envelope on the model with outcome $Y_{new,i}$, exposure $X_{new,i}$ and IV, Z_i, similar as in Section 5.2.1 and 5.2.2. For instance, let $Z_1, Z_3, Z_2/\sqrt{Z_1 Z_3}$ follow uniform distribution $U(0,1)$. Suppose X follows a normal distribution $N(0, ((Z_1, Z_2)^T, (Z_2, Z_3)^T))$ plus a noise ε_X, which is independent with Z, and $Y = \beta X + \alpha Z + \varepsilon_Y$. In such model, we have the linear models $X_{new} = Z_1 + Z_2 + Z_3 + Var(\varepsilon_X)$ and $Y_{new} = \Psi X_{new} + \varepsilon_{new}$. If there is no linear relationship between X_{new} and IV, we can adopt the semiparametric envelope proposed in Chapter 4 (see more discussion in Section 5.5).

We have the similar asymptotic results using first stage and second envelope method respectively.

Proposition D.1.1 *Under Model* (D.1)*, and suppose Assumptions 5.1.1, 5.1.2, D.1.1 and D.1.2 hold. Suppose X_{new} and Z satisfy the conditions in Proposition 5.2.2. Then $\sqrt{n}\{vec(\widehat{\beta}_{1st,env}^*) - vec(\beta)\}$ converges to a normal distribution with mean 0 and variance $V_{1st,env}^* = \{vech(I_p)^T(\delta\Sigma_Z\delta^T)^{-1}vech(I_p)\}\Sigma_{\varepsilon_{new}} = V_{2SLS}^*$, where*

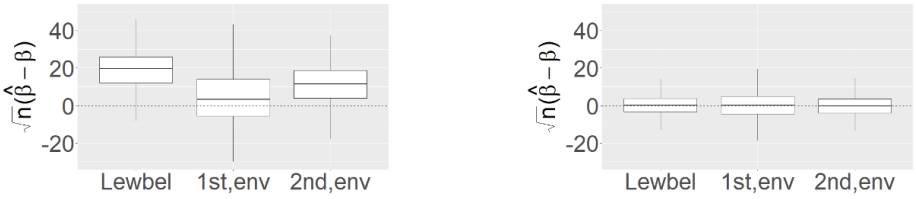
$\Sigma_{\varepsilon_{new}} = cov(\varepsilon_{new})$.

Proposition D.1.2 *Under Model* (D.1)*, and suppose Assumptions 5.1.1, 5.1.2, D.1.1 and D.1.2 hold. Suppose Z, X_{new} and Y_{new} satisfy all the conditions in Proposition 5.2.4. Suppose $\Sigma_Z = \Gamma\Delta\Gamma + \Gamma_0\Delta_0\Gamma_0$. Then $\sqrt{n}(\widehat{\beta}^*_{2nd,env} - \beta)$ converges in distribution to a normal distribution with mean 0 and variance $V^*_{2nd,env}$, where $V^*_{2nd,env} = \{vech(I_p)^T(\delta\Sigma_Z\delta^T)^{-1}\delta\Gamma\Delta\Gamma^T\delta^T(\delta\Sigma_Z\delta^T)^{-1}vech(I_p)\}\Sigma_{\varepsilon_{new}} + \{vech(I_p)^T(\delta\Sigma_Z\delta^T)^{-1}\delta\Gamma_0\Delta_0\Gamma_0^T\delta^T(\delta\Sigma_Z\delta^T)^{-1}vech(I_p)\}\beta Cov(X_{new}|Z)\beta^T$.*

Corollary D.1.1 *Under the conditions in Proposition D.1.2,* $V^*_{2SLS} - V^*_{2nd,env} = \{vech(I_p)^T(\delta\Sigma_Z\delta^T)^{-1}\delta\Gamma_0\Delta_0\Gamma_0^T\delta^T(\delta\Sigma_Z\delta^T)^{-1}vech(I_p)\}(\Sigma_{\varepsilon_{new}} - \beta Cov(X_{new}|Z)\beta^T)$.

D.2 Simulations for Invalid IVs

We investigate the MSE of second stage envelope under Model (D.1) when Assumption 5.1.3 is violated. We generate a sample of size $n = 5000$ and set $r = 1, p = 2$ and $k = 20$. We first generate $Z_i \sim c_i\chi_1, i = 1, 3, 4, \ldots, 20$, where $c_i = 1$ for $i = 1, 3$ and $c_i = 0.3$ for $i \neq 1, 3$. Let $Z_2 \sim U(0, \sqrt{Z_1 Z_3})$. Then, we generate confounder $\varepsilon_X = (\varepsilon_1, \varepsilon_2)^T \sim N(0, diag(10, 1))$, $\varepsilon_Y = \varepsilon_1$ and $X^* \sim N(0, \Sigma_{X^*})$, where $\Sigma_{X^*} = \{(Z_1, Z_2)^T, (Z_2, Z_3)^T\}$. We set $X = X^* + \varepsilon_X$ and $Y = \beta X + \alpha Z + \varepsilon_Y$, where $\beta = (-0.5, 0.5)$, $\alpha^T = (\alpha_1, \ldots, \alpha_{15})^T \in \mathbb{R}^{15}$ and α_j are independently generated from $N(0, 100)$ and held fixed. Thus, $E(X_{new}|Z) = (I_3, 0_{3\times 17})Z$. We apply 2SLS and the second stage envelope method as proposed in Section D.1. Figure D.1 shows the boxplot of $\sqrt{n}(\widehat{\beta}^*_{2SLS} - \beta)$, $\sqrt{n}(\widehat{\beta}^*_{1st,env} - \beta)$ and $\sqrt{n}(\widehat{\beta}^*_{2nd,env} - \beta)$ over 500 simulations. The eigenvalues of variance differences $var(\widehat{\beta}^*_{2SLS}) - var(\widehat{\beta}^*_{1st,env})$ and $var(\widehat{\beta}^*_{2SLS}) - var(\widehat{\beta}^*_{2nd,env})$ are $(-1.28, -10.13)$ and $(2.86, 0.81)$. The estimate bias $(\sqrt{n}\|E(\widehat{\beta} - \beta)\|_2)$ of 2SLS, first stage envelope and second stage envelope are 0.38, 0.39, 0.41 and 0.19. The MSEs $(nE[\|\widehat{\beta} - \beta\|_2^2])$ for these three methods are 87.23, 110.73, 110.88 and 21.91. This shows appealing performance of the second stage envelope estimator even when the IVs are invalid.



