Envelope Methods with Ignorable Missing Data

Linquan Ma[1,2], Lan Liu[2] and Wei Yang [3]

[1]Department of Statistics, University of Wisconsin - Madison, Madison, Wisconsin, USA

[2]School of Statistics, University of Minnesota at Twin Cities, Minneapolis, Minnesota, USA

[3]Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Abstract

Envelope method was recently proposed as a method to reduce the dimension of responses in multivariate regressions. However, when there exists missing data, the envelope method using the complete case observations may lead to biased and inefficient results. In this paper, we generalize the envelope estimation when the predictors and/or the responses are missing at random. Specifically, we incorporate the envelope structure in the expectation-maximization (EM) algorithm. As the parameters under the envelope method are not pointwise identifiable, the EM algorithm for the envelope method was not straightforward and requires a special decomposition. Our method is guaranteed to be more efficient, or at least as efficient as, the standard EM algorithm. Moreover, our method has the potential to outperform the full data MLE. We give asymptotic properties of our method under both normal and non-normal cases. The efficiency gain over the standard EM is confirmed in simulation studies and in an application to the Chronic Renal Insufficiency Cohort (CRIC) study.

Keywords: EM-algorithm; Efficiency gain; Sufficient dimension reduction; Missing data; Multivariate regression.

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 a condition, 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, 2018). The envelope conditions 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), model-free envelope (Cook and Zhang, 2015a), and mixed effects envelope (Shi et al., 2020). 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 inefficient or possibly biased results. We assume the missingness mechanism is MAR throughout this paper.

In this paper, 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 a 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 paper is among the first few in the dimension reduction literature to discuss the case where both responses and predictors are subject to missingness.

We organize the paper as follows. In Section 2, we introduce the notations and review the envelope models. In Section 3, we present the observed data likelihood and clarify the difficulty of applying the envelope method directly. In Section 4, we propose an EM envelope algorithm. Simulations are given in Section 5, where we compare the EM envelope method with the existing methods. In Section 6, we apply the EM envelope to the Chronic Renal Insufficiency Cohort (CRIC) data. In Section 7, we present a brief discussion. Section 8 contains the link to our R package.

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

$$\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\boldsymbol{\varepsilon}_i$ are identically and independently (i.i.d) distributed with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}$, and $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$. We firstly assume the normality of the error when deriving the EM envelope estimator. We extend later (Propositions 2 and 3) the robustness property of our estimator when the normality is possibly violated. Let $R_{X_{ij}} = 1$ if X_{ij} is observed and $R_{X_{ij}} = 0$ if otherwise, for $j = 1, \ldots, p$. Similarly, let $R_{Y_{ik}}$ denote the missing indicator for Y_{ik}, for $k = 1, \ldots, r$. 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 1 (ignorability)**.** $\mathbf{R}_i \perp\!\!\!\perp \mathbf{D}_{i,mis} \mid \mathbf{D}_{i,obs}$.

Assumption 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. A matrix $\mathbf{O} \in \mathbb{R}^{r\times r}$ is orthonormal 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. Consider an orthonormal matrix $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{r\times r}$ such that

Condition 1. $\text{span}(\boldsymbol{\beta}) \subseteq \text{span}(\mathbf{\Gamma})$,

Condition 2. $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$,

where $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$, $\mathbf{\Gamma}_0 \in \mathbb{R}^{r\times(r-u)}$, and $0 \le u \le r$. The subspace $\text{span}(\mathbf{\Gamma})$ satisfying Conditions 1 and 2 is not unique, but Cook et al. (2010) defined the envelope to be the smallest subspace satisfying these conditions. The dimension u is known as the envelope dimension. Notice the decomposition of $\mathbf{\Sigma}$ is equivalent to $\text{cor}(\mathbf{\Gamma}_0^T\mathbf{Y}, \mathbf{\Gamma}^T\mathbf{Y} \mid \mathbf{X}) = 0$. From $\text{span}(\boldsymbol{\beta}) \subseteq \text{span}(\mathbf{\Gamma})$, the regression parameter can be written as $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}$, where $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$. Therefore, the envelope model can also be written as follows:

$$\mathbf{Y}_i = \mathbf{\Gamma}\boldsymbol{\eta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i, \quad \mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T. \tag{2}$$

The null correlation only guarantees the information of $\mathbf{\Gamma}_0^T\mathbf{Y}$ is immaterial in the first two moments. Under the normality assumption of the error, Conditions 1–2 are equivalent to the following two conditions:

Condition 3. $\mathbf{\Gamma}_0^T\mathbf{Y} \perp\!\!\!\perp \mathbf{X}$.

Condition 4. $\mathbf{\Gamma}^T\mathbf{Y} \perp\!\!\!\perp \mathbf{\Gamma}_0^T\mathbf{Y} \mid \mathbf{X}$.

Conditions 3–4 are equivalent to $\boldsymbol{\Gamma}_0^T\mathbf{Y} \perp\!\!\!\perp (\boldsymbol{\Gamma}^T\mathbf{Y}, \mathbf{X})$.

Although the original envelope was developed using Conditions 1–2, we directly define envelope using Conditions 3–4. The envelope under Conditions 3–4 is in general no smaller than that defined by Conditions 1–2. We prefer Conditions 3–4 because the interpretation of the envelope is more straightforward especially when the normality is violated.

We give a simple example for the envelope model. 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, $\boldsymbol{\Gamma} = (1, -1)^T/\sqrt{2}$, and $\boldsymbol{\Gamma}_0 = (1, 1)^T/\sqrt{2}$. The combinations of responses that are involved in the regression, $\boldsymbol{\Gamma}^T\mathbf{Y}$, is called the material part of $\mathbf{Y}$, and the part that is uninvolved, $\boldsymbol{\Gamma}_0^T\mathbf{Y}$, is called the immaterial part of $\mathbf{Y}$. Hence, the main focus of the envelope method is to find the column space of $\boldsymbol{\Gamma}$, i.e., span($\boldsymbol{\Gamma}$), that fully contains the information of $\boldsymbol{\beta}$, i.e., find an envelope of $\boldsymbol{\beta}$.

Once an estimate of the basis $\boldsymbol{\Gamma}$, $\hat{\boldsymbol{\Gamma}}$, is obtained, $\hat{\boldsymbol{\beta}}_{env}$ is obtained by projecting the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ onto the estimated envelope space, $\hat{\boldsymbol{\beta}}_{env} = \mathbf{P}_{\hat{\boldsymbol{\Gamma}}}\hat{\boldsymbol{\beta}}$, where $\mathbf{P}_{\mathbf{A}}$ stands for the projection matrix for the matrix $\mathbf{A}$.

Figure 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(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 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(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(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(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.

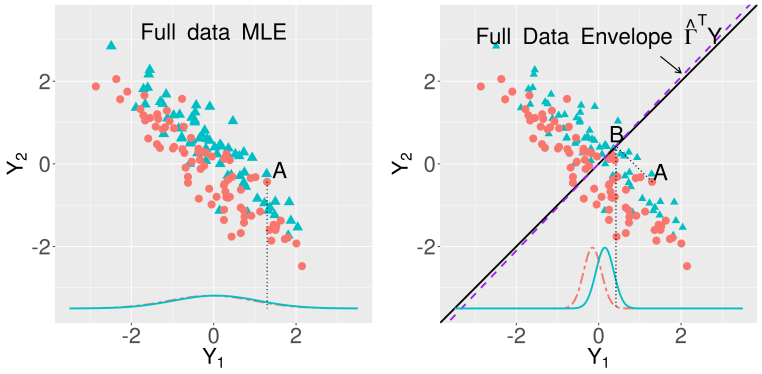
Now, consider the case where the predictors $\mathbf{X}$ are fully observed but some values of the responses are missing (see Figure 2). The missingness mechanism is as follows. For an individual i for $i = 1, \ldots, 150$, if $X_i = 1$ and if Y_{i1} is among the largest 30 $Y_{i'1}$ for

$i^{'} = 1, \ldots, 150$, 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 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 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 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.

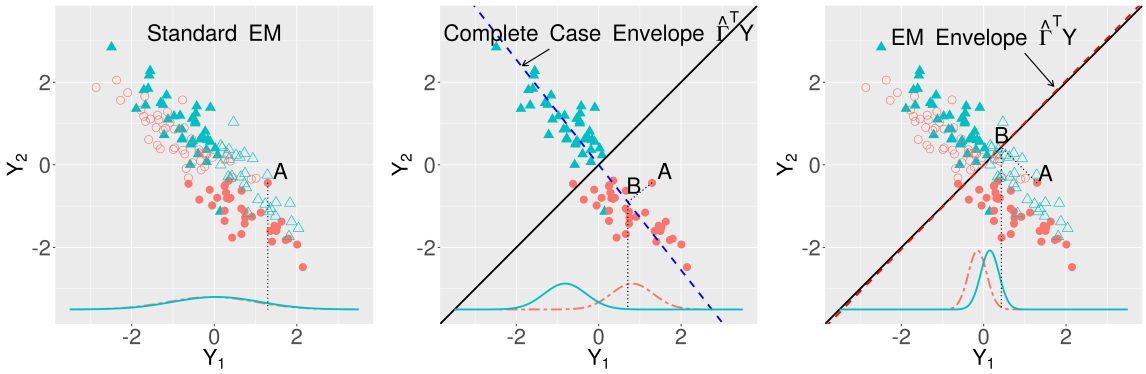
Our method is shown in Figure 3(c). Different from the complete case analysis, we use both the complete cases and the partially missing information. Our proposed method is asymptotically unbiased when the missing pattern is MAR. The estimated envelope direction is shown as the red dashed line. Our method 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.

Figure 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 the envelope method are shown at the bottom of each subfigure.



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

Figure 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

3 The Observed Data Likelihood

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

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

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 $\int f(\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis} \mid \mathbf{x}_i; \boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}_0, \boldsymbol{\Omega}) d\mathbf{y}_{i,mis} = f(\mathbf{y}_{i,obs} \mid \mathbf{x}_i; \boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}_0, \boldsymbol{\Omega})$, we can simplify the observed data likelihood as

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

The first part of the observed data likelihood corresponds to the likelihood of individuals with fully observed predictors. The second part corresponds to the likelihood of individuals with missing predictors. Hence, the 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 3 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 first order derivative of the objective function 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 challenging to calculate the maximum likelihood estimates under an envelope structure. 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.

4 The EM Envelope

4.1 The EM updates

Let $l_{full}(\boldsymbol{\phi} \mid L) = \log L_{full}(\boldsymbol{\phi} \mid L)$ denote the log of full data likelihood, where $\boldsymbol{\phi} = (\boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}_0, \boldsymbol{\Omega}, \boldsymbol{\rho})$. Then, the logarithm of full data likelihood of $(\mathbf{X}, \mathbf{Y})$ is

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

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

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

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 $\boldsymbol{\phi} = (\boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\rho})$ as the new parameters for the reparameterization. Hence, we have

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

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

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

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

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

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 $Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}_t)$ as $Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}_t) = Q_1(\boldsymbol{\rho} \mid \boldsymbol{\phi}_t) + Q_2(\boldsymbol{\beta},\boldsymbol{\Sigma} \mid \boldsymbol{\phi}_t)$, where $Q_1(\boldsymbol{\rho} \mid \boldsymbol{\phi}_t) = \mathbb{E}[2\log\{f_x(\mathbf{X}_i \mid \boldsymbol{\rho})\} \mid \mathbf{D}_{obs};\boldsymbol{\phi}_t] + 2C$, and $Q_2(\boldsymbol{\beta},\boldsymbol{\Sigma} \mid \boldsymbol{\phi}_t) = -n\log|\boldsymbol{\Sigma}| - \text{tr}\{\boldsymbol{\Sigma}^{-1}\big(\mathbf{A}_{1,t} - 2\mathbf{A}_{2,t}\boldsymbol{\beta}^T + \boldsymbol{\beta}\mathbf{A}_{3,t}\boldsymbol{\beta}^T\big)\}$. As $Q_1(\boldsymbol{\rho} \mid \boldsymbol{\phi}_t)$ only involves $\boldsymbol{\rho}$, the maximizer of $Q_1(\boldsymbol{\rho} \mid \boldsymbol{\phi}_t)$ is $\boldsymbol{\rho}_{t+1} = \arg\max_{\boldsymbol{\rho}\in\boldsymbol{\Pi}} \mathbb{E}[2\log\{f_x(\mathbf{x}_i \mid \boldsymbol{\rho})\} \mid \mathbf{D}_{obs};\boldsymbol{\phi}_t]$, where $\boldsymbol{\Pi}$ is the parameter space of $\boldsymbol{\rho}$.

