ENVELOPE MODEL FOR MULTIVARIATE LINEAR REGRESSION WITH ELLIPTICAL ERROR

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fullfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

---

by
Gunes Alkan
August 2021

Examining Committee Members:

Dr. Yuexiao Dong, Dissertation Advisor, Department of Statistical Science
Dr. Pallavi Chitturi, Dissertation Examination Chair, Department of Statistical Science
Dr. Kuang-Yao Lee, Department of Statistical Science
Dr. Cencheng Shen, External Member, University of Deleware

ABSTRACT

ENVELOPE MODEL FOR MULTIVARIATE LINEAR REGRESSION WITH ELLIPTICAL ERROR

Gunes Alkan

DOCTOR OF PHILOSOPHY

Temple University, August 2021

Dr. Yuexiao Dong, Advisory Chair

In recent years, the need for models which can accommodate higher order covariates have increased greatly. We first consider linear regression with vector-valued response Y and tensor-valued predictors X. Envelope models (Cook et al., 2010) can significantly improve the estimation efficiency of the regression coefficients by linking the regression mean with the covariance of the regression error. Most existing tensor regression models assume that the conditional distribution of Y given X follows a normal distribution, which may be violated in practice. In Chapter 2, we propose an envelope multivariate linear regression model with tensor-valued predictors and elliptically contoured error distributions. The proposed estimator is more robust to violations of the error normality assumption, and it is more efficient than the estimators without considering the underlying envelope structure. We compare the new proposal with existing estimators in extensive simulation studies. In Chapter 3, we explore how the missing data problem can be addressed for multivariate linear regression setting with envelopes and elliptical error. A popular and efficient approach, multiple imputation is implemented with bootstrapped expectation-maximization (EM) algorithm to fill the missing data, which is then followed with an adjustment in estimating regression coefficients. Simulations with synthetic data as well as real data are presented to establish the superiority of the adjusted multiple imputation method proposed.

DEDICATION

Beni her zorlukta destekleyen, beni kendime inandıracak kadar bana inanan annem, Emine Çorlu Alkan'a.

Her şeyin en güzelini haketmene rağmen, hayat belki sana hakettiğin tüm şansları tanımadı. Ancak ben tüm şanslarımı seninle paylasmaya hazırım. Annem, uzun yıllarımı verdiğim, tum birikimimi, çabamı, terimi ve göz yaşımı döktüğüm bu doktora tezini yalnız ve yalnızca sana ithaf ediyorum.

"Kaşım gözümden çok içim bir parçan
Annem sen benim yanıma kalansın."

ACKNOWLEDGMENTS

First and foremost, I'd like to thank my advisor, Dr. Yuexiao Dong for his guidance and support throughout the last two years. I am very thankful to have had the chance to work with and learn from such an intelligent and inspiring researcher. I'd like to also give special thanks to Dr. Pallavi Chitturi, who from my first year in the program believed in me, encouraged me and has always been extremely kind and understanding towards me. She has contributed to my overall growth much more than words can express.

I would like to also express gratitude to my committee members, Dr. Kuang-Yao Lee and Dr. Cencheng Shen for their valuable time and comments. Moreover, to my dearest friends, Abdul, Shinjini, Nahid, Mike and all my fellow graduate students, without your presence, life at Temple would not have been the same. I cannot wait to see you on the other side.

Furthermore, I want to thank Dr. Gulnihal Meral who helped me gain the confidence that brought me to another country to pursue my graduate studies. I learned that I was awarded with a sponsorship, while I was a student of hers. Upon hearing the news, I immediately called her and she told me to go. Without her encouragement to accept it, I am not sure if I would have pursued education in the United States. I deeply appreciate her support and wisdom that she very generously shared in the short time we worked together in Zonguldak.

I also would like to thank Dr. Yusuf Kaya who was a great and inspiring professor during my undergraduate education. I'd like to acknowledge his support. He helped me more than he probably knows.