(a) First coordinate

(b) Second coordinate

Figure D.1: Box-plot of two coordinates of the 2SLS and the second stage envelope estimators when IVs are invalid

D.3 Algorithms of 2SLS and ILS

Algorithm 6: Two Stage Least Squares

Stage 1: Regress X on Z. Then, we have $\widehat{X} = Z(Z^TZ)^{-1}Z^TX$.

Stage 2: Regress Y on $\widehat{X}$, where $\widehat{X}$ is the predicted exposure from the first stage.

Calculation: $\widehat{\beta}_{2SLS} = Y^T\widehat{X}(\widehat{X}^T\widehat{X})^{-1} = Y^TP_ZX(X^TP_ZX)^{-1}$.

Algorithm 7: Indirect Least Squares

Stage 1: The same as traditional 2SLS

Stage 2: Regress Y on Z. Then, we have $\widehat{\delta} = (Z^TZ)^{-1}Z^TY$.

Calculation: $\widehat{\beta}_{ILS} = \widehat{\lambda}Z^TZ\widehat{\delta}^T(\widehat{\delta}Z^TZ\widehat{\delta}^T)^{-1} = Y^TP_ZX\left(X^TP_ZX\right)^{-1}$.

D.4 Proof of Proposition 5.2.1

Note that $\sqrt{n}(\widehat{\beta}_{2SLS} - \beta) = (\varepsilon_Y^TZ/\sqrt{n})\widehat{\delta}_{2SLS}^T\left(\widehat{\delta}_{2SLS}\widehat{\Sigma}_Z\widehat{\delta}_{2SLS}^T\right)^{-1}$, where $\widehat{\delta}_{2SLS} = X^TZ(Z^TZ)^{-1}$. Because $(X_i^T, Z_i^T)^T$, $i = 1, \ldots, n$, are i.i.d. with finite second moment, $\widehat{\delta}_{2SLS}$ converges to δ in probability.

Then $\sqrt{n}\{\text{vec}(\widehat{\beta}_{2SLS}) - \text{vec}(\beta)\} = \{\left(\widehat{\delta}_{2SLS}\widehat{\Sigma}_Z\widehat{\delta}_{2SLS}^T\right)^{-1}\widehat{\delta}_{2SLS} \otimes I_r\}\text{vec}(\varepsilon_Y^TZ/\sqrt{n}) = \{\left(\widehat{\delta}_{2SLS}\widehat{\Sigma}_Z\widehat{\delta}_{2SLS}^T\right)^{-1}\widehat{\delta}_{2SLS} \otimes I_r\}\sum_{i=1}^n Z_i \otimes \varepsilon_{Y_i}/\sqrt{n}$. Because Z_i and ε_{Y_i} have finite fourth moments, $Z_i \otimes \varepsilon_{Y_i}$ has finite variance $\Sigma_Z \otimes \Sigma$. Thus, by Central Limit Theorem, $\text{vec}(\sum_{i=1}^n Z_i \otimes \varepsilon_{Y_i}/\sqrt{n})$ converges to a normal distribution with mean 0 and variance $\Sigma_Z \otimes \Sigma$. By Slutsky's theorem, we have $\sqrt{n}\{\text{vec}(\widehat{\beta}_{2SLS}) - \text{vec}(\beta)\}$ converges to a normal distribution with mean 0 and variance $\{(\delta\Sigma_Z\delta^T)^{-1}\delta \otimes I_r\}(\Sigma_Z \otimes \Sigma_{\varepsilon_Y})\{\delta^T(\delta\Sigma_Z\delta^T)^{-1} \otimes I_r\} = (\delta\Sigma_Z\delta^T)^{-1} \otimes \Sigma_{\varepsilon_Y} = V$.

D.5 Proof of Proposition 5.2.2

Note that $\sqrt{n}(\widehat{\beta}_{1st,env}-\beta)=(\varepsilon_Y^TZ/\sqrt{n})\widehat{\delta}_{1st,env}^T\left(\widehat{\delta}_{1st,env}\widehat{\Sigma}_Z\widehat{\delta}_{1st,env}^T\right)^{-1}$ and $\widehat{\delta}$ converges to δ. We adapt similar approach as the proof of Proposition 5.2.1.

We have $\sqrt{n}\{\text{vec}(\widehat{\beta}_{1st,env})-\text{vec}(\beta)\}=\{\left(\widehat{\delta}_{1st,env}\widehat{\Sigma}_Z\widehat{\delta}_{1st,env}^T\right)^{-1}\widehat{\delta}_{1st,env}\otimes I_r\}\text{vec}(\varepsilon_Y^TZ/\sqrt{n})=\{\left(\widehat{\delta}_{1st,env}\widehat{\Sigma}_Z\widehat{\delta}_{1st,env}^T\right)^{-1}\widehat{\delta}_{1st,env}\otimes I_r\}\sum_{i=1}^n Z_i\otimes\varepsilon_{Y_i}/\sqrt{n}$. Because Z_i and ε_{Y_i} have finite fourth moments, $Z_i\otimes\varepsilon_{Y_i}$ has finite variance $\Sigma_Z\otimes\Sigma$. Thus, by Central Limit Theorem, $\text{vec}(\sum_{i=1}^n Z_i\otimes\varepsilon_{Y_i}/\sqrt{n})$ converges to a normal distribution with mean 0 and variance $\Sigma_Z\otimes\Sigma$. By Slutsky's theorem, we have $\sqrt{n}\{\text{vec}(\widehat{\beta}_{1st,env})-\text{vec}(\beta)\}$ converges to a normal distribution with mean 0 and variance $\{(\delta\Sigma_Z\delta^T)^{-1}\delta\otimes I_r\}(\Sigma_Z\otimes\Sigma_{\varepsilon_Y})\{\delta^T(\delta\Sigma_Z\delta^T)^{-1}\otimes I_r\}=(\delta\Sigma_Z\delta^T)^{-1}\otimes\Sigma_{\varepsilon_Y}=V$.