To find the maximizer of $Q_2(\boldsymbol{\beta},\boldsymbol{\Sigma} \mid \boldsymbol{\phi}_t)$, note under the envelope conditions 3–4, 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 $\text{Span}(\boldsymbol{\beta}) \subseteq \text{Span}(\boldsymbol{\Sigma}_1)$. 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:

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

where $\det_0(\mathbf{A})$ denotes the product of its non-zero eigenvalues. Further, we have $Q_2(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \boldsymbol{\phi}_t) = Q_{2,1}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_1 \mid \boldsymbol{\phi}_t) + Q_{2,2}(\boldsymbol{\Sigma}_2 \mid \boldsymbol{\phi}_t)$, where $Q_{2,1}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_1 \mid \boldsymbol{\phi}_t) = -n \log \det_0 \boldsymbol{\Sigma}_1 - \mathrm{tr}\{\boldsymbol{\Sigma}_1^{\dagger}(\mathbf{A}_{1,t} - 2\mathbf{A}_{2,t}\boldsymbol{\beta}^T + \boldsymbol{\beta}\mathbf{A}_{3,t}\boldsymbol{\beta}^T)\}$, and $Q_{2,2}(\boldsymbol{\Sigma}_2 \mid \boldsymbol{\phi}_t) = -n \log \det_0 \boldsymbol{\Sigma}_2 - \mathrm{tr}\big(\boldsymbol{\Sigma}_2^{\dagger}\mathbf{A}_{1,t}\big)$. 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 $Q_{2,1}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_1 \mid \boldsymbol{\phi}_t)$ subjects to $\mathrm{Span}(\boldsymbol{\beta}) \subseteq \mathrm{Span}(\boldsymbol{\Sigma}_1)$ 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 $Q_{2,1}(\boldsymbol{\beta}_{t+1}, \boldsymbol{\Sigma}_1 \mid \boldsymbol{\phi}_t) = -n \log \det_0 \boldsymbol{\Sigma}_1 - \mathrm{tr}\{\boldsymbol{\Sigma}_1^{\dagger}(\mathbf{A}_{1,t} - \mathbf{A}_{2,t}\mathbf{A}_{3,t}^{-1}\mathbf{A}_{2,t}^T)\}$.

In order to maximize $Q_{2,1}(\boldsymbol{\beta}_{t+1}, \boldsymbol{\Sigma}_1 \mid \boldsymbol{\phi}_t)$, $Q_{2,2}(\boldsymbol{\Sigma}_2 \mid \boldsymbol{\phi}_t)$ over $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, we use the Lemma 4.3 in Cook et al. (2010), which is reviewed as Lemma 5 in the Appendix. Suppose matrix $\boldsymbol{\Gamma}$ is given, then by Lemma 5, 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, $Q_{2,1}(\boldsymbol{\beta}_{t+1}, \boldsymbol{\Sigma}_{1,t+1} \mid \boldsymbol{\phi}_t) = C_1 - n \log \det_0\{\mathbf{P}_{\boldsymbol{\Gamma}}(\mathbf{A}_{1,t} - \mathbf{A}_{2,t}\mathbf{A}_{3,t}^{-1}\mathbf{A}_{2,t}^T)\mathbf{P}_{\boldsymbol{\Gamma}}\}$, $Q_{2,2}(\boldsymbol{\Sigma}_{2,t+1} \mid \boldsymbol{\phi}_t) = C_2 - n \log \det_0\big(\mathbf{Q}_{\boldsymbol{\Gamma}}\mathbf{A}_{1,t}\mathbf{Q}_{\boldsymbol{\Gamma}}\big)$, 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 $\mathrm{Span}(\boldsymbol{\Gamma})$, we only need to estimate the span of the column space of $\boldsymbol{\Gamma}$, which is identifiable. The MLE of $\mathrm{Span}(\boldsymbol{\Gamma})$ can be obtained using full Grassmannian optimization (Cook et al., 2010, 2016).

4.2 Selection of the envelope dimension

The selection of the envelope dimension can be viewed as a diagnostic or model selection under the envelope framework. 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 $\mathbb{E}\{\log f(\mathbf{D}_{obs} \mid \boldsymbol{\phi}) \mid \mathbf{D}_{obs}; \boldsymbol{\phi}_t\} = Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}_t) - H(\boldsymbol{\phi} \mid \boldsymbol{\phi}_t)$, where $H(\boldsymbol{\phi} \mid \boldsymbol{\phi}_t) = \mathbb{E}\{\log f(\mathbf{D}_{mis} \mid \mathbf{D}_{obs}; \boldsymbol{\phi}) \mid \mathbf{D}_{obs}; \boldsymbol{\phi}_t\}$ and $Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}_t)$ was defined in Section 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 $\text{BIC}_{H,Q} = -2Q(\hat{\boldsymbol{\phi}} \mid \hat{\boldsymbol{\phi}}) + 2H(\hat{\boldsymbol{\phi}} \mid \hat{\boldsymbol{\phi}}) + pu \log n$. 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 $H(\hat{\boldsymbol{\phi}} \mid \hat{\boldsymbol{\phi}})$, where $\text{BIC}_Q = -2Q(\hat{\boldsymbol{\phi}} \mid \hat{\boldsymbol{\phi}}) + pu \log n$. When the proportion of missing information is small, the use of BIC_Q is adequate.

The information criterion relies on the correct specification of the distribution. Alternatively, we can generalize a bootstrap method for choosing the envelope dimension u, which is more robust to misspecification of distributions. A similar bootstrap method was proposed by Ye and Weiss (2003); Dong and Li (2010) and has been widely used for selecting the dimension of the central space in the dimension reduction literature (Li and Wang, 2007; Yin et al., 2008; Zhu and Zeng, 2006). We propose to first fix the dimension u for the basis matrix $\boldsymbol{\Gamma}$ and then bootstrap data b times to get a sequence of 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 of dimension u^* that contains $\text{span}(\boldsymbol{\Gamma})$, and thus, the estimate should suffer from large variability as compared to the estimate of the original data $\hat{\boldsymbol{\Gamma}}$. Therefore, we choose the largest dimension u^* such that the bootstrap estimated space is the most similar to $\hat{\boldsymbol{\Gamma}}$. To evaluate the variability of

$\hat{\mathbf{\Gamma}}^1, \ldots, \hat{\mathbf{\Gamma}}^b$, we use the *vector correlation coefficient* q^2 proposed by Hotelling (1936). 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

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

Additionally, 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.

4.3 Asymptotics

The following propositions guarantee the efficiency gain and asymptotic normality of the EM envelope estimator. Specifically, Proposition 1 establishes the asymptotic property when the densities of both $\boldsymbol{\varepsilon}$ and $\mathbf{X}$ are correctly specified and that of $\boldsymbol{\varepsilon}$ is normal. Proposition 2 extends the result to the case where the distribution of $\mathbf{X}$ is correctly specified but $\boldsymbol{\varepsilon}$ has a misspecified normal working density. Proposition 3 extends the result further to the case where $\boldsymbol{\varepsilon}$ and $\mathbf{X}$ both have a misspecified normal working density. Let l^* denote the log-likelihood under working model. Let $s_n(\boldsymbol{\phi}) = \nabla l^*(\boldsymbol{\phi})$ and $\mathbf{M}_n(\boldsymbol{\phi}) = -\mathbb{E}\{\nabla^2 l^*(\boldsymbol{\phi})\}$, where ∇ denote the gradient with respect to a general parameter $\boldsymbol{\phi}$. We state our regularity conditions first.

(A1) (Observed likelihood) L_{obs} is unimodal, i.e, the probability distribution has a single maximum, in the parameter space $\mathbf{\Phi}$ with only one point $\boldsymbol{\phi}_0$ such that $\partial Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}_t)/\partial \boldsymbol{\phi}|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} = 0$, and that $\partial Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}_t)/\partial \boldsymbol{\phi}$ is continuous in $\boldsymbol{\phi}$ and $\boldsymbol{\phi}_t$.

(A2) (Finite moments) The error term $\boldsymbol{\varepsilon}_i$ and covariates $\mathbf{X}_i$ have finite $(4+\delta)$-th moment for some $\delta > 0$.

(A3) (Eigenvalues) $\underline{\lim}_n \lambda_-\{n^{-1}\mathrm{Var}(s_n(\boldsymbol{\phi}))\} > 0$ and $\underline{\lim}_n \lambda_-\{n^{-1}\mathbf{M}_n(\boldsymbol{\phi})\} > 0$, where $\underline{\lim}$ and $\lambda_-(\cdot)$ stands for the lower limit and the smallest eigenvalue.

(B1) (Equicontinuous) $\nabla s_n(\boldsymbol{\phi})$ is equicontinuous on any compact subset of $\boldsymbol{\Phi}$.

(B2) (Uniqueness) $\lim_{n\to\infty} \mathbb{E}\{n^{-1}s_n(\boldsymbol{\phi})\} = 0$ has a unique solution at the true parameter value.

Conditions (A1)–(A3), (B1)–(B2) are mild regularity conditions. We proved the following examples in the Appendix that (B1)–(B2) hold when $\mathbf{X}_i$ follows normal or Binomial distribution and the working model for $\boldsymbol{\varepsilon}_i$ is normal.

Example 1. Under Model (1), suppose Assumption 1 holds, if the distribution of $\mathbf{X}_i$ is normal, then regularity conditions (B1)–(B2) hold.

Example 2. Under Model (1), suppose Assumption 1 holds, if $\mathbf{X}_i$ follows Binomial distribution, then regularity conditions (B1)–(B2) hold.

The parameter of the envelope model is $\boldsymbol{\phi} = (\boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\rho})$. We are interested in the property of the parameters $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\rho}$, which are functions of $\boldsymbol{\phi}$. From (2), we have $\mathbf{h}(\boldsymbol{\phi}) = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\rho}) = (\boldsymbol{\Gamma}\boldsymbol{\eta}, \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \boldsymbol{\rho}) = \{\mathbf{h}_1(\boldsymbol{\phi}), \mathbf{h}_2(\boldsymbol{\phi}), \mathbf{h}_3(\boldsymbol{\phi})\}$. Let $\boldsymbol{\theta} = \mathbf{h}(\boldsymbol{\phi})$ denote our parameter of interest, $\hat{\boldsymbol{\theta}}_{em\cdot env}$ and $\hat{\boldsymbol{\theta}}_{em\cdot std}$ denote the EM envelope and the standard EM estimators as the EM sequence converges. The following propositions can be proved using the results in Shapiro (1986).

Proposition 1. Under Model (1), suppose Assumption 1, Conditions 3–4, and (A1) hold, assume the distributions of $\boldsymbol{\varepsilon}_i$ and $\mathbf{X}_i$ are both correctly specified and $\boldsymbol{\varepsilon}_i$ follows a normal distribution, then $\sqrt{n}(\hat{\boldsymbol{\theta}}_{em\cdot std} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_{std})$ and $\sqrt{n}(\hat{\boldsymbol{\theta}}_{em\cdot env} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_{env})$ as

$n \to \infty$, where $\mathbf{V}_{env} = \mathbf{G}(\mathbf{G}^T\mathbf{V}_{std}^{-1}\mathbf{G})^\dagger\mathbf{G}^T$ and $\mathbf{G}$ is given by

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

Matrices $\mathbf{C}_r$ and $\mathbf{E}_u$ are defined in the Appendix. Hence, $\mathbf{V}_{env} - \mathbf{V}_{std} \geq 0$, which indicates the efficiency gain of the EM envelope estimator.

When the envelope dimension $u = r$, the envelope reduces to the standard maximum likelihood estimate. That is, even when the envelope assumptions do not hold, the EM envelope estimator performs as well as the standard EM estimator. Also, following a similar argument as in Cook et al. (2010), if the variability of the immaterial part is relatively large, then the efficiency gain would be substantial.

Propositions 2 and 3 below extend Proposition 1 and provide the asymptotics of missing data envelope estimator when the normality of $\boldsymbol{\varepsilon}_i$ is violated. Lemmas 1–4 provide asymptotics for the standard estimator.

Lemma 1. Under Model (1), suppose Assumption 1 holds, when $\boldsymbol{\varepsilon}_i$ is misspecified to follow a normal distribution, if (A1)–(A2) and (B1)–(B2) hold, then $\hat{\boldsymbol{\theta}}_{em\cdot std} \xrightarrow{p} \boldsymbol{\theta}$ as $n \to \infty$.