To my friends Nese and Fatma, I'd like to thank you both for your unconditional support, endless hours of patient listening, and Zoom study sessions. Sizsiz yapamazdim!

To my family who allowed me to explore and cave my way and find new worlds.

To my brother, Murat, thank you for always being on my side, believing in me and being my biggest cheer leader. Knowing that you will always be on the other side of the phone no matter what happened has been the biggest source of comfort. Ikinci annem, en iyi arkadasim, teyzem Suzan'a, butun bunlar olurken benim yanibasimda oldugunu bilmek, bu uzun doktora yolunu bir nebze de olsa kolaylastirdi. Sonsuz minnettarim. Ayrica babam Ibrahim Alkan, teyzelerim Serife ve Fatos, kuzenim Ezgi ve tum aileme bu zorlu surecte beni yureklendirdikleri, bana inandiklari, benim icin dualar ettikleri icin ictenlikle tesekkur ediyorum. Ayrica, evden her ayrilisimda bana cok calismami soyleyen anneannemi ve degerli sozlerini gittigim her yere kalbimde tasiyor ve onu sonsuz ozlemle, minnetle hatirliyorum. (To my second mom, my best friend, my aunt Suzan, having you by my side made life during Ph.D. much easier. Also, I'd like to thank my father Ibrahim Alkan, my aunts Serife and Fatma, my cousin Ezgi and all other family members, who encouraged me, prayed for me, and stood by my side all along.)

Finally, I'd like to thank my life partner, my husband, my love Eric. Thank you for allowing me to be myself even when I struggle. Thank you for being by my side, literally, whenever I was scared and was dramatic enough to think that I was having a heart attack!!! This program was long and challenging, and having you made each step a bit more bearable. Thank you for trusting me Eric, thank you for turning this country into a home for me, and thank you for making me a better person. Your intelligence, love, and curiosity astonishes me every day and makes me excited about the days to come. You mean so much to me!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

AN OVERVIEW OF ENVELOPES, TENSORS AND ELLIPTICAL DISTRIBUTION FAMILY

To explain what motivated us to work on the proposed method in this dissertation, we explore some of the current theory. As the focus of our new method in Chapter 2 is on higher dimensional variables, also known as tensors, we will first explain some tensor operators and examine existing tensor regression models. Then, we will provide a thorough examination of the envelope models, and explain how they were developed to reduce dimension. We will also briefly discuss envelope estimation. Finally, as our goal is to eliminate normality assumption in regression setting further and to allow error term to have a more generalized distribution, namely elliptically contoured distribution, we will study this distribution family along with its properties.

1.1 Tensor Decompositions, Applications and Regression

In order to have a deep understanding of multivariate regression with tensor predictors, we first need to look into the notation and some results of tensor operations. A tensor is a multidimensional array. Throughout this section, multidimensional array, n-way array and tensor will be used interchangeably.

A $(p_1 \times p_2 \times \cdots \times p_N)$ tensor over $\mathbb{R}$ is denoted as $\mathcal{A}$ with entries $\mathcal{A}_{i_1,i_2,\ldots,i_N}$ or $\mathcal{A}_{i_1 i_2 \ldots i_N}$ where $i_j \in \{1, 2, \ldots, p_j\}$ and $j = 1, 2, \ldots N$. Number of dimensions (ways, modes) is also known as the "order" of that tensor. For example, $\mathcal{A}_{i_1,i_2,\ldots,i_N}$ is a N-way tensor or a tensor of order N. A one-way tensor ($N = 1$) is a vector, and it is often denoted by lowercase letters, e.g., $a \in \mathbb{R}^{p_1}$. A two-way tensor ($N = 2$) is a matrix, and it is denoted by uppercase letters, e.g., $A \in \mathbb{R}^{p_1 \times p_2}$.

a_i represents i^{th} element of a vector a, a_{ij} represents $(i, j)^{th}$ element of a matrix A, and a_{ijk} represents $(i, j, k)^{th}$ element of a (three-way) tensor $\mathcal{A}$.