D.6 Proof of Lemma 5.2.1

This proposition is a direct generalization of Buse (1992) when Y is multivariate and predictors are projected on the given envelope. The proof is similar.

Proof. Note that

$$\widehat{\beta}_{1st,env}-\beta=\frac{\varepsilon_Y^TP_{Z\Phi}X}{n}(\frac{X^TP_{Z\Phi}X}{n})^{-1}=\frac{\varepsilon_Y^TP_{Z\Phi}X}{n}(\frac{\delta Z^TZ\delta^T}{n}+\frac{\delta Z\varepsilon_X}{n}+\frac{\varepsilon_X^TZ\delta^T}{n}+\frac{\varepsilon_X^TP_{Z\Phi}\varepsilon_X}{n})^{-1} \tag{D.3}$$

$$:=\frac{\varepsilon_Y^TP_{Z\Phi}X}{n}(A+B+C+D)^{-1}. \tag{D.4}$$

The expectation considered here is conditional expectation giving $Z\Phi$. By multivariate Central limit theorem (CLT), we know that $\boldsymbol{\varepsilon}^TZ\delta^T/n=o_p(n^{-1/2+\epsilon})$, for any fixed $\epsilon\in(0,1/8)$. Thus, $A\to\delta\Sigma_Z\delta^T$ in probability, $B=o_p(n^{-1/2+\epsilon})$, $C=o_p(n^{-1/2+\epsilon})$ and $D=\frac{\varepsilon_X^TZ\Phi}{n}(\frac{\Phi^TZ^TZ\Phi}{n})^{-1}\frac{Z\Phi^T\varepsilon_X}{n}=o_p(n^{-1+2\epsilon})$. Suppose $S=A+B+C+D$ and $r=S^{-1}-\{A^{-1}-A^{-1}(S-A)A^{-1}\}$. We aim to prove r is $o_p(n^{-1+2\epsilon})$. Because $rS=(S-A)A^{-1}(S-A)A^{-1}=o_p(n^{-1+2\epsilon})$ and $S\to\delta\Sigma_Z\delta^T$ in probability, we have $r=o_p(n^{-1+2\epsilon})$ and $S^{-1}=A^{-1}-A^{-1}(S-A)A^{-1}+o_p(n^{-1+2\epsilon})=A^{-1}-A^{-1}(B+C)A^{-1}+o_p(n^{-1+2\epsilon})$.

Now consider $\frac{\varepsilon_Y^TP_{Z\Phi}X}{n}=\frac{\varepsilon_Y^TZ\delta^T}{n}+\frac{\varepsilon_Y^TZ\Phi(\Phi^TZ^TZ\Phi)^{-1}Z\Phi^T\varepsilon_X}{n}:=E+F$. Similarly, by multivariate CLT, $E=o_p(n^{-1/2+\epsilon})$ and $F=o_p(n^{-1+2\epsilon})$. Then $\widehat{\beta}_{2SLS}-\beta=(E+F)\{A^{-1}-A^{-1}(B+C)A^{-1}+o_p(n^{-1+2\epsilon})\}=EA^{-1}-EA^{-1}(B+C)A^{-1}+FA^{-1}+o_p(n^{-1+2\epsilon})$.

Then we calculate the expectation of the four leading terms.

First, $E\{EA^{-1}\} = E\{\varepsilon_Y^T Z\delta(\delta Z^T Z\delta^T)^{-1}\} = 0$ because of giving $Z\Phi$. Then,

$$E(FA^{-1} - EA^{-1}BA^{-1}) = E\{\varepsilon_Y^T Z\Phi(\Phi^T Z^T Z\Phi)^{-1}Z\Phi^T\varepsilon_X A^{-1} - \varepsilon_Y^T Z\delta^T A^{-1}\delta Z\Phi^T\varepsilon_X A^{-1}\} \tag{D.5}$$

$$= E[\varepsilon_Y^T\{Z\Phi(\Phi^T Z^T Z\Phi)^{-1}Z\Phi^T - Z\delta^T A^{-1}\delta Z\Phi^T\}\varepsilon_X A^{-1}]. \tag{D.6}$$

Note that $P_1 = Z\Phi(\Phi^T Z^T Z\Phi)^{-1}Z\Phi^T - Z\delta^T A^{-1}\delta Z\Phi^T$ is an idempotent symmetric matrix with $tr(P_1) = u - p$. Then $E(\varepsilon_Y^T P_1\varepsilon_X A^{-1}) = \alpha_1^T E(U^T P_1 U)\alpha_2 A^{-1}$. The (i,j) element of $E(U^T P_1 U)$ is $E(U_i^T P_1 U_j) = tr(P_1)\sigma_{i,j} = (u-p)\sigma_{i,j}$, where U_l represents the lth column of U and $\sigma_{i,j}$ is the (i,j) element of Σ_U. Then $E(\varepsilon_Y^T P_1\varepsilon_X A^{-1}) = (u-p)\alpha_1^T\Sigma_U\alpha_2 A^{-1}$. The last term $E(EA^{-1}CA^{-1}) = E\{\varepsilon_Y^T Z\delta^T A^{-1}\varepsilon_X^T Z\delta^T A^{-1}\} = E\{\alpha_1^T U^T Z\delta^T A^{-1}\alpha_2^T U^T Z\delta^T A^{-1}\}$. Consider the (i,j) element of $E(U^T Z\delta^T A^{-1}\alpha_2^T U^T)$, we obtain that $E(EA^{-1}CA^{-1}) = \alpha_1^T\Sigma_U$
$(Z\delta^T A^{-1}\alpha_2^T)^T Z\delta^T A^{-1} = \alpha_1^T\Sigma_U\alpha_2 A^{-1}$. Thus, $\widehat{\beta}_{2SLS} - \beta = \widehat{b} + o_p(n^{-1+2\epsilon})$, where $E(\widehat{b}) = (u-p-1)\Sigma_{\varepsilon_Y,\varepsilon_X}(\delta Z^T Z\delta^T)^{-1} = \frac{u-p-1}{n}\Sigma_{\varepsilon_Y,\varepsilon_X}(\delta\Sigma_Z\delta^T)^{-1} + o_p(n^{-1})$ and $\Sigma_{\varepsilon_Y,\varepsilon_X}$ is the covariance matrix of ε_{Y_i} and ε_{X_i}. ■

D.7 Proof of Proposition 5.2.3

Proof. Suppose for any matrix $A \in \mathbb{R}^{p\times p}$, $\text{vech}(A) = C_p\text{vec}(A)$ and $\text{vec}(A) = E_p\text{vech}(A)$, where E_p is the expansion matrix and C_p is the contraction matrix.