Lemma 2. Under Model (1), suppose Assumption 1 holds, when $\boldsymbol{\varepsilon}_i$ is misspecified to follow a normal distribution, if (A1)–(A3) and (B1)–(B2) hold, then $\sqrt{n}(\hat{\boldsymbol{\theta}}_{em\cdot std} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \tilde{\mathbf{V}}_{std})$ as $n \to \infty$, where $\tilde{\mathbf{V}}_{std} = \mathbf{M}_n(\boldsymbol{\theta})^{-1}\mathrm{Var}\{s_n(\boldsymbol{\theta})\}\mathbf{M}_n(\boldsymbol{\theta})^{-1}$.

Proposition 2. Under Model (1), suppose Assumption 1, Conditions 3–4, (A1)–(A3), and (B1)–(B2) hold, if the distribution of $\mathbf{X}_i$ is correctly specified and $\boldsymbol{\varepsilon}_i$ is misspecified to follow a normal distribution, we have $\sqrt{n}(\hat{\boldsymbol{\theta}}_{em\cdot env} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \tilde{\mathbf{V}}_{env})$ as $n \to \infty$, where $\tilde{\mathbf{V}}_{env} = \mathbf{P}_{\mathbf{G}(\mathbf{J})}\tilde{\mathbf{V}}_{std}\mathbf{P}_{\mathbf{G}(\mathbf{J})}^T$, $\mathbf{P}_{\mathbf{G}(\mathbf{J})} = \mathbf{G}(\mathbf{G}^T\mathbf{J}\mathbf{G})^\dagger\mathbf{G}^T\mathbf{J}$, $\mathbf{G}$ is defined in Proposition 1 and the definition of the symmetric matrix $\mathbf{J}$ is given in the Appendix.

Lemma 3. Under Model (1), suppose Assumption 1 holds, when $\boldsymbol{\varepsilon}_i$ and $\mathbf{X}_i$ are misspecified to follow a normal distribution, if (A1)–(A2) hold, $\hat{\boldsymbol{\theta}}_{em\cdot std} \xrightarrow{p} \boldsymbol{\theta}$ as $n \to \infty$.

Lemma 4. Under Model (1), suppose Assumption 1 holds, when $\boldsymbol{\varepsilon}_i$ and $\mathbf{X}_i$ are misspecified to follow a normal distribution, if (A1)–(A3) hold, $\sqrt{n}(\hat{\boldsymbol{\theta}}_{em\cdot std} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \tilde{\mathbf{V}}_{std})$ as $n \to \infty$, where $\tilde{\mathbf{V}}_{std} = \mathbf{M}_n(\boldsymbol{\theta})^{-1}\text{Var}\{s_n(\boldsymbol{\theta})\}\mathbf{M}_n(\boldsymbol{\theta})^{-1}$.

Proposition 3. Under Model (1), suppose Assumption 1, Conditions 3–4, (A1)–(A3) hold, if $\boldsymbol{\varepsilon}_i$ and $\mathbf{X}_i$ are both misspecified to follow a normal distribution, we have $\sqrt{n}(\hat{\boldsymbol{\theta}}_{em\cdot env} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \tilde{\mathbf{V}}_{env})$ as $n \to \infty$, where $\tilde{\mathbf{V}}_{env} = \mathbf{P}_{\mathbf{G}(\mathbf{J})}\tilde{\mathbf{V}}_{std}\mathbf{P}^T_{\mathbf{G}(\mathbf{J})}$, and $\mathbf{P}_{\mathbf{G}(\mathbf{J})} = \mathbf{G}(\mathbf{G}^T\mathbf{J}\mathbf{G})^{\dagger}\mathbf{G}^T\mathbf{J}$.

5 Simulations

5.1 Normal errors

Jia et al. (2010) compared the envelope method with some competitor estimators such as 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, to avoid duplication, we do not consider those competitor estimators here. In this subsection, we compare six different estimators: the EM envelope estimator $\hat{\boldsymbol{\beta}}_{em\cdot env}$, the complete case (CC) envelope estimator $\hat{\boldsymbol{\beta}}_{cc\cdot env}$, the full data envelope $\hat{\boldsymbol{\beta}}_{full\cdot env}$, the standard EM estimator $\hat{\boldsymbol{\beta}}_{em\cdot std}$, the standard complete case (CC) estimator $\hat{\boldsymbol{\beta}}_{cc\cdot std}$, and the full data MLE $\hat{\boldsymbol{\beta}}_{full\cdot std}$. 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. 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 QR decomposition, 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 $\mathbf{N} \in \mathbb{R}^{p\times p}$ where each element is independently from $U(-10, 10)$, 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 $\mathbf{X}_i \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and $\mathbf{Y}_i \mid \mathbf{X}_i \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{\beta}\mathbf{X}, \boldsymbol{\Sigma}_\varepsilon)$ 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 $\text{logit}P(R_{X_{i,4}} = 1 \mid x_{i,1}, x_{i,2}, x_{i,3}) = 1 - x_{i,1} - 2x_{i,2} - 3x_{i,3}$, $\text{logit}P(R_{X_{i,3}} = 1 \mid x_{i,1}, x_{i,4}) = 1 - x_{i,1} - 2x_{i,4}$, and $\text{logit}P(R_{X_{i,5}} = 1 \mid x_{i,1}) = 1 - x_{i,1}$. Also, set five missingness mechanisms for the responses as $\text{logit}P(R_{Y_{i,2}} = 1, R_{Y_{i,4}} = 1 \mid x_{i,1}, y_{i,8}, y_{i,9}) = 2 - x_{i,1} - y_{i,8} - 3y_{i,9}$, $\text{logit}P(R_{Y_{i,3}} = 1 \mid x_{i,2}, y_{i,4}, y_{i,6}) = 1 - x_{i,2} - 3y_{i,4} - y_{i,6}$, $\text{logit}P(R_{Y_{i,7}} = 1, R_{Y_{i,8}} = 1, R_{Y_{i,9}} = 1 \mid y_{i,1}, y_{i,2}, y_{i,3}) = 2 - 2y_{i,1} - y_{i,2} - 3y_{i,3}$, $\text{logit}P(R_{Y_{i,1}} = 1, R_{Y_{i,10}} = 1 \mid x_{i,1}, x_{i,2}) = 1 - x_{i,1} - x_{i,2}$ and $\text{logit}P(R_{Y_{i,5}} = 1, R_{Y_{i,6}} = 1 \mid x_{i,1}, x_{i,2}, y_{i,1}, y_{i,10}) = 1 - x_{i,1} - x_{i,2} - y_{i,1} - y_{i,10}$. 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 $\hat{\boldsymbol{\beta}}_{em\cdot env}$, $\hat{\boldsymbol{\beta}}_{cc\cdot env}$, $\hat{\boldsymbol{\beta}}_{full\cdot env}$, $\hat{\boldsymbol{\beta}}_{em\cdot std}$, $\hat{\boldsymbol{\beta}}_{cc\cdot std}$, and $\hat{\boldsymbol{\beta}}_{full\cdot std}$, where $\hat{\boldsymbol{\beta}}_{em\cdot env}$ is calculated from the EM envelope algorithm using BIC_Q to select the envelope dimension.

Step 5. Repeat Steps 2–4 for 1000 times.

Under the missingness mechanisms above, each predictor suffers from about 10%–15% missingness and each response about 5%–10%. In our simulations, 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. Details about the algorithm are in the Appendix. 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 3 below and Table 1 in the Appendix. For the EM envelope estimator, by using $\mathrm{BIC}_{\mathrm{Q}}$ 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 1. Similar to the illustrative example in Section 2, 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. The complete case data may not have an envelope structure, although in finite sample cases we can usually find one. 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.

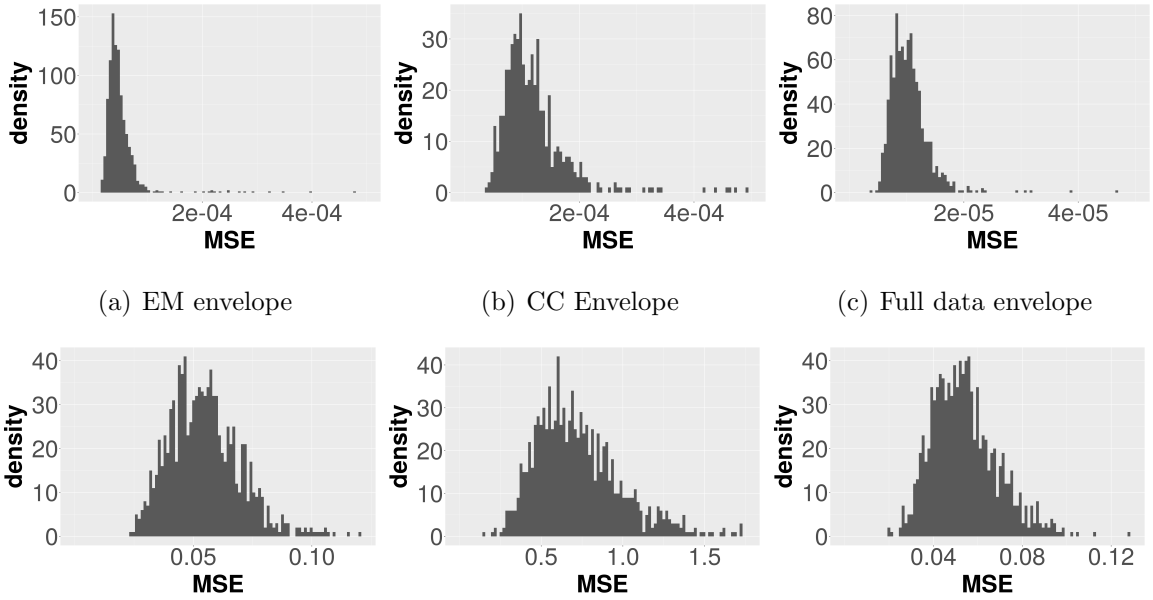
We carried out another simulation study, where the steps were the same as above, except we replaced $\boldsymbol{\Omega}_0 = 1000\mathbf{I}_q$ with $\boldsymbol{\Omega}_0 = 10\mathbf{I}_q$ in Step 2. This is a case where the variance of the immaterial part is not as large. The median MSEs of the EM envelope, 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 8 and Table 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.

5.2 Non-normal errors

In order to investigate the performance of our estimator under the scenario of Propositions 2 and 3, we carried out four additional sets of simulations to compare $\hat{\boldsymbol{\beta}}_{em\cdot env}$ and $\hat{\boldsymbol{\beta}}_{em\cdot std}$ as well as other estimators when the error term $\boldsymbol{\varepsilon}_i$ is not normally distributed. Specifically, we consider two scenarios: (i) Correctly specified the distribution of $\mathbf{X}_i$ and (ii) Misspecified the distribution of $\mathbf{X}_i$. The simulations under scenario (i) are carried out in the following steps.

Step 1*. Set $n = 500$, $r = 10$, $p = 5$, and $u = 2$. Generate parameters $\tilde{\boldsymbol{\Gamma}} \in \mathbb{R}^{r\times u}$, $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{r\times p}$, where the elements are drawed independently from $U(0,1)$ and $U(-10,10)$. By

Figure 3: 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 $\boldsymbol{\Omega}_0 = 1000\mathbf{I}_q$.



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

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

QR decomposition, 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 $\mathbf{N} \in \mathbb{R}^{p\times p}$ where each element is independently from $U(-10, 10)$, and set $\boldsymbol{\Sigma}_x = \mathbf{N}\mathbf{N}^T$.

Step 2*. Generate the full data $(\mathbf{X}_i, \mathbf{Y}_i)$ for each individual i. We generate $\mathbf{X}_{ij} \overset{i.i.d}{\sim} 25\text{Ber}(0.5)$ where $j = 1, \ldots 5$. In order to satisfy the independence conditions $\boldsymbol{\Gamma}_0^T\mathbf{Y}_i \perp\!\!\!\perp \mathbf{X}_i$ and $\boldsymbol{\Gamma}^T\mathbf{Y}_i \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{Y}_i \mid \mathbf{X}_i$, we firstly draw $\boldsymbol{\varepsilon}_{i1} \in \mathbb{R}^u$ and $\boldsymbol{\varepsilon}_{i2} \in \mathbb{R}^{r-u}$ independently from two distributions $t_5(\mathbf{0}, \mathbf{I}_u)$ and $t_5(\mathbf{0}, 1000\mathbf{I}_{r-u})$. Then we set $\boldsymbol{\varepsilon}_i = \boldsymbol{\Gamma}\boldsymbol{\varepsilon}_{i1} + \boldsymbol{\Gamma}_0\boldsymbol{\varepsilon}_{i2}$ and $\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\varepsilon}_i$.