In order to indicate all entries/ elements on a mode, we use colon. That is, $a_{:j}$ denotes all elements of the first mode/ order of a matrix A, that is all rows and the j^{th} column, whereas $a_{i:}$ represents i^{th} row of the matrix. Also, a vector a_j can be a shorter alternative representation of the j^{th} column of a matrix.

Furthermore, we use "fibers" when representing all elements in only one mode and fixing all others of a tensor. For instance, a three-way tensor A_{ijk} has three different fibers: column fiber denoted by $a_{:jk}$, row fiber denoted by $a_{i:k}$ and tube fiber denoted by $a_{ij:}$. Following are figures of fibers from Kolda and Bader (2009).

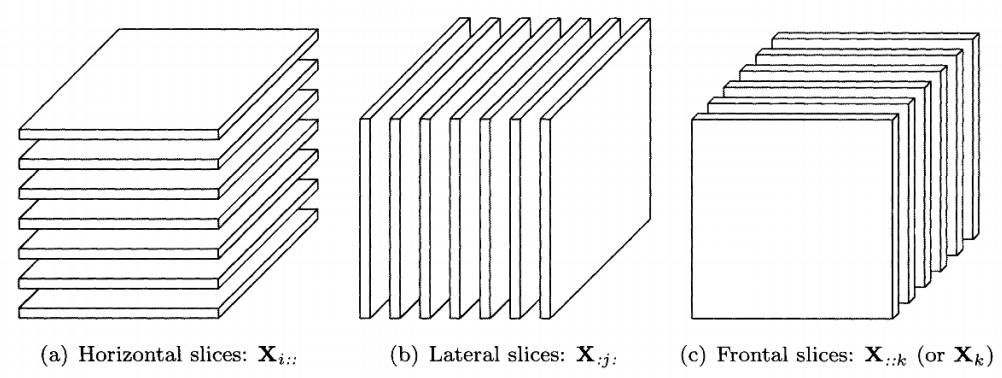

Figure 1.1: Slices of a three-way tensor X_{ijk}

On the other hand, when focusing on two-dimensional parts of a tensor $\mathcal{A}$; that is, when representing all elements in only two modes and fixing all others, we work with "slices". Horizontal slices of a three-way tensor $\mathcal{A}_{ijk}$ is denoted by $\mathcal{A}_{i::}$, lateral slices of a three-way tensor $\mathcal{A}_{ijk}$ is denoted by $\mathcal{A}_{:j:}$, and frontal slices of a three-way tensor $\mathcal{A}_{ijk}$ is denoted by $\mathcal{A}_{::k}$. Below are figures of slices from Kolda and Bader (2009) for a three-way array.

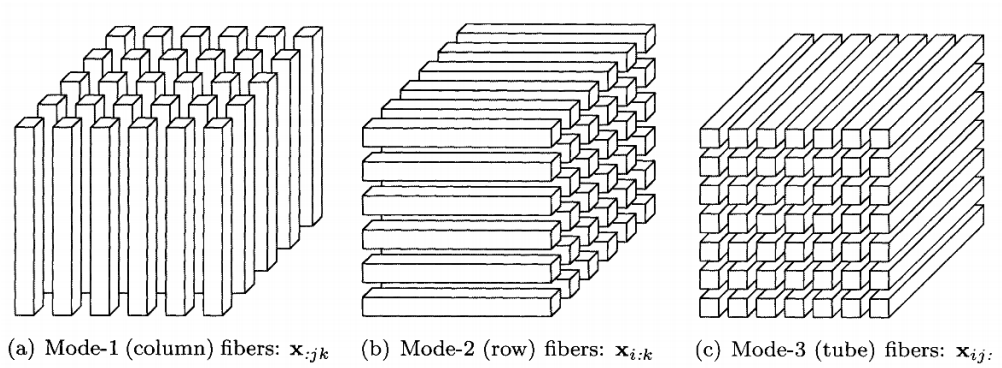

Figure 1.2: Slices of a three-way tensor X_{ijk}

1.1.1 An Overview of Standard Tensor Operations

Let $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_N}$ be N^{th} order tensors (N-mode arrays). Following are some important operations/ definitions.