Because $\widehat{\beta}_{2nd,env} = \widehat{\lambda}_{2nd,env}\widehat{\Sigma}_Z\widehat{\delta}_{2SLS}^T(\widehat{\delta}_{2SLS}\widehat{\Sigma}_Z\widehat{\delta}_{2SLS}^T)^{-1}$, we have $\sqrt{n}\{\text{vec}(\widehat{\beta}_{2nd,env}) - \text{vec}(\beta)\} = (\widehat{\delta}_{2SLS}\widehat{\Sigma}_Z\widehat{\delta}_{2SLS}^T)^{-1}\widehat{\delta}_{2SLS}\widehat{\Sigma}_Z \otimes I_r[\sqrt{n}\{\text{vec}(\widehat{\lambda}_{2nd,env}) - (I_k\otimes\beta)\text{vec}(\widehat{\delta}_{2SLS})]$. Then by Slutsky's theorem, we only need to calculate the asymptotic distribution of $\sqrt{n}\{\text{vec}(\widehat{\delta}_{2SLS}^T - \delta^T)^T, \text{vec}(\widehat{\lambda}_{2nd,env}^T - \lambda^T)^T\}^T$.

Let $F = \log(\Sigma_C) + tr(S_C\Sigma_C^{-1}) + \|X^T - Z^T\delta^T\|_2^2/n$, where $C = (Z, Y)$,
$\Sigma_C = ((\Sigma_Z, \Sigma_Z\lambda^T)^T, (\lambda\Sigma_Z, \Sigma_Y)^T)$, S_Z is the sample estimator for Σ_Z. Because Σ_Y doesn't change in the envelope model and is not the target parameter, we don't consider Σ_Y then. The parameters of the envelope model are

$$\phi = (\text{vec}^T(\delta^T), \text{vec}^T(\eta), \text{vec}^T(\Gamma), \text{vech}^T(\Omega), \text{vech}^T(\Omega_0)),$$

and the unconstrained parameters of 2SLS model are

$$h(\phi) = (\text{vec}^T(\delta^T), \text{vec}^T(\lambda^T), \text{vech}^T(\Sigma_Z)).$$

Thus, $\Delta = \frac{\partial h}{\partial \phi} = \begin{pmatrix} I_{pk} & \\ & H \end{pmatrix}$, where

$$H = \begin{pmatrix} I_p \otimes \Gamma & \eta \otimes I_k & 0 & 0 \\ 0 & 2C_k(\Gamma\Omega \otimes I_k - \Gamma \otimes \Gamma_0\Omega_0\Gamma_0^T) & C_k(\Gamma \otimes \Gamma)E_u & C_k(\Gamma_0 \otimes \Gamma_0)E_{k-u}. \end{pmatrix}.$$

Also,

$$V_0 = \frac{\partial^2 F}{\partial h(\phi)\partial h(\phi)^T} = \begin{pmatrix} I_p \otimes \Sigma_Z & 0 \\ 0 & J \end{pmatrix},$$

where $J = \begin{pmatrix} \Sigma_{Y|Z}^{-1} \otimes \Sigma_Z & 0 \\ 0 & E_k^T(\Sigma_Z^{-1} \otimes \Sigma_Z^{-1})E_k/2 \end{pmatrix}$.

The 2SLS estimator $\sqrt{n}\{\text{vec}(\widehat{\lambda}_{2SLS}^T - \lambda^T)^T, \text{vec}(\widehat{\delta}_{2SLS}^T - \delta^T)^T\}$ converges to a normal distribution with mean 0 and variance $V = \begin{pmatrix} \Sigma_{\varepsilon_X} \otimes \Sigma_Z^{-1} & V_{12} \\ V_{12}^T & J^{-1} \end{pmatrix}$, where $V_{12} = (\Sigma_{X,Y|Z} \otimes \Sigma_Z^{-1}, 0)$.

By Shapiro (1986), $\sqrt{n}\{\text{vec}(\widehat{\delta}_{2SLS}^T - \delta^T)^T, \text{vec}(\widehat{\lambda}_{2nd,env}^T - \lambda^T)^T\}^T$ converges to a normal distribution with mean 0 and variance $V_{2nd,env} = PVP^T$, where $P = \Delta(\Delta^T V_0 \Delta)^- \Delta^T V_0$. Then $V_{2nd,env} = \begin{pmatrix} \Sigma_{X|Z} \otimes \Sigma_Z^{-1} & (\Sigma_{X,Y|Z}\Sigma_{Y|Z}^{-1} \otimes \Sigma_Z^{-1})V_1 \\ * & V_1 \end{pmatrix}$, where $V_1 = \Sigma_{Y|Z} \otimes \Gamma\Omega^{-1}\Gamma^T + (\eta \otimes \Gamma_0)M^\dagger(\eta^T \otimes \Gamma_0^T)$ and $M = \eta^T\Sigma_{Y|Z}^{-1}\eta \otimes \Omega_0 + \Omega \otimes \Omega_0^{-1} + \Omega^{-1} \otimes \Omega_0 - 2I_u \otimes I_{k-u}$. Thus by Slutsky's theorem, we can obtain the asymptotic distribution of $\text{vec}(\widehat{\beta}_{2nd})$ presented.