Step 3*. Generate missingness same as Step 3.

Step 4*. Calculate $\hat{\boldsymbol{\beta}}_{em\cdot env}$, $\hat{\boldsymbol{\beta}}_{cc\cdot env}$, $\hat{\boldsymbol{\beta}}_{full\cdot env}$, $\hat{\boldsymbol{\beta}}_{em\cdot std}$, $\hat{\boldsymbol{\beta}}_{cc\cdot std}$, and $\hat{\boldsymbol{\beta}}_{full\cdot std}$. We calculate $\hat{\boldsymbol{\beta}}_{em\cdot std}$ and $\hat{\boldsymbol{\beta}}_{em\cdot env}$ using normal working model for $\boldsymbol{\varepsilon}_i$ and Bernoulli model for $\mathbf{X}_i$ using the parameter updates derived in Example 5. The dimension of the envelope of $\hat{\boldsymbol{\beta}}_{em\cdot env}$,

$\hat{\boldsymbol{\beta}}_{full\cdot env}$ and $\hat{\boldsymbol{\beta}}_{cc\cdot env}$ are obtained through the bootstrap method with 20 iterations.

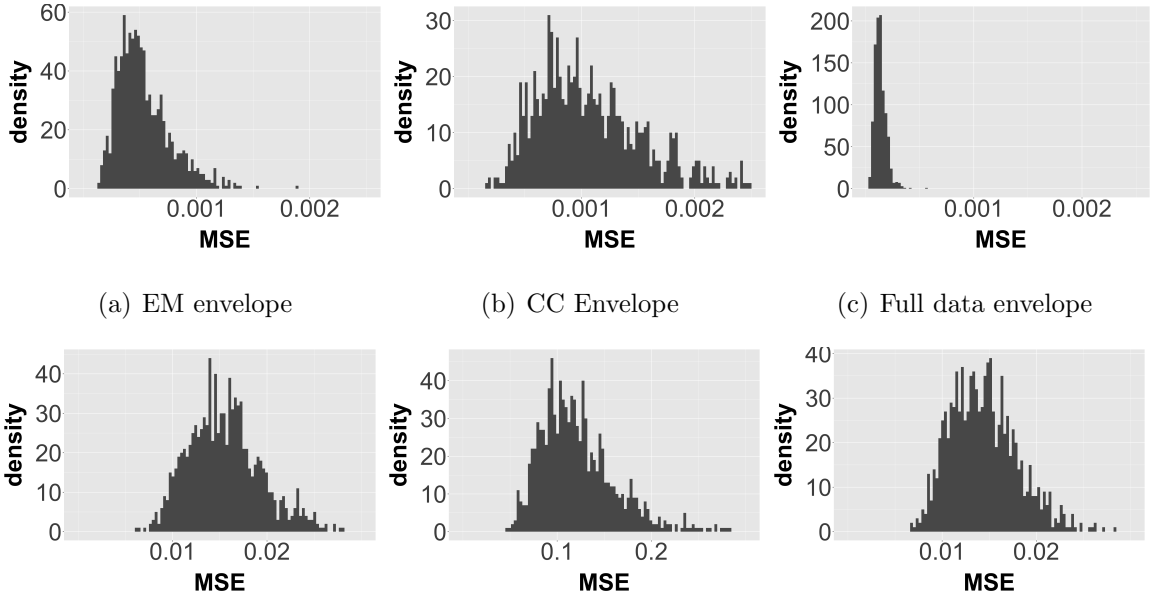
Step 5*. Repeat Steps 2*–4* for 1000 times.

Using the above missingness mechanism, the predictors and responses suffers from about 13% missingness. Although the normality of $\boldsymbol{\varepsilon}_i$ is violated, the data was still generated under a nontrivial envelope structure defined by Conditions 3–4 with the envelope dimension $u = 2$.

We use boostrap to choose the envelope dimensions for $\hat{\boldsymbol{\beta}}_{em\cdot env}$, $\hat{\boldsymbol{\beta}}_{full\cdot env}$ and $\hat{\boldsymbol{\beta}}_{cc\cdot env}$. All the envelope dimensions are correctly specified for $\hat{\boldsymbol{\beta}}_{em\cdot env}$ and $\hat{\boldsymbol{\beta}}_{full\cdot env}$. Following Theorem 2 in Su and Cook (2012) and Proposition 2, once the envelope dimension is correctly specified, the full data envelope with a misspecified working normal density is still consistent although it no longer provides the MLE. As for $\hat{\boldsymbol{\beta}}_{cc\cdot env}$, the correct envelope dimension $u = 2$ is selected 903 out of 1000 times, $u = 3$ is selected 94 times, and it chose $u = 4$ for the rest of 3 times. We observe the bootstrap method requires more computational time than the likelihood method, but is more robust in selecting the envelope dimension. It is worth noticing that for the complete case, even if the envelope dimension is correctly specified for most of the time, the resulting estimator usually suffers from bias. Under current missingness mechanism, the bias for the complete case estimator is relatively small. Therefore, all three envelope estimators have better performances than the standard estimators with full, complete and all data, because the variance of the immaterial part is much larger than that of the material part. The median MSEs are 4.84×10^{-4}, 1.52×10^{-2}, 1.07×10^{-3}, 0.11, 1.28×10^{-4}, and 1.41×10^{-2} for $\hat{\boldsymbol{\beta}}_{em\cdot env}$, $\hat{\boldsymbol{\beta}}_{em\cdot std}$, $\hat{\boldsymbol{\beta}}_{cc\cdot env}$, $\hat{\boldsymbol{\beta}}_{cc\cdot std}$, $\hat{\boldsymbol{\beta}}_{full\cdot env}$, $\hat{\boldsymbol{\beta}}_{full\cdot std}$. Detailed comparisons of the simulation results are given in Figure 4 below and Table 3 in the Appendix. We see that when the error term follows multivariate t distribution, as long as the envelope independence conditions hold, our EM envelope estimator empirically outperforms the standard estimator. Also, the EM envelope outperforms the full data MLE, suggesting that in practice, our method has the potential to recover the efficiency loss from missing data.

The simulation under scenario (ii) is similar to that under scenario (i). In Step 2*, we generate $\mathbf{X}_i \overset{i.i.d}{\sim} t_5(\mathbf{0}, \boldsymbol{\Sigma}_x)$, where $t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represent the multivariate t distribution with

Figure 4: 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 the error term $\boldsymbol{\epsilon}_i$ follows t-distribution and $\mathbf{X}_i$ follows Bernoulli distribution.



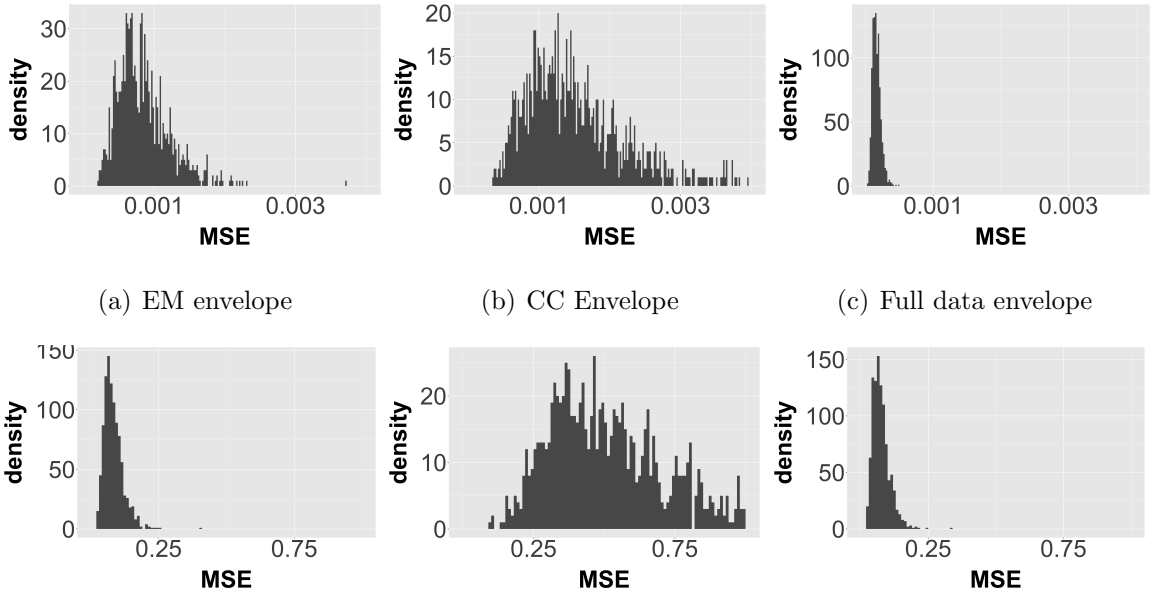
(a) EM envelope (b) CC Envelope (c) Full data envelope

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

location parameter $\boldsymbol{\mu}$, scale parameter $\boldsymbol{\Sigma}$ and degrees of freedom ν, $\boldsymbol{\Sigma}_x = \mathbf{N}\mathbf{N}^T$. and each element of $\mathbf{N}$ is independently from $U(-10, 10)$. In Step 4*, $\hat{\boldsymbol{\beta}}_{em\cdot std}$ and $\hat{\boldsymbol{\beta}}_{em\cdot env}$ are obtained using normal working model for both $\boldsymbol{\varepsilon}_i$ and $\mathbf{X}_i$.

All the envelope dimensions for $\hat{\boldsymbol{\beta}}_{em\cdot env}$ and $\hat{\boldsymbol{\beta}}_{full\cdot env}$ are correctly estimated through the bootstrap method. The dimension for $\hat{\boldsymbol{\beta}}_{cc\cdot env}$ is selected correctly for 90.5% of the time, while the rest 9.5% yields an estimated dimension $u > 3$. All three envelope estimators have better performances than the standard estimators with full, complete and all data because the variation of the immaterial part is much larger than the material part. The median MSEs are 7.96×10^{-4}, 7.61×10^{-2}, 1.38×10^{-3}, 0.50, 1.52×10^{-4}, and 6.96×10^{-2} for $\hat{\boldsymbol{\beta}}_{em\cdot env}$, $\hat{\boldsymbol{\beta}}_{em\cdot std}$, $\hat{\boldsymbol{\beta}}_{cc\cdot env}$, $\hat{\boldsymbol{\beta}}_{cc\cdot std}$, $\hat{\boldsymbol{\beta}}_{full\cdot env}$, $\hat{\boldsymbol{\beta}}_{full\cdot std}$. Detailed comparison of the simulation results are given in Figure 5 below and Table 4 in the Appendix.

Figure 5: 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 the error term $\boldsymbol{\epsilon}_i$ and $\mathbf{X}_i$ follows t-distribution.



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

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

We carried out another two sets of simulations where the data generating steps were the same as above, but we changed the distribution of $\boldsymbol{\varepsilon}_{i1} \in \mathbb{R}^u$ and $\boldsymbol{\varepsilon}_{i2} \in \mathbb{R}^{r-u}$. Firstly, we generate each element of $\boldsymbol{\varepsilon}_{i1}$, $\boldsymbol{\varepsilon}_{i2}$ independently from $U(-1,1)$ and $U(-10,10)$. Under this setting, the median MSEs are 2.82×10^{-4}, 1.59×10^{-3}, 1.37×10^{-3}, 1.00×10^{-2}, 2.14×10^{-4}, 1.45×10^{-3} for $\hat{\boldsymbol{\beta}}_{em\cdot env}$, $\hat{\boldsymbol{\beta}}_{em\cdot std}$, $\hat{\boldsymbol{\beta}}_{cc\cdot env}$, $\hat{\boldsymbol{\beta}}_{cc\cdot std}$, $\hat{\boldsymbol{\beta}}_{full\cdot env}$, $\hat{\boldsymbol{\beta}}_{full\cdot std}$. When each element of $\boldsymbol{\varepsilon}_{i1}$, $\boldsymbol{\varepsilon}_{i2}$ are generated independently from Laplace$(0,1)$ and Laplace$(0,20)$, the median MSEs are 1.45×10^{-3}, 3.75×10^{-2}, 2.92×10^{-3}, 0.246, 3.38×10^{-4} and 3.41×10^{-2} for $\hat{\boldsymbol{\beta}}_{em\cdot env}$, $\hat{\boldsymbol{\beta}}_{em\cdot std}$, $\hat{\boldsymbol{\beta}}_{cc\cdot env}$, $\hat{\boldsymbol{\beta}}_{cc\cdot std}$, $\hat{\boldsymbol{\beta}}_{full\cdot env}$, $\hat{\boldsymbol{\beta}}_{full\cdot std}$. Detailed results are provided in Table 5 and 6 in the Apendix. Under both settings, we see substantial empirical efficiency gains by using our method.

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 CKD patients with a 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 the 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, troponin, 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.

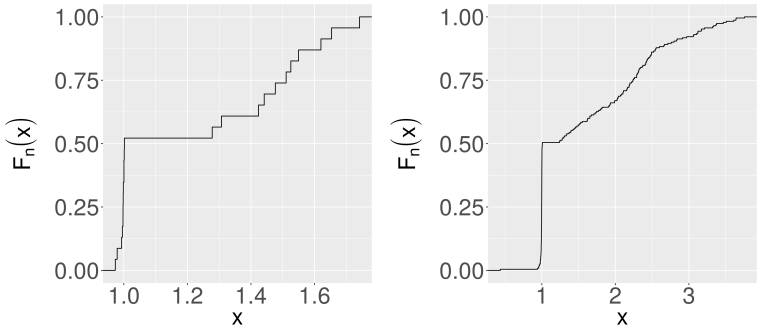
We first assessed the difference in the distributions of baseline biomarkers versus the ESRD status, unadjusted for the established biomarkers. All the biomarkers except the EGFR had some missingness ranging from <1% to 6%. Also, as for the predictors, hemoglobin and BMI had a relatively low missing 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 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 Louis (1982) suggested, the closed form of the asymptotic variance for the standard EM estimator is in general hard 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 6 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 6) using our method than using the standard EM in-

dicating the efficiency gain using our method, which aligns with our theory. 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 aforementioned 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 7 and Table 8 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 7. The point estimates and the standard errors are again in general smaller using our method as compared with using the standard EM. Figure 7 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 6 and Figure 7, 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

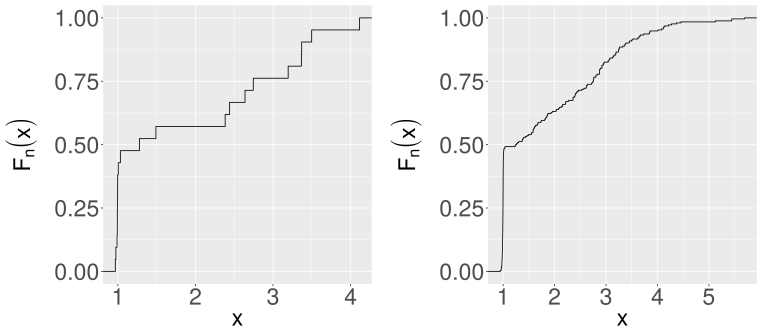
discrepancy between the standard EM and our method: our method found HS CRP, aldosterone, and C-peptide significant which were not shown in standard EM; whereas standard EM found NGAL, which was not found in our method. As our method is more efficient for finite sample, the results of which are more precise than those of the standard EM.

Figure 6: 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 7: 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

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 exists a linear combination of predictors that do 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.

As pointed out by one reviewer, the original envelope formulation uses a decomposition of the variance of the error term. The independence between the material and immaterial part is only guaranteed under normality. The null covariance only guarantees that the information of $\mathbf{\Gamma}_0^T\mathbf{Y}$ is immaterial in the first two moments, rather than all moments which is implied by independency. Motivated by such an observation, we explored alternative ways to guarantee independence in a separate paper (Wang et al., 2020). Specifically, we modified the envelope method by imposing the independence conditions directly and used semiparametric methods to derive the semiparametric efficiency bound. The missing data under this newly defined envelope model can be handled using semiparametric estimating equations (Robins and Rotnitzky, 1995; Robins et al., 1994; Sun et al., 2018; Sun and Liu, 2018). We leave extensions of our missing data estimation methods to semiparametric inference to future research.

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 the 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 the 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.

8 Software

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

A 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 $(\mathbf{Y}_{i,obs}^T, \mathbf{X}_{i,obs}^T)^T$ also follows a normal distribution.

Example 3. Suppose the predictors and responses are normally distributed as $\mathbf{Y}_i|\mathbf{X}_i \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{\beta}\mathbf{X}_i, \boldsymbol{\Sigma})$ and $\mathbf{X}_i \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. Then, $(\mathbf{Y}_{i,obs}^T, \mathbf{X}_{i,obs}^T)^T$ follows a normal distribution $\mathcal{N}(\boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}_i^*)$, 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 3

Note that $\mathbf{Y}_i|\mathbf{X}_i \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{\beta}\mathbf{X}_i, \boldsymbol{\Sigma})$ and $\mathbf{X}_i \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$; hence, $(\mathbf{X}_i^T, \mathbf{Y}_i^T)^T \overset{i.i.d}{\sim} \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, 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 elsewhere, 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, $(\mathbf{X}_{i,obs}^T, \mathbf{Y}_{i,obs}^T, \mathbf{X}_{i,mis}^T, \mathbf{Y}_{i,mis}^T)^T$ follows $\mathcal{N}(\mathbf{B}_i\tilde{\boldsymbol{\mu}}, \mathbf{B}_i\tilde{\boldsymbol{\Sigma}}\mathbf{B}_i^T)$. Therefore, by the property of normal distribution, $(\mathbf{X}_{i,obs}^T, \mathbf{Y}_{i,obs}^T)^T \sim \mathcal{N}(\mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\mu}}, \mathbf{S}_i\mathbf{B}_i\tilde{\boldsymbol{\Sigma}}\mathbf{B}_i^T\mathbf{S}_i^T)$, 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 update 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 $\boldsymbol{\rho}_t$.

Example 4. Under Model (1) and assume $\mathbf{X}_i \overset{i.i.d}{\sim} \mathcal{N}_p(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. 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 $\boldsymbol{\Sigma}_{x,t+1} = \{\mathbf{A}_{3,t} - 2\mathbb{E}(\mathbf{X}_i|\mathbf{D}_{i,obs}; \boldsymbol{\theta}_t)\boldsymbol{\mu}_{x,t+1}\}/n + \boldsymbol{\mu}_{x,t+1}\boldsymbol{\mu}_{x,t+1}^T$.

Derivation of Example 4

The likelihood 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,

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

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 5, 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 normally distributed, following a similar derivation as in the Example 3, 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 $\mathcal{N}\{(\boldsymbol{\mu}_{i,1}^T, \boldsymbol{\mu}_{i,2}^T)^T, \begin{pmatrix} \boldsymbol{\Sigma}_{i1} & \boldsymbol{\Sigma}_{i2} \\ \boldsymbol{\Sigma}_{i2}^T & \boldsymbol{\Sigma}_{i3} \end{pmatrix}\}$, 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 $\mathcal{N}\{\boldsymbol{\mu}_{i,1} + \boldsymbol{\Sigma}_{i2}\boldsymbol{\Sigma}_{i3}^{-1}(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,2}), \boldsymbol{\Sigma}_{i1} - \boldsymbol{\Sigma}_{i2}\boldsymbol{\Sigma}_{i3}^{-1}\boldsymbol{\Sigma}_{i2}^T\}$. Therefore,

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

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

and

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

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

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

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

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

Example 5. Under model (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.

Proof of Example 5

Let $\boldsymbol{\beta}_{i,obs}$ denote the submatrix of $\boldsymbol{\beta}$ where 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 $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{i,obs})$.

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

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

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 $\dfrac{\pi_t q_t}{1-\pi_t+\pi_t q_t}$, 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 $\mathbb{E}(X_i|\mathbf{D}_{i,obs}) = \dfrac{\pi_t q_t}{1-\pi_t+\pi_t q_t}$ if otherwise. Denote $\tilde{\pi}_i = \left(\dfrac{\pi_t q_t}{1-\pi_t+\pi_t q_t}\right)^{1-R_{X_i}} X_i^{R_{X_i}}$, 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 for 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 exists 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, $\mathrm{Var}(\mathbf{y}_{i,mis}^T, \mathbf{y}_{i,obs}^T)^T = \mathbf{B}_i \boldsymbol{\Sigma} \mathbf{B}_i^T = \begin{pmatrix} \boldsymbol{\Sigma}_{i1} & \boldsymbol{\Sigma}_{i2} \\ \boldsymbol{\Sigma}_{i2}^T & \boldsymbol{\Sigma}_{i3} \end{pmatrix}$, where $\boldsymbol{\Sigma}_{i1} = \mathrm{Var}(\mathbf{Y}_{i,mis})$, $\boldsymbol{\Sigma}_{i2} = \mathrm{Cov}(\mathbf{Y}_{i,mis}, \mathbf{Y}_{i,mis})$, and $\boldsymbol{\Sigma}_{i3} = \mathrm{Var}(\mathbf{Y}_{i,obs})$.

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

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

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.

Proof of Example 1

Firstly we prove the case where $\mathbf{X}_i$ follows normal distribution. Since the working model for $\boldsymbol{\varepsilon}_i$ is also normal, the estimator $\hat{\boldsymbol{\theta}}_{obs \cdot std}$ is obtained by maximizing the following observed data likelihood under the working model:

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

From Example 1 and notations therein, we have

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

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 $\hat{\boldsymbol{\theta}}_{obs \cdot std}$ is the solution to the following generalized estimating equation (GEE):

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

where l_i is the log-likelihood of each observation under the working model, and $\psi(\mathbf{D}_{i,obs}, \boldsymbol{\theta}) = \partial l_i / \partial \boldsymbol{\theta}$. During the proof, we are calculating the expectation given the observed data pattern.

Denote $\mathbf{M}_1 = \dfrac{\partial \tilde{\boldsymbol{\mu}}}{\partial \boldsymbol{\mu}_x^T}$, $\mathbf{M}_2 = \dfrac{\partial \tilde{\boldsymbol{\mu}}}{\partial \boldsymbol{\beta}^T}$, $\mathbf{M}_3 = \dfrac{\partial \mathrm{vec}(\tilde{\boldsymbol{\Sigma}})}{\partial \mathrm{vech}(\boldsymbol{\Sigma})^T}$, $\mathbf{M}_4 = \dfrac{\partial \mathrm{vec}(\tilde{\boldsymbol{\Sigma}})}{\partial \boldsymbol{\beta}^T}$, and $\mathbf{M}_5 = \dfrac{\partial \mathrm{vec}(\tilde{\boldsymbol{\Sigma}})}{\partial \mathrm{vech}(\boldsymbol{\Sigma}_x)^T}$. We have

$$\frac{\partial l_i}{\partial \boldsymbol{\mu}_x^T} = (\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T \boldsymbol{\Sigma}_{i,obs}^{-1} \mathbf{S}_i \mathbf{B}_i \mathbf{M}_1,$$

$$\frac{\partial l_i}{\partial \mathrm{vech}(\boldsymbol{\Sigma}_x)^T} = -\frac{1}{2}\mathrm{vec}(\boldsymbol{\Sigma}_{i,obs}^{-1})^T (\mathbf{B}_i^T \mathbf{S}_i^T \otimes \mathbf{B}_i^T \mathbf{S}_i^T)\mathbf{M}_5 + \frac{1}{2}(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T \otimes$$
$$(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T (\boldsymbol{\Sigma}_{i,obs}^{-1} \otimes \boldsymbol{\Sigma}_{i,obs}^{-1})(\mathbf{B}_i^T \mathbf{S}_i^T \otimes \mathbf{B}_i^T \mathbf{S}_i^T)\mathbf{M}_5,$$

$$\frac{\partial l_i}{\partial \mathrm{vech}(\boldsymbol{\Sigma})^T} = -\frac{1}{2}\mathrm{vec}(\boldsymbol{\Sigma}_{i,obs}^{-1})^T (\mathbf{B}_i^T \mathbf{S}_i^T \otimes \mathbf{B}_i^T \mathbf{S}_i^T)\mathbf{M}_3 + \frac{1}{2}(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T \otimes$$
$$(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T (\boldsymbol{\Sigma}_{i,obs}^{-1} \otimes \boldsymbol{\Sigma}_{i,obs}^{-1})(\mathbf{B}_i^T \mathbf{S}_i^T \otimes \mathbf{B}_i^T \mathbf{S}_i^T)\mathbf{M}_3,$$

$$\begin{aligned}\frac{\partial l_i}{\partial \boldsymbol{\beta}^T} &= \frac{\partial l_i}{\partial \mathrm{vec}(\boldsymbol{\Sigma}_{i,obs})^T}\frac{\partial \mathrm{vec}(\boldsymbol{\Sigma}_{i,obs})}{\partial \boldsymbol{\beta}^T} + \frac{\partial l_i}{\partial \boldsymbol{\mu}_{i,obs}^T}\frac{\partial \boldsymbol{\mu}_{i,obs}}{\partial \boldsymbol{\beta}^T} \\ &= -\frac{1}{2}\mathrm{vec}(\boldsymbol{\Sigma}_{i,obs}^{-1})^T (\mathbf{B}_i^T \mathbf{S}_i^T \otimes \mathbf{B}_i^T \mathbf{S}_i^T)\mathbf{M}_4 + \frac{1}{2}(\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T \otimes (\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T \\ &\quad (\boldsymbol{\Sigma}_{i,obs}^{-1} \otimes \boldsymbol{\Sigma}_{i,obs}^{-1})(\mathbf{B}_i^T \mathbf{S}_i^T \otimes \mathbf{B}_i^T \mathbf{S}_i^T)\mathbf{M}_4 + (\mathbf{D}_{i,obs} - \boldsymbol{\mu}_{i,obs})^T \boldsymbol{\Sigma}_{i,obs}^{-1} \mathbf{S}_i \mathbf{B}_i \mathbf{M}_2,\end{aligned}$$

and

$$\psi(\mathbf{D}_{i,obs}, \boldsymbol{\theta}) = \left(\frac{\partial l_i}{\partial \boldsymbol{\mu}_x^T}, \frac{\partial l_i}{\partial \mathrm{vech}(\boldsymbol{\Sigma}_x)^T}, \frac{\partial l_i}{\partial \mathrm{vech}(\boldsymbol{\Sigma})^T}, \frac{\partial l_i}{\partial \boldsymbol{\beta}^T}\right)^T.$$

We need to show (B1) 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 $\psi(\mathbf{D}_{i,obs}, \boldsymbol{\theta})$ is equicontinuous on any compact set of the parameter space.