■

D.8 Proof of Proposition 5.2.4

Proof. Under Model (5.1), suppose $\Gamma \in \mathbb{R}^{k\times q}$ is a basis of the second stage envelope $\mathcal{E}_{\Sigma_Z}\{span(\lambda^T)\}$ and $\Gamma_0 \in \mathbb{R}^{k\times(k-q)}$ is a basis of its orthogonal subspace. Then, $\Sigma_Z = \Gamma\Delta\Gamma^T + \Gamma_0\Delta_0\Gamma_0^T$ and $\lambda^T = \Gamma\eta$, where $\Delta = \Gamma^T\Sigma_Z\Gamma^T$, $\Delta_0 = \Gamma_0^T\Sigma_Z\Gamma_0^T$ and $\eta \in \mathbb{R}^{q\times r}$.

We have $\sqrt{n}(\widehat{\beta} - \beta) = \sqrt{n}(\widehat{\lambda}_{2nd,env} - \beta\widehat{\delta}_{2SLS})S_Z\widehat{\delta}_{2SLS}^T(\widehat{\delta}_{2SLS}S_Z\widehat{\delta}_{2SLS}^T)^{-1}$, where $S_Z = Z^TZ/n$, $\widehat{\delta}_{2SLS} = X^TZ(Z^TZ)^{-1}$ and $\widehat{\lambda}_{2nd,env} = Y^TZ\Gamma(\Gamma^TZ^TZ\Gamma)^{-1}\Gamma^T$. We now focus on the asymptotic behavior of $\sqrt{n}\text{vec}(\widehat{\lambda}_{2nd,env} - \beta\widehat{\delta}_{2SLS})$. Note that $\sqrt{n}\text{vec}(\widehat{\lambda}_{2nd,env} - \beta\widehat{\delta}_{2SLS}) = \sqrt{n}\text{vec}\{Y^TZ\Gamma(\Gamma^TZ^TZ\Gamma)^{-1}\Gamma^T - \beta X^TZ(Z^TZ)^{-1}\} = \text{vec}\{\{(\beta\varepsilon_X + \varepsilon_Y)^TZ/\sqrt{n}\}\Gamma(\Gamma^TS_Z\Gamma)^{-1}\Gamma^T - (\beta\varepsilon_X^TZ/\sqrt{n})S_Z^{-1}\}$. By Central Limit The-

orem, we have $\|(\beta\varepsilon_X + \varepsilon_Y)^T Z/\sqrt{n}\|_F = O_p(1)$ and $\|\beta\varepsilon_X^T Z/\sqrt{n}\|_F = O_p(1)$. Because $\|S_Z - \Sigma_Z\|_F = o_p(1)$, we have $\text{vec}\{\{(\beta\varepsilon_X + \varepsilon_Y)^T Z/\sqrt{n}\}\Gamma(\Gamma^T S_Z \Gamma)^{-1}\Gamma^T$
$-(\beta\varepsilon_X^T Z/\sqrt{n})S_Z^{-1}\} = \text{vec}\{\{(\beta\varepsilon_X + \varepsilon_Y)^T Z/\sqrt{n}\}\Gamma(\Gamma^T\Sigma_Z\Gamma)^{-1}\Gamma^T - (\beta\varepsilon_X^T Z/\sqrt{n})\Sigma_Z^{-1}\} +$
$o_p(1) = \text{vec}\{(\varepsilon_Y^T Z/\sqrt{n})\Gamma\Delta^{-1}\Gamma^T - (\beta\varepsilon_X^T Z/\sqrt{n})\Gamma_0\Delta_0^{-1}\Gamma_0^T\} = \{\Gamma\Delta^{-1}\Gamma^T \otimes (0_{r\times p}, I_r) -$
$\Gamma_0\Delta_0^{-1}\Gamma_0^T \otimes (\beta, 0_{r\times r})\}\text{vec}\{(\varepsilon_X, \varepsilon_Y)^T Z/\sqrt{n}\} + o_p(1)$. By Slutsky's theorem, note that $\text{vec}\{(\varepsilon_X, \varepsilon_Y)^T Z/\sqrt{n}\}$ converges to $N(0, V)$ in distribution by Central Limit Theorem, we have $\sqrt{n}\text{vec}(\widehat{\lambda}_{2nd,env} - \beta\widehat{\delta}_{2SLS})$ converges to $N(0, UVU^T)$ in distribution, where $V = \Sigma_Z \otimes \Sigma$, $\Sigma = cov((\varepsilon_X^T, \varepsilon_Y^T)^T)$, and $U = \Gamma\Delta^{-1}\Gamma^T \otimes (0_{r\times p}, I_r) -$ $\Gamma_0\Delta_0^{-1}\Gamma_0^T \otimes (\beta, 0_{r\times r})$. We can simplify UVU^T as $\Gamma\Delta^{-1}\Gamma^T \otimes \Sigma_{\varepsilon_Y} + \Gamma_0\Delta_0^{-1}\Gamma_0^T \otimes$ $\beta\Sigma_{\varepsilon_X}\beta^T$. Thus, because S_Z and $\widehat{\delta}_{2SLS}$ converge to Σ_Z and δ in probability respectively, by Slutsky's theorem, $\sqrt{n}(\widehat{\beta}_{2nd,env} - \beta)$ converges in distribution to a normal distribution with mean 0 and variance $(\delta\Sigma_Z\delta^T)^{-1}\delta\Gamma\Delta\Gamma^T\delta^T(\delta\Sigma_Z\delta^T)^{-1} \otimes \Sigma_{\varepsilon_Y} +$ $(\delta\Sigma_Z\delta^T)^{-1}\delta\Gamma_0\Delta_0\Gamma_0^T\delta^T(\delta\Sigma_Z\delta^T)^{-1} \otimes \beta\Sigma_{\varepsilon_X}\beta^T$.

■

D.9 Proof of Lemma D.1.1

Proof. Note that

$$E[\{X_i - E(X_i|Z_i)\}Y_i^T|X_i, Z_i, U_i] = \{X_i - E(X_i|Z_i)\}X_i^T\beta^T + \{X_i - E(X_i|Z_i)\}\beta_Z(Z_i, U_i)^T + \{X_i - E(X_i|Z_i)\}\beta_U(U_i)^T := A_i + B_i + C_i.$$

The conditional expectation of the second and third term

$$\begin{aligned} E(B_i|Z_i, U_i) &= E[\{X_i - E(X_i|Z_i)\}\beta_Z(Z_i, U_i)^T|Z_i, U_i] \\ &= \{E(X_i|Z_i, U_i) - E(X_i|Z_i)\}\beta_Z(Z_i, U_i)^T \\ &= \{\alpha_Z(Z_i, U_i) + \alpha_U(U_i) - E(X_i|Z_i)\}\beta_Z(Z_i, U_i)^T, \end{aligned}$$

$$E(C_i|Z_i, U_i) = E[\{X_i - E(X_i|Z_i)\}\beta_U(U_i)^T|Z_i, U_i] = [\alpha_Z(Z_i, U_i) + \alpha_U(U_i) - E\{\alpha_Z(Z_i, U_i)|Z_i\}]\beta_U(U_i)^T.$$

Thus, by Assumption D.1.1,

$$
\begin{aligned}
E(B_i|Z_i) &= E[\{X_i - E(X_i|Z_i)\}\beta_Z(Z_i,U_i)^T|Z_i] = cov\{\alpha_Z(Z_i,U_i) + \alpha_U(U_i), \beta_Z(Z_i,U_i)|Z_i = 0,\\
E(C_i|Z_i) &= E[\{\alpha_Z(Z_i,U_i) + \alpha_U(U_i) - E\{\alpha_Z(Z_i,U_i)|Z_i\}\}\beta_U(U_i)^T|Z_i]\\
&= cov\{\alpha_Z(Z_i,U_i), \beta_U(U_i)|Z_i\} + E\{\alpha_U(U_i)\beta_U(U_i)^T|Z_i\} = 0.
\end{aligned}
$$

Then

$$
\begin{aligned}
E(Z_iY_{i,new}^T) &= E\{Z_ivec(A_i + B_i + C_i)^T\} = E[E\{Z_ivec(A_i)^T|Z_i\}] = E[E\{Z_ivech(A_i)^TE_p^T|Z_i\}]\\
&= E[Z_ivech\{\{X_i - E(X_i|Z_i)\}X_i^T\}^T]E_p^T(\beta^T \otimes I_p) = E(Z_iX_{i,new}^T)\Psi^T,
\end{aligned}
$$

where $\Psi = (\beta \otimes I_p)E_p$. ■

He did not know that the new life would not be given him for nothing... But that is the beginning of a new story – the story of the gradual renewal of a man, the story of his gradual regeneration, of his passing from one world into another, of his initiation into a new unknown life. That might be the subject of a new story, but our present story is ended.

(F. Dostoevsky, Crime and Punishment, Epilogue.)

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control 19*, 716–723.

Anderson, A., W. Yang, R. Townsend, Q. Pan, G. Chertow, J. Kusek, J. Charleston, J. He, R. Kallem, J. Lash, et al. (2015). Time-updated systolic blood pressure and the progression of chronic kidney disease: a cohort study. *Annals of Internal Medicine 162*, 258–265.

Anderson, T. and H. Rubin (1949). Estimators of the parameters of a single equation in a complete set of stochastic equations. *Annals of Mathematical Statistics 21*, 570–581.

Andrews, I. and J. Stock (2018). Weak instruments and what to do about them.

Angrist, J., G. Imbens, and D. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association 91*, 444–455.

Angrist, J. D. and A. B. Keueger (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics 106*, 979–1014.

Bansal, N., M. Keane, P. Delafontaine, D. Dries, E. Foster, C. Gadegbeku, A. Go, L. Hamm, J. Kusek, A. Ojo, et al. (2013). A longitudinal study of left ventricular function and structure from CKD to ESRD: the CRIC study. *Clinical Journal of the American Society of Nephrology 8*, 355–362.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica 62*, 657–81.

Bi, X. and A. Qu (2015). Sufficient dimension reduction for longitudinal data. *Statistica Sinica 25*, 787–807.

Blundell, R., L. Dearden, and B. Sianesi (2005). Evaluating the effect of education on earnings: models, methods and results from the national child development survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 168*, 473–512.

Bound, J., D. A. Jaeger, and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association 90*, 443–450.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Breiman, L. and J. H. Friedman (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59*, 3–54.

Budoff, M., D. Rader, M. Reilly, E. Mohler, J. Lash, W. Yang, L. Rosen, M. Glenn, V. Teal, and H. Feldman (2011). Relationship of estimated GFR and coronary artery calcification in the CRIC (Chronic Renal Insufficiency Cohort) study. *American Journal of Kidney Diseases 58*, 519–526.

Buse, A. (1992). The bias of instrumental variable estimators. *Econometrica: Journal of the Econometric Society 60*, 173–180.

Capuano, V., A. Bambacaro, T. D'Arminio, G. Vecchio, and L. Cappuccio (2003). Correlation between body mass index and others risk factors for cardiovascular disease in women compared with men. *Monaldi Archives for Chest Disease 60*, 295–300.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics 3*, 1801–1863.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica 69*, 1127–1160.

Casella, G. and R. L. Berger (2021). *Statistical inference*. Cengage Learning.

Chamberlain, G. and G. Imbens (2004). Random effects estimators with many instrumental variables. *Econometrica 72*, 295–306.

Chaussé, P. (2010). Computing generalized method of moments and generalized empirical likelihood with r. *Journal of Statistical Software 34*, 1–35.

Chen, Q., J. G. Ibrahim, M.-H. Chen, and P. Senchaudhuri (2008). Theory and inference for regression models with missing responses and covariates. *Journal of multivariate analysis 99*, 1302–1331.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2 1*, 1–4.

Cook, R. (2018a). *An introduction to envelopes: dimension reduction for efficient estimation in multivariate statistics*, Volume 401. John Wiley & Sons.

Cook, R. and L. Ni (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association 100*, 410–428.

Cook, R. and S. Weisberg (1991a). Comment. *Journal of the American Statistical Association 86*, 328–332.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*, Volume 482. John Wiley & Sons.