By taking the derivative of $\psi(\mathbf{D}_{i,obs}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, we will see that $\dfrac{\partial \psi}{\partial \boldsymbol{\theta}}$ is continuous in $\boldsymbol{\theta}$ and $\mathbf{D}_{i,obs}$. Hence, when the parameter space Θ is compact and $\|\mathbf{D}_{i,obs}\| \leq c$, $\dfrac{\partial \psi}{\partial \boldsymbol{\theta}}$ is uniformly bounded. Therefore, $\psi(\mathbf{D}_{i,obs}, \boldsymbol{\theta})$ is equicontinuous. That is, regularity condition (B1) holds.

Next, we prove condition (B2) holds. That is, the solution of

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\mathbb{E}\{\psi(\mathbf{D}_{i,obs},\boldsymbol{\theta})\}=0 \tag{3}$$

is unique at $\boldsymbol{\theta}=\boldsymbol{\theta}_0$. Since we assumed a fixed missing mechanism and $(\mathbf{X}_i,\mathbf{Y}_i)$ is of length $p+r$, there are at most $2^{p+r}-1$ observed data patterns. Let m denote the total number of observed data patterns, $\mathbf{D}^*_{i,obs}$ denote the i-th observed data pattern with probability p^*_i for $i=1,\ldots,m$ satisfying $\sum_{i=1}^m p^*_i=1$. For example, if for the i-th observed data pattern only X_1 is missing, then $\mathbf{D}^*_{i,obs}=(X_2,\ldots,X_p,\mathbf{Y})$. Hence,

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\mathbb{E}\{\psi(\mathbf{D}_{i,obs},\boldsymbol{\theta})\}=\sum_{i=1}^{m}p^*_i\mathbb{E}\{\psi(\mathbf{D}^*_{i,obs},\boldsymbol{\theta})\}. \tag{4}$$

Let $\boldsymbol{\theta}_0=(\boldsymbol{\mu}_{0x},\boldsymbol{\Sigma}_{0x},\boldsymbol{\Sigma}_0,\boldsymbol{\beta}_0)$ denote the true parameter value, $\tilde{\boldsymbol{\mu}}_0=(\boldsymbol{\mu}_{0x}^T,\boldsymbol{\mu}_{0x}^T\boldsymbol{\beta}_0^T)^T$ and $\tilde{\boldsymbol{\Sigma}}_0=\begin{pmatrix}\boldsymbol{\Sigma}_{0x} & \boldsymbol{\Sigma}_{0x}\boldsymbol{\beta}_0\\ \boldsymbol{\beta}_0^T\boldsymbol{\Sigma}_{0x} & \boldsymbol{\Sigma}_0+\boldsymbol{\beta}_0^T\boldsymbol{\Sigma}_{0x}\boldsymbol{\beta}_0\end{pmatrix}$. Let $\mathbf{S}^*_i$, $\mathbf{B}^*_i$ denote the corresponding matrices in Example 1 for the observed data $\mathbf{D}^*_{i,obs}$. Let l^*_i denote the log-likelihood of the i-th observed data pattern under the working model, then

$$\frac{\partial l^*_i}{\partial\tilde{\boldsymbol{\mu}}^T}=(\mathbf{D}^*_{i,obs}-\mathbf{S}^*_i\mathbf{B}^*_i\tilde{\boldsymbol{\mu}})^T(\mathbf{S}^*_i\mathbf{B}^*_i\tilde{\boldsymbol{\Sigma}}\mathbf{B}^{*T}_i\mathbf{S}^{*T}_i)^{-1}\mathbf{S}^*_i\mathbf{B}^*_i$$

By $\mathbb{E}(\mathbf{D}^*_{i,obs})=\mathbf{S}^*_i\mathbf{B}^*_i\tilde{\boldsymbol{\mu}}_0$ and (4), we have

$$\begin{aligned}\sum_{i=1}^{m}p^*_i\mathbb{E}\left(\frac{\partial l^*_i}{\partial\tilde{\boldsymbol{\mu}}^T}\right)&=(\tilde{\boldsymbol{\mu}}_0-\tilde{\boldsymbol{\mu}})^T\sum_{i=1}^{m}p^*_i(\mathbf{S}^*_i\mathbf{B}^*_i)^T(\mathbf{S}^*_i\mathbf{B}^*_i\tilde{\boldsymbol{\Sigma}}\mathbf{B}^{*T}_i\mathbf{S}^{*T}_i)^{-1}\mathbf{S}^*_i\mathbf{B}^*_i\\&=(\tilde{\boldsymbol{\mu}}_0-\tilde{\boldsymbol{\mu}})^T\sum_{i=1}^{m}p^*_i\mathbf{P}_{\mathbf{B}^{*T}_i\mathbf{S}^{*T}_i(\tilde{\boldsymbol{\Sigma}})}\tilde{\boldsymbol{\Sigma}}^{-1}=0,\end{aligned}$$

where $\mathbf{P}_{\mathbf{B}(\boldsymbol{\Sigma})}\equiv\mathbf{B}(\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})^{-1}\mathbf{B}^T\boldsymbol{\Sigma}$ represents the projection onto span$(\mathbf{B})$ relative to $\boldsymbol{\Sigma}$. In order to show the above estimating equation has a unique solution at $\tilde{\boldsymbol{\mu}}=\tilde{\boldsymbol{\mu}}_0$, we only need to show $\sum_{i=1}^m p^*_i\mathbf{P}_{\mathbf{B}^{*T}_i\mathbf{S}^{*T}_i(\tilde{\boldsymbol{\Sigma}})}$ is full rank. Let q^*_i denote the probability of X_i is observed if $i\le p$, and the probability of Y_{i-p} is observed if $i>p$. Then

$$\sum_{i=1}^{m} p_i^* \mathbf{P}_{\mathbf{B}_i^{*T}\mathbf{S}_i^{*T}(\tilde{\mathbf{\Sigma}})} = \sum_{i=1}^{p+r} q_i^* \mathbf{P}_{\mathbf{e}_i(\tilde{\mathbf{\Sigma}})},$$

where $\mathbf{e}_i$ is the vector of length $p+r$ where the i-th index equals 1 and equals 0 otherwise. Since there is no predictor or response with missing rate 100%, $q_i^* > 0$ for all $1 \le i \le p+r$, the above matrix is full rank. That is, $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}_0$ is the unique solution.

Since $\tilde{\boldsymbol{\mu}} = (\boldsymbol{\mu}_{0x}^T, \boldsymbol{\mu}_{0x}^T\boldsymbol{\beta}_0^T)^T$, the solution for $\boldsymbol{\mu}_x$ must be unique and $\boldsymbol{\mu}_x = \boldsymbol{\mu}_{0x}$.

Recall that

$$\mathbb{E}(\mathbf{D}_{i,obs}^*) = \mathbf{S}_i^*\mathbf{B}_i^*\tilde{\boldsymbol{\mu}}_0 = \boldsymbol{\mu}_{0i,obs}^*,$$

and

$$\mathrm{Var}(\mathbf{D}_{i,obs}^*) = \mathbf{S}_i^*\mathbf{B}_i^*\tilde{\mathbf{\Sigma}}_0\mathbf{B}_i^{*T}\mathbf{S}_i^{*T} = \mathbf{\Sigma}_{0i,obs}^*,$$

we have

$$\mathbb{E}\{(\mathbf{D}_{i,obs}^* - \boldsymbol{\mu}_{0i,obs}^*)^T \otimes (\mathbf{D}_{i,obs}^* - \boldsymbol{\mu}_{0i,obs}^*)^T(\mathbf{\Sigma}_{i,obs}^{*-1} \otimes \mathbf{\Sigma}_{i,obs}^{*-1})\} = \mathrm{vec}(\mathbf{\Sigma}_{i,obs}^{*-1}\mathbf{\Sigma}_{0i,obs}^*\mathbf{\Sigma}_{i,obs}^{*-1})^T.$$

Therefore,

$$\begin{aligned}
\frac{\partial l_i^*}{\partial \mathrm{vech}(\tilde{\mathbf{\Sigma}})^T}\Big|_{\tilde{\boldsymbol{\mu}}=\tilde{\boldsymbol{\mu}}_{0x}} &= \frac{1}{2}\{\mathrm{vec}(\mathbf{\Sigma}_{i,obs}^{*-1}\mathbf{\Sigma}_{0i,obs}^*\mathbf{\Sigma}_{i,obs}^{*-1})^T - \mathrm{vec}(\mathbf{\Sigma}_{i,obs}^{*-1})^T\}(\mathbf{B}_i^T\mathbf{S}_i^T \otimes \mathbf{B}_i^T\mathbf{S}_i^T) \\
&= \frac{1}{2}\mathrm{vec}\{\mathbf{\Sigma}_{i,obs}^{*-1}(\mathbf{\Sigma}_{0i,obs}^* - \mathbf{\Sigma}_{i,obs}^*)\mathbf{\Sigma}_{i,obs}^{*-1}\}^T(\mathbf{B}_i^T\mathbf{S}_i^T \otimes \mathbf{B}_i^T\mathbf{S}_i^T) \\
&= \frac{1}{2}\mathrm{vec}(\mathbf{\Sigma}_{0i,obs}^* - \mathbf{\Sigma}_{i,obs}^*)^T(\mathbf{\Sigma}_{i,obs}^{*-1} \otimes \mathbf{\Sigma}_{i,obs}^{*-1})(\mathbf{B}_i^T\mathbf{S}_i^T \otimes \mathbf{B}_i^T\mathbf{S}_i^T) \\
&= \frac{1}{2}\mathrm{vec}(\tilde{\mathbf{\Sigma}}_0 - \tilde{\mathbf{\Sigma}})^T(\mathbf{S}_i^*\mathbf{B}_i^* \otimes \mathbf{S}_i^*\mathbf{B}_i^*)(\mathbf{\Sigma}_{i,obs}^{*-1} \otimes \mathbf{\Sigma}_{i,obs}^{*-1})(\mathbf{B}_i^T\mathbf{S}_i^T \otimes \mathbf{B}_i^T\mathbf{S}_i^T) \\
&= \frac{1}{2}\mathrm{vec}(\tilde{\mathbf{\Sigma}}_0 - \tilde{\mathbf{\Sigma}})^T\{\mathbf{S}_i^*\mathbf{B}_i^*(\mathbf{S}_i^*\mathbf{B}_i^*\tilde{\mathbf{\Sigma}}\mathbf{B}_i^{*T}\mathbf{S}_i^{*T})^{-1}\mathbf{B}_i^T\mathbf{S}_i^T \otimes \mathbf{S}_i^*\mathbf{B}_i^*(\mathbf{S}_i^*\mathbf{B}_i^*\tilde{\mathbf{\Sigma}}\mathbf{B}_i^{*T}\mathbf{S}_i^{*T})^{-1}\mathbf{B}_i^T\mathbf{S}_i^T\} \\
&= \frac{1}{2}\mathrm{vec}(\tilde{\mathbf{\Sigma}}_0 - \tilde{\mathbf{\Sigma}})^T(\mathbf{P}_{\mathbf{B}_i^{*T}\mathbf{S}_i^{*T}(\tilde{\mathbf{\Sigma}})} \otimes \mathbf{P}_{\mathbf{B}_i^{*T}\mathbf{S}_i^{*T}(\tilde{\mathbf{\Sigma}})}).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\sum_{i=1}^{m} p_i^*\mathbb{E}\left(\frac{\partial l_i^*}{\partial \mathrm{vech}(\tilde{\mathbf{\Sigma}})^T}\Big|_{\tilde{\boldsymbol{\mu}}=\tilde{\boldsymbol{\mu}}_0}\right) \\
=&\frac{1}{2}\mathrm{vec}(\tilde{\mathbf{\Sigma}}_0 - \tilde{\mathbf{\Sigma}})^T\sum_{i=1}^{m} p_i^*(\mathbf{P}_{\mathbf{B}_i^{*T}\mathbf{S}_i^{*T}(\tilde{\mathbf{\Sigma}})} \otimes \mathbf{P}_{\mathbf{B}_i^{*T}\mathbf{S}_i^{*T}(\tilde{\mathbf{\Sigma}})}) = 0.
\end{aligned}$$

Similarly, $\sum_{i=1}^{m} p_i^*(\mathbf{P}_{\mathbf{B}_i^{*T}\mathbf{S}_i^{*T}(\tilde{\boldsymbol{\Sigma}})} \otimes \mathbf{P}_{\mathbf{B}_i^{*T}\mathbf{S}_i^{*T}(\tilde{\boldsymbol{\Sigma}})})$ is full rank. Hence, the above equation implies $\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{\Sigma}}_0$ is the unique solution. That is, $\boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_{0x}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$, and $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ are unique solutions.

Therefore, the solution for (3) is unique, so that (B2) holds.

Proof of Example 2

When X_i follows Binomial distribution with m trials and success probability p. Without loss of generality, among the n samples, we let the first n_0 samples to be the case where the covariate X_i is not missing. Then, the observed data likelihood can be written as

$$
\begin{aligned}
L(\boldsymbol{\theta}) = &\prod_{i=1}^{n_0} \int (2\pi)^{-\frac{r}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{y}_i - x_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - x_i\boldsymbol{\beta})\}\binom{m}{x_i} p^{x_i}(1-p)^{m-x_i} d\mathbf{y}_{i,mis} \\
&\cdot \prod_{i=n_0+1}^{n} \int\int (2\pi)^{-\frac{r}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{y}_i - x_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - x_i\boldsymbol{\beta})\} \\
&\quad \cdot \binom{m}{x_i} p^{x_i}(1-p)^{m-x_i} dx_i d\mathbf{y}_{i,mis} \\
= &\prod_{i=1}^{n_0} (2\pi)^{-\frac{r}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{y}_{i,obs} - x_i\boldsymbol{\beta}_{i,obs})^T\boldsymbol{\Sigma}_{i,obs}^{-1}(\mathbf{y}_{i,obs} - x_i\boldsymbol{\beta}_{i,obs})\}\binom{m}{x_i} p^{x_i}(1-p)^{m-x_i} \\
&\cdot \prod_{i=n_0+1}^{n} \sum_{k=0}^{m} (2\pi)^{-\frac{r}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{y}_{i,obs} - k\boldsymbol{\beta}_{i,obs})^T\boldsymbol{\Sigma}_{i,obs}^{-1}(\mathbf{y}_{i,obs} - k\boldsymbol{\beta}_{i,obs})\} \\
&\quad \binom{m}{k} p^k(1-p)^{m-k}.
\end{aligned}
$$

Hence, $L(\boldsymbol{\theta})$ can also be expressed using normal densities:

$$
\begin{aligned}
L(\boldsymbol{\theta}) = &\prod_{i=1}^{n_0} \left\{ \boldsymbol{\phi}(x_i\boldsymbol{\beta}_{i,obs}, \boldsymbol{\Sigma}_{i,obs}) \binom{m}{x_i} p^{x_i}(1-p)^{m-x_i} \right\} \\
&\cdot \prod_{i=n_0+1}^{n} \left\{ \sum_{k=0}^{m} \boldsymbol{\phi}(k\boldsymbol{\beta}_{i,obs}, \boldsymbol{\Sigma}_{i,obs}) \binom{m}{k} p^k(1-p)^{m-k} \right\},
\end{aligned}
$$

which is a Gaussian mixture model. It is easy to show that $\dfrac{\partial \psi_i}{\partial \boldsymbol{\theta}}$ is continuous in $\boldsymbol{\theta}$ using the same technique. Hence, following the same proof procedure, we know that (B1)–(B2) holds when $\mathbf{X}_i$ follows Binomial distribution.

B Proof of Propositions

Proof of Proposition 1

The parameter of the envelope model is $\boldsymbol{\phi} = (\boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\rho})$. A more rigorous notation would be $\boldsymbol{\phi} = \{\text{vec}(\boldsymbol{\eta}), \text{vec}(\boldsymbol{\Gamma}), \text{vech}(\boldsymbol{\Omega}), \text{vech}(\boldsymbol{\Omega}_0), \text{vec}(\boldsymbol{\rho})\}$, where the vectorization operator $\text{vec} : \mathbb{R}^{r\times p} \to \mathbb{R}^{rp}$ stacks the columns of the matrix. Also, for symmetric matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$, we use the "vech" operator: $\mathbb{R}^{r\times r} \to \mathbb{R}^{r(r+1)/2}$, which stacks the unique elements lies on or below the diagonal by column. Following the notations in Henderson and Searle (1979), we let $\mathbf{C}_r \in \mathbb{R}^{r(r+1)/2\times r^2}$ and $\mathbf{E}_r \in \mathbb{R}^{r^2\times r(r+1)/2}$ denote the "contraction" and "expansion" matrices such that $\text{vech}(\mathbf{A}) = \mathbf{C}_r\text{vec}(\mathbf{A})$ and $\text{vec}(\mathbf{A}) = \mathbf{E}_r\text{vech}(\mathbf{A})$ for any symmetric matrix $\mathbf{A}$ of size r.

Recall we let $\boldsymbol{\phi} = (\boldsymbol{\eta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\rho})$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\rho})$ denote the parameters under the envelope model and the standard model. Since regularity condition (A1) holds, by Corollary 1 of Wu (1983), we know $\hat{\boldsymbol{\theta}}_{em\cdot std}$ and $\hat{\boldsymbol{\theta}}_{em\cdot env}$ are the observed MLE.

We can find function $\mathbf{h}$ such that

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

By matrix differentiation, the gradient matrix $\mathbf{G} = \dfrac{\partial \mathbf{h}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$ have the following form

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

Because of the over-parameterization of $\boldsymbol{\theta}$, the gradient matrix $\mathbf{G}$ is not of full rank. By Proposition 3.1 in Shapiro (1986), we have

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

Hence,

$$\mathbf{V}_{std} - \mathbf{V}_{env} = \mathbf{V}_{std}^{\frac{1}{2}}\{\mathbf{I} - \mathbf{V}_{std}^{-\frac{1}{2}}\mathbf{G}(\mathbf{G}^T\mathbf{V}_{std}^{-1}\mathbf{G})^{\dagger}\mathbf{G}^T\mathbf{V}_{std}^{-\frac{1}{2}}\}\mathbf{V}_{std}^{\frac{1}{2}}.$$

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

Proof of Lemma 1

Under Model (1), since condition (A1) holds, $\hat{\boldsymbol{\theta}}_{em\cdot std}$ is the the same as the observed data MLE. Since regularity conditions (A2), (B1)–(B2) hold, by Proposition 5.5 in Shao (2003), $\hat{\boldsymbol{\theta}}_{em\cdot std} \xrightarrow{p} \boldsymbol{\theta}$ as $n \to \infty$.

Proof of Lemma 2

In additional to the conditions in Lemma 1, we also have condition (A3) holds. Hence, by Theorem 5.14, $\sqrt{n}(\hat{\boldsymbol{\theta}}_{em\cdot std}-\boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \tilde{\mathbf{V}}_{std})$ as $n \to \infty$, where $\tilde{\mathbf{V}}_{std} = \mathbf{M}_n(\boldsymbol{\theta})^{-1}\text{Var}\{s_n(\boldsymbol{\theta})\}\mathbf{M}_n(\boldsymbol{\theta})^{-1}$.

Proof of Proposition 2

From Lemma 1 and 2, we know that $\hat{\boldsymbol{\theta}}_{em\cdot std}$ is consistent and asymptotically normal. Then, we can use Proposition 4.1 in (Shapiro, 1986) to prove this proposition.

Shapiro's $\boldsymbol{\xi}$ in our context is $\boldsymbol{\xi} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\rho})$. Following the proof in Su and Cook (2012), we give the minimum discrepancy function as $f_{MDF} = l_{max} - l$, where l is the logarithm of the misspecified likelihood function 2, and l_{max} is obtained by substituting $\hat{\boldsymbol{\theta}}_{em\cdot std}$ for $\boldsymbol{\theta}$ in 2. There must be one-to-one functions f_1 from $\boldsymbol{\theta}$ to $\boldsymbol{\xi}$ and f_2 from $\hat{\boldsymbol{\theta}}_{em\cdot std}$ to $\mathbf{x}$ so that $\boldsymbol{\xi} = f_1(\boldsymbol{\theta})$ and $\mathbf{x} = f_2(\hat{\boldsymbol{\theta}}_{em\cdot std})$. As f_{MDF} is constructed by the normal likelihood, it satisfies the four conditions required by Shapiro (1986). Let $\mathbf{J} = \frac{1}{2}\frac{\partial^2 f_{MDF}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$. Then, because $\hat{\boldsymbol{\theta}}_{em\cdot std}$ is

obtained by minimizing f_{MDF}, by Proposition 4.1 of Shapiro (1986), we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{em\cdot env} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \tilde{\mathbf{V}}_{env}),$$

where $\tilde{\mathbf{V}}_{env} = \mathbf{G}(\mathbf{G}^T\mathbf{J}\mathbf{G})^\dagger\mathbf{G}^T\mathbf{J}\mathbf{V}_{std}\mathbf{J}\mathbf{G}(\mathbf{G}^T\mathbf{J}\mathbf{G})^\dagger\mathbf{G}^T$.

Proof of Lemma 3

We use Proposition 5.5 in Shao (2003) to prove consistency. In the proof of Example 4, we showed the regularity conditions (B1)–(B2) hold when $\mathbf{X}_i$ is modeled using a normal distribution.

Moreover, since both $\boldsymbol{\varepsilon}_i$ and $\mathbf{X}_i$ have finite $(4+\delta)$-th moment from Condition (A2), we have $\mathbb{E}\{\sup_{\theta\in\Theta} \|\psi_i(\mathbf{D}_{i,obs}, \boldsymbol{\theta})\|\}^2 < \infty$, and $\mathbb{E}\|\mathbf{D}_{i,obs}\| < \infty$. Therefore, the conditions of Lemma 5.3 in Shao (2003) holds. Since the observed data MLE $\hat{\boldsymbol{\theta}}_{obs\cdot std}$ is always $\mathcal{O}(1)$, by Proposition 5.5 in Shao (2003), $\hat{\boldsymbol{\theta}}_{obs\cdot std} \xrightarrow{p} \boldsymbol{\theta}_0$ as $n \to \infty$.

Proof of Lemma 4

In order to prove the asymptotic normality of $\hat{\boldsymbol{\theta}}_{em\cdot std}$, we only need to show $\sqrt{n}(\hat{\boldsymbol{\theta}}_{obs\cdot std} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \tilde{\mathbf{V}}_{std})$ because of condition (A1). We prove that 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, by condition (A3), $\liminf_n \lambda_-\{n^{-1}\mathrm{Var}(s_n(\boldsymbol{\theta}))\} > 0$ and $\liminf_n \lambda_-\{n^{-1}\mathbf{M}_n(\boldsymbol{\theta})\} > 0$ holds. Therefore,

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

Proof of Proposition 3

From Lemma 3–4, we know the standard estimator $\hat{\boldsymbol{\theta}}_{em\cdot std}$ is consistent and asymptotical normal under the normal working model. Hence, the proof of Proposition 3 is the same as

the proof of Proposition 2. We omit the proof here.

C Lemma and algorithms

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

Lemma 5. 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 1: 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 2: 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 1 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.

D Additional tables and figures

References

Anderson, A., Yang, W., Townsend, R., Pan, Q., Chertow, G., Kusek, J., Charleston, J., He, J., Kallem, R., Lash, J., 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.