Cook, R. D. (2018b). Principal Components, Sufficient Dimension Reduction, and Envelopes. *Annual Review of Statistics and Its Application 5*, 533–559.

Cook, R. D. and L. Forzani (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association 104*, 197–208.

Cook, R. D., L. Forzani, and L. Liu (2021). Envelopes for multivariate linear regression with linearly constrained coefficients. *arXiv preprint arXiv:2101.00514*.

Cook, R. D., L. Forzani, and Z. Su (2016). A note on fast envelope estimation. *Journal of Multivariate Analysis 150*, 42–54.

Cook, R. D., L. Forzani, and X. Zhang (2015). Envelopes and reduced-rank regression. *Biometrika 102*, 439–456.

Cook, R. D., I. Helland, and Z. Su (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*, 851–877.

Cook, R. D., B. Li, and F. Chiaromonte (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica 20*, 927–960.

Cook, R. D. and Z. Su (2013). Scaled envelopes: scale-invariant and efficient estimation in multivariate linear regression. *Biometrika 100*, 939–954.

Cook, R. D. and S. Weisberg (1991b). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association 86*, 328–332.

Cook, R. D. and X. Zhang (2015a). Foundations for envelope models and methods. *Journal of the American Statistical Association 110*, 599–611.

Cook, R. D. and X. Zhang (2015b). Simultaneous envelopes for multivariate linear regression. *Technometrics 57*, 11–25.

Cook, R. D. and X. Zhang (2016). Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics 25*, 284–300.

Cook, R. D. and X. Zhang (2018). Fast envelope algorithms. *Statistica Sinica 28*, 1179–1197.

De Jong, S. (1993). Simpls: an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems 18*, 251–263.

Donald, S. G. and W. K. Newey (2001). Choosing the number of instruments. *Econometrica 69*, 1161–1191.

Dong, Y. and B. Li (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika 97*, 279–294.

Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in r. *Journal of statistical software 21*, 1–16.

Eck, D. J. and R. D. Cook (2017). Weighted envelope estimation to handle variability in model selection. *Biometrika 104*, 743–749.

Feldman, H., L. Appel, G. Chertow, D. Cifelli, B. Cizman, J. Daugirdas, J. Fink, E. Franklin-Becker, A. Go, L. Hamm, et al. (2003). The chronic renal insufficiency cohort (CRIC) study: design and methods. *Journal of the American Society of Nephrology 14*, S148–S153.

Ferguson, J., G. Matthews, R. Townsend, D. Raj, P. Kanetsky, M. Budoff, M. Fischer, S. Rosas, R. Kanthety, M. Rahman, et al. (2013). Candidate gene association study of coronary artery calcification in chronic kidney disease: findings from the CRIC study (Chronic Renal Insufficiency Cohort). *Journal of the American College of Cardiology 62*, 789–798.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics 7*(2), 179–188.

Foster, M. C., J. Coresh, J. V. Bonventre, V. S. Sabbisetti, S. S. Waikar, T. E. Mifflin, R. G. Nelson, M. Grams, H. I. Feldman, R. S. Vasan, et al. (2015). Urinary biomarkers and risk of esrd in the atherosclerosis risk in communities study. *Clinical Journal of the American Society of Nephrology 10*, 1956–1963.

Fung, W., X. He, L. Liu, and P. Shi (2002). Dimension reduction based on canonical correlation. *Statistica Sinica*, 1093–1113.

Goldberger, A. (1972). Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society 40*, 979–1001.

Hall, A. R. (2004). *Generalized method of moments*. OUP Oxford.

Hall, A. R. and F. P. Peixe (2003). A consistent method for the selection of relevant instruments. *Econometric reviews 22*, 269–287.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory 24*, 726–748.

Hayfield, T. and J. S. Racine (2008). Nonparametric econometrics: The np package. *Journal of statistical software 27*, 1–32.

He, J., M. Reilly, W. Yang, J. Chen, A. Go, J. Lash, M. Rahman, C. DeFilippi, C. Gadegbeku, R. Kanthety, et al. (2012). Risk factors for coronary artery calcium among patients with chronic kidney disease (from the Chronic Renal Insufficiency Cohort study). *American Journal of Cardiology 110*, 1735–1741.

Heckman, J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources 32*, 441–462.

Henderson, H. V. and S. Searle (1979). Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics. *Canadian Journal of Statistics 7*, 65–81.

Herd, P., D. Carr, and C. Roan (2014). Cohort profile: Wisconsin longitudinal study (wls). *International journal of epidemiology 43*(1), 34–41.

Hernán, M. and J. M. Robins (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology 17*, 360–372.

Hotelling, H. (1936a). Relations between two sets of variates. *Biometrika 28*(3-4), 321–377.

Hotelling, H. (1936b). Relations between two sets of variates. *Biometrika 28*, 321–377.

Hristache, M. and V. Patilea (2017). Conditional moment models with data missing at random. *Biometrika 104*, 735–742.

Hughes, J. and M. Haran (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*, 139–159.

Ibrahim, J. G., H. Zhu, and N. Tang (2008). Model selection criteria for missing-data problems using the em algorithm. *Journal of the American Statistical Association 103*, 1648–1658.

Imbens, G. and J. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*, 467–475.

Imbens, G. and P. Rosenbaum (2005). Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 168*, 109–126.

Inker, L. A., J. Coresh, Y. Sang, C.-y. Hsu, M. C. Foster, J. H. Eckfeldt, A. B. Karger, R. G. Nelson, X. Liu, M. Sarnak, et al. (2017). Filtration markers as

predictors of ESRD and mortality: individual participant data meta-analysis. *Clinical Journal of the American Society of Nephrology 12*, 69–78.

Jia, J., Y. Benjamini, C. Lim, G. Raskutti, and B. Yu (2010). Envelope models for parsimonious and efficient multivariate linear regression comment.

Jones, R. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine 30*, 3050–3056.

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica 70*, 1781–1803.

Kleibergen, F. (2005). Testing parameters in GMM without assuming that they are identified. *Econometrica 73*, 1103–1123.

Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics 30*, 67–80.

Li, B. and Y. Dong (2009). Dimension reduction for nonelliptically distributed predictors. *The Annals of Statistics 37*, 1272–1298.

Li, B. and S. Wang (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association 102*, 997–1008.

Li, B., H. Zha, and F. Chiaromonte (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics 33*, 1580–1616.

Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*, 316–327.

Li, K. (1992). On principal hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association 87*, 1025–1039.

Li, L. and X. Zhang (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association 112*, 1131–1146.

Li, Q. and J. S. Racine (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association 87*, 1227–1237.

Little, R. J. and D. B. Rubin (2014). *Statistical Analysis with Missing Data*, Volume 333. John Wiley & Sons.

Liu, L., W. Miao, B. Sun, J. Robins, and E. Tchetgen Tchetgen (2017). Identification and inference for marginal average treatment effect on the treated with an instrumental variable. *Statistica Sinica*.

Ma, L., H. Kang, and L. Liu (2022+). Improving instrumental variable estimation by removing redundant instruments. *Technical Report*.

Ma, L., L. Liu, and W. Yang (2021). Envelope method with ignorable missing data. *Electronic Journal of Statistics 15*(2), 4420–4461.

Ma, Y. and L. Zhu (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association 107*, 168–179.

Ma, Y. and L. Zhu (2013). Efficient estimation in sufficient dimension reduction. *Annals of statistics 41*, 250.

Ma, Y. and L. Zhu (2014). On estimation efficiency of the central mean subspace. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(5), 885–901.

Madjid, M. and O. Fatemi (2013). Components of the complete blood count as risk predictors for coronary heart disease: in-depth review and update. *Texas Heart Institute Journal 40*, 17–29.

Magus, J. and H. Neudecker (1984). Matrix differential calculus with applications to simple, hadamard, and kronecker products. Technical report.

Matsouaka, R. and E. Tchetgen Tchetgen (2017). Instrumental variable estimation of causal odds ratios using structural nested mean models. *Biostatistics 18*, 465–476.

Meng, X.-L. and D. B. Rubin (1991). Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association 86*, 899–909.

Moreira, M. (2003). A conditional likelihood ratio test for structural models. *Econometrica 71*, 1027–1048.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

Nocedal, J. and S. Wright (2006). *Numerical optimization*. Springer Science & Business Media.

Park, M., C.-Y. Hsu, A. S. Go, H. I. Feldman, D. Xie, X. Zhang, T. Mifflin, S. S. Waikar, V. S. Sabbisetti, J. V. Bonventre, et al. (2017). Urine kidney injury biomarkers and risks of cardiovascular disease events and all-cause death: The CRIC study. *Clinical Journal of the American Society of Nephrology 12*, 761–771.

Park, Y., Z. Su, and H. Zhu (2017). Groupwise envelope models for imaging genetic analysis. *Biometrics 73*, 1243–1253.

Pfeiffer, R., L. Forzani, and E. Bura (2012). Sufficient dimension reduction for longitudinally measured predictors. *Statistics in Medicine 31*, 2414–2427.

Rekabdarkolaee, H. M., Q. Wang, Z. Naji, and M. Fuente (2020). New parsimonious multivariate spatial model: spatial envelope. *Statistica Sinica 30*, 1583–1604.

Robins, J. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS 113*, 159.

Robins, J. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods 23*, 2379–2412.

Robins, J. and A. Rotnitzky (2004). Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika 91*, 763–783.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics 6*, 461–464.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, Volume 162. John Wiley & Sons.

Shao, J. (2003). *Mathematical Statistics*. Springer Science & Business Media.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association 81*, 142–149.

Shi, Y., L. Ma, and L. Liu (2020). Mixed effects envelope models. *Stat 9*(1), e313.

Small, D. and P. Rosenbaum (2008). War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association 103*, 924–933.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*, 557–586.

Su, Z. and R. D. Cook (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika 98*, 133–146.

Su, Z. and R. D. Cook (2012). Inner envelopes: efficient estimation in multivariate linear regression. *Biometrika 99*, 687–702.

Su, Z. and R. D. Cook (2013). Estimation of multivariate means with heteroscedastic errors using envelope models. *Statistica Sinica 23*, 213–230.

Su, Z., G. Zhu, X. Chen, and Y. Yang (2016). Sparse envelope model: efficient estimation and response variable selection in multivariate linear regression. *Biometrika 103*, 579–593.

Sun, B., L. Liu, W. Miao, K. Wirth, J. Robins, and E. Tchetgen Tchetgen (2017). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*.

Tan, Z. (2010). Marginal and nested structural models using instrumental variables. *Journal of the American Statistical Association 105*, 157–169.

Tchetgen Tchetgen, E., B. Sun, and S. Walter (2017). The genius approach to robust mendelian randomization inference. *arXiv preprint arXiv:1709.07779*.

Vansteelandt, S. and E. Goetghebeur (2003). Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*, 817–835.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, 391–420.

Wright, S. (1928). Appendix to the tariff on animal and vegetable oils. *New York: MacMillan.(1934)," The Method of Path Coefficients," Annals of Mathematical Statistics 5*, 161–215.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics 11*, 95–103.

Ye, Z. and R. E. Weiss (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association 98*, 968–979.

Zhang, X., C. Wang, and Y. Wu (2018). Functional envelope for model-free sufficient dimension reduction. *Journal of Multivariate Analysis 163*, 37–50.

Zhou, J. and X. He (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *The Annals of Statistics 36*, 1649–1668.

Zhou, L., J. Huang, J. Martinez, A. Maity, V. Baladandayuthapani, and R. Carroll (2010). Reduced rank mixed effects models for spatially correlated hierarchical functional data. *Journal of the American Statistical Association 105*, 390–400.

Zhu, G. and Z. Su (2020). Envelope-based sparse partial least squares. *The Annals of Statistics 48*, 161–182.

Zhu, L., T. Wang, L. Zhu, and L. Ferré (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika 97*, 295–304.

Zhu, L., L. Zhu, and Z. Feng (2010). Dimension reduction in regressions via average partial mean estimation. *Journal of the American Statistical Association 105*, 1455–1466.