Table 1: Summary of MSE when $\boldsymbol{\varepsilon}_i$ and $\mathbf{X}_i$ are correctly specified using a normal distribution and $\boldsymbol{\Omega}_0 = 1000\mathbf{I}_q$

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

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

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

Bansal, N., Keane, M., Delafontaine, P., Dries, D., Foster, E., Gadegbeku, C., Go, A., Hamm, L., Kusek, J., Ojo, A., 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.

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59:3–54.

Table 3: Summary of MSE when $\boldsymbol{\varepsilon}_i$ follows t-distribution and $\mathbf{X}_i$ follows Bernoulli distribution

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
$\hat{\boldsymbol{\beta}}_{em\cdot env}$	1.39e-04	3.64e-04	4.84e-04	5.32e-04	6.60e-04	1.90e-03
$\hat{\boldsymbol{\beta}}_{cc\cdot env}$	1.66e-04	7.42e-04	1.07e-03	6.11e-03	1.54e-03	0.236
$\hat{\boldsymbol{\beta}}_{full\cdot env}$	2.89e-05	9.80e-05	1.28e-04	1.36e-04	1.64e-04	5.50e-04
$\hat{\boldsymbol{\beta}}_{em\cdot std}$	6.21e-03	1.27e-02	1.52e-02	1.56e-02	1.77e-02	3.61e-02
$\hat{\boldsymbol{\beta}}_{cc\cdot std}$	4.80e-02	9.32e-02	0.115	0.123	0.143	0.518
$\hat{\boldsymbol{\beta}}_{full\cdot std}$	6.60e-03	1.17e-02	1.41e-02	1.44e-02	1.66e-02	3.26e-02

Table 4: Summary of MSE when $\boldsymbol{\varepsilon}_i$ and $\mathbf{X}$ follows t-distribution

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
$\hat{\boldsymbol{\beta}}_{em\cdot env}$	2.14e-04	6.00e-04	7.96e-04	8.50e-04	1.04e-03	3.72e-03
$\hat{\boldsymbol{\beta}}_{cc\cdot env}$	3.48e-04	9.93e-04	1.38e-03	1.53e-03	1.89e-03	5.67e-03
$\hat{\boldsymbol{\beta}}_{full\cdot env}$	3.41e-05	1.17e-04	1.52e-04	1.62e-04	1.96e-04	4.98e-04
$\hat{\boldsymbol{\beta}}_{em\cdot std}$	2.36e-02	5.79e-02	7.61e-02	8.29e-02	0.101	0.407
$\hat{\boldsymbol{\beta}}_{cc\cdot std}$	9.37e-02	0.363	0.500	0.567	0.683	3.70
$\hat{\boldsymbol{\beta}}_{full\cdot std}$	2.11e-02	5.24e-02	6.96e-02	7.56e-02	9.10e-02	0.338

Table 5: Summary of MSE when $\boldsymbol{\varepsilon}_i$ follows uniform distribution and $\mathbf{X}_i$ follows t-distribution

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
$\hat{\boldsymbol{\beta}}_{em\cdot env}$	7.05e-05	2.14e-04	2.82e-04	3.00e-04	3.61e-03	1.00e-03
$\hat{\boldsymbol{\beta}}_{cc\cdot env}$	1.70e-04	9.89e-04	1.37e-03	1.54e-03	1.93e-03	6.53e-03
$\hat{\boldsymbol{\beta}}_{full\cdot env}$	5.34e-05	1.59e-04	2.13e-04	2.29e-04	2.83e-04	7.99e-04
$\hat{\boldsymbol{\beta}}_{em\cdot std}$	4.22e-04	1.24e-03	1.59e-03	1.68e-03	2.06e-03	4.81e-03
$\hat{\boldsymbol{\beta}}_{cc\cdot std}$	2.27e-03	7.64e-03	1.00e-02	1.11e-02	1.34e-02	4.45e-02
$\hat{\boldsymbol{\beta}}_{full\cdot std}$	4.48e-04	1.14e-03	1.45e-03	1.53e-03	1.84e-03	4.12e-03

Table 6: Summary of MSE when $\boldsymbol{\varepsilon}_i$ follows Laplacian distribution and $\mathbf{X}_i$ follows t-distribution

	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
$\hat{\boldsymbol{\beta}}_{em\cdot env}$	3.59e-04	1.10e-03	1.45e-03	1.57e-03	1.94e-03	5.85e-03
$\hat{\boldsymbol{\beta}}_{cc\cdot env}$	5.40e-04	2.16e-03	2.92e-03	3.20e-03	3.98e-03	1.09e-02
$\hat{\boldsymbol{\beta}}_{full\cdot env}$	9.61e-05	2.57e-04	3.38e-04	3.56e-04	4.40e-04	9.81e-04
$\hat{\boldsymbol{\beta}}_{em\cdot std}$	7.41e-03	2.92e-02	3.75e-02	4.07e-02	4.97e-02	0.101
$\hat{\boldsymbol{\beta}}_{cc\cdot std}$	5.33e-03	0.179	0.246	0.274	0.340	0.908
$\hat{\boldsymbol{\beta}}_{full\cdot std}$	9.38e-03	2.74e-02	3.41e-02	3.71e-02	4.56e-02	9.87e-02

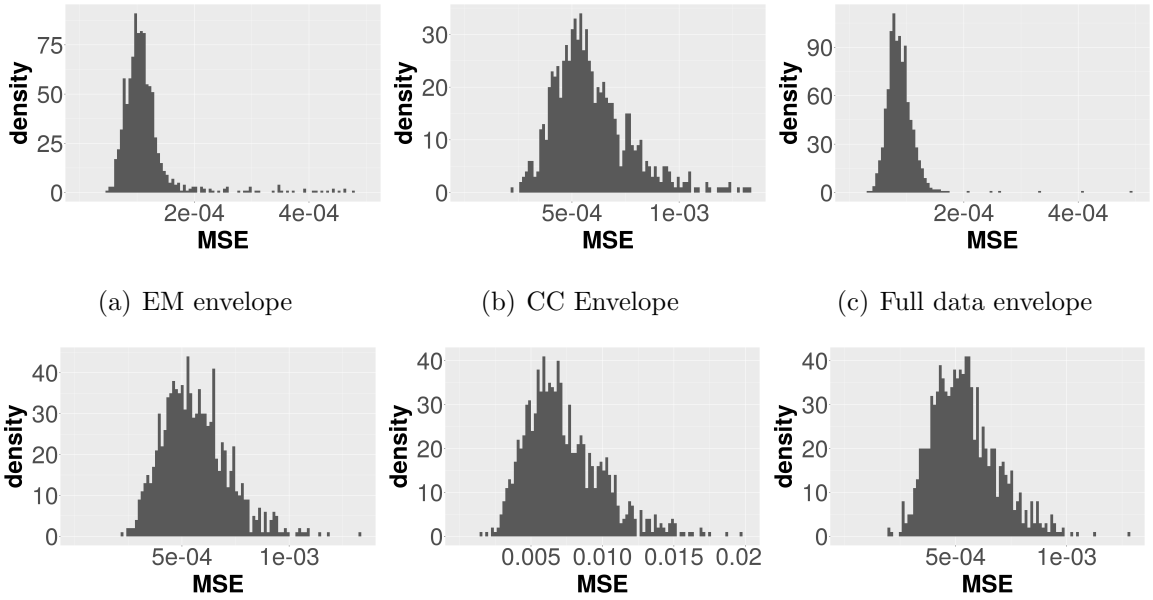
Table 7: The point estimates, bootstrap standard errors, confidence intervals and $p-$values for the difference among patients with and without ESRD on biomarkers adjusted for the established biomarkers

	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 8: 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 8: 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

Budoff, M., Rader, D., Reilly, M., Mohler, E., Lash, J., Yang, W., Rosen, L., Glenn, M., Teal, V., and Feldman, H. (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.

Capuano, V., Bambacaro, A., D'Arminio, T., Vecchio, G., and Cappuccio, L. (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.

Chen, Q., Ibrahim, J. G., Chen, M.-H., and Senchaudhuri, P. (2008). Theory and inference for regression models with missing responses and covariates. *Journal of multivariate analysis*, 99:1302–1331.

Cook, R. D. (2018). Principal Components, Sufficient Dimension Reduction, and Envelopes. *Annual Review of Statistics and Its Application*, 5:533–559.

Cook, R. D., Forzani, L., and Su, Z. (2016). A note on fast envelope estimation. *Journal of Multivariate Analysis*, 150:42–54.

Cook, R. D., Forzani, L., and Zhang, X. (2015). Envelopes and reduced-rank regression. *Biometrika*, 102:439–456.

Cook, R. D., Helland, I., and Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:851–877.

Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, 20:927–960.

Cook, R. D. and Su, Z. (2013). Scaled envelopes: scale-invariant and efficient estimation in multivariate linear regression. *Biometrika*, 100:939–954.

Cook, R. D. and Zhang, X. (2015a). Foundations for envelope models and methods. *Journal of the American Statistical Association*, 110:599–611.

Cook, R. D. and Zhang, X. (2015b). Simultaneous envelopes for multivariate linear regression. *Technometrics*, 57:11–25.

Cook, R. D. and Zhang, X. (2016). Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics*, 25:284–300.

Cook, R. D. and Zhang, X. (2018). Fast envelope algorithms. *Statistica Sinica*, 28:1179–1197.

Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika*, 97:279–294.

Eck, D. J. and Cook, R. D. (2017). Weighted envelope estimation to handle variability in model selection. *Biometrika*, 104:743–749.

Feldman, H., Appel, L., Chertow, G., Cifelli, D., Cizman, B., Daugirdas, J., Fink, J., Franklin-Becker, E., Go, A., Hamm, L., 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., Matthews, G., Townsend, R., Raj, D., Kanetsky, P., Budoff, M., Fischer, M., Rosas, S., Kanthety, R., Rahman, M., 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.

Foster, M. C., Coresh, J., Bonventre, J. V., Sabbisetti, V. S., Waikar, S. S., Mifflin, T. E., Nelson, R. G., Grams, M., Feldman, H. I., Vasan, R. S., 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.

He, J., Reilly, M., Yang, W., Chen, J., Go, A., Lash, J., Rahman, M., DeFilippi, C., Gadegbeku, C., Kanthety, R., 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.

Henderson, H. V. and Searle, S. (1979). Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics. *Canadian Journal of Statistics*, 7:65–81.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28:321–377.

Hristache, M. and Patilea, V. (2017). Conditional moment models with data missing at random. *Biometrika*, 104:735–742.

Ibrahim, J. G., Zhu, H., and Tang, N. (2008). Model selection criteria for missing-data problems using the em algorithm. *Journal of the American Statistical Association*, 103:1648–1658.

Inker, L. A., Coresh, J., Sang, Y., Hsu, C.-y., Foster, M. C., Eckfeldt, J. H., Karger, A. B., Nelson, R. G., Liu, X., Sarnak, M., 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., Benjamini, Y., Lim, C., Raskutti, G., and Yu, B. (2010). Envelope models for parsimonious and efficient multivariate linear regression comment.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102:997–1008.

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112:1131–1146.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, 87:1227–1237.

Little, R. J. and Rubin, D. B. (2014). *Statistical Analysis with Missing Data*, volume 333. John Wiley & Sons.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44:226–233.

Madjid, M. and Fatemi, O. (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.

Meng, X.-L. and Rubin, D. B. (1991). Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86:899–909.

Park, M., Hsu, C.-Y., Go, A. S., Feldman, H. I., Xie, D., Zhang, X., Mifflin, T., Waikar, S. S., Sabbisetti, V. S., Bonventre, J. V., 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.

Robins, J. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90:122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89:846–866.

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., Ma, L., and Liu, L. (2020). Mixed effects envelope models. *Stat*.

Su, Z. and Cook, R. D. (2012). Inner envelopes: efficient estimation in multivariate linear regression. *Biometrika*, 99:687–702.

Su, Z., Zhu, G., Chen, X., and Yang, Y. (2016). Sparse envelope model: efficient estimation and response variable selection in multivariate linear regression. *Biometrika*, 103:579–593.

Sun, B., Liu, L., Miao, W.and Wirth, K., Robins, J., and Tchetgen Tchetgen, E. (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28:1965–1983.

Sun, Z. and Liu, L. (2018+). Semiparametric inference with missing not at random confounders. *Statistica Sinica, in press*.

Wang, J., Chen, H., and Liu, L. (2020). Semiparametric envelope-based partial least square. *in preparation*.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, 11:95–103.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98:968–979.

Yin, X., Li, B., and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99(8):1733–1757.

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, 101(476):1638–1651.