- The *inner product* of $\mathcal{A}$ and $\mathcal{B}$ is the sum of the products of each element of $\mathcal{A}$ and $\mathcal{B}$:

$$\mathcal{A} \cdot \mathcal{B} = \langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i_1=1}^{p_1} \sum_{i_2=1}^{p_2} \cdots \sum_{i_1=N}^{p_N} a_{i_1 i_2 \ldots i_N} b_{i_1 i_2 \ldots i_N}.$$

- *Matricization* is the process of unfolding elements of a tensor into a matrix. More specifically, *mode-n* matricization of a tensor $\mathcal{A} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_N}$, denoted by $A_{(n)}$, is rearranged into a matrix in a way such that mode-n fibers are the columns of the created matrix.

  More formally, the $(i_1 i_2 \ldots i_N)^{th}$ element of the $\mathcal{A}$ is mapped into $(i_n, j)^{th}$ element of the $A_{(n)}$ such that

$$j = 1 + \sum_{\substack{k=1 \\ k \neq n}}^{N} (i_k - 1) J_k \quad \text{with} \quad J_k = \prod_{\substack{m=1 \\ m \neq n}}^{k-1} I_m$$

  where $A_{(n)} \in \mathbb{R}^{p_n \times \prod_{\substack{j=1 \\ j \neq n}}^{N} p_j}$ with $n = 1, 2, \ldots, N$.

  Understanding n-mode matricization may be simpler with an example. Let $\mathcal{X} \in \mathbb{R}^{2 \times 3 \times 2}$ with the frontal slices as follows

$$X_1 = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}, \quad X_2 = \begin{bmatrix} 7 & 9 & 11 \\ 8 & 10 & 12 \end{bmatrix}$$

  Then, mode-n matricization of $\mathcal{X}$ is as follows

$$X_{(1)} = \begin{bmatrix} 1 & 3 & 5 & 7 & 9 & 11 \\ 2 & 4 & 6 & 8 & 10 & 12 \end{bmatrix} \in \mathbb{R}^{2 \times 6},$$

$$X_{(2)} = \begin{bmatrix} 1 & 2 & 7 & 8 \\ 3 & 4 & 9 & 10 \\ 5 & 6 & 11 & 12 \end{bmatrix} \in \mathbb{R}^{3 \times 4},$$

$$X_{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 7 & 8 & 9 & 10 & 11 & 12 \end{bmatrix} \in \mathbb{R}^{2 \times 6}$$

It is also noteworthy that we can unfold a tensor into a vector, which is called *vectorization*.

Using the example above, $\mathcal{X}$ can be vectorized as $\mathcal{X}^v := vec(\mathcal{X}) = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ 12 \end{bmatrix}$.

1.1.2 Tucker Decomposition

Since fundamental tensor properties have been examined, we can proceed with discussing a multilinear operator, the *Tucker operator*.

As discussed in Kolda (2006), the Tucker operator represents multi-mode product. Let $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ and assume we have matrices $\mathbf{A}^{(n)} \in \mathbb{R}^{J_n \times I_n}$ for $n = 1, 2, \ldots, N$. Then, the Tucker operator is defined as follows.

$$[\![\mathcal{X}; \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(N)}]\!] \equiv \mathcal{X} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \cdots \times_N \mathbf{A}^{(N)} \tag{1.1.1}$$

which results in a $(J_1 \times J_2 \times \cdots \times J_N)$ tensor. Various properties of this operator can be found in Kolda (2006).

In Tucker (1963), the Tucker decomposition was introduced for the first time. Since then there have been many names used for it such as three-mode factor analysis (3MFA/ Tucker3), N-mode SVD etc. As discussed in Kolda and Bader (2009), the Tucker decomposition is in a way a version of higher-order principal component analysis. It decomposes an n^{th}-order tensor into a core tensor and $n-$factor matrices that are orthogonal to the core tensor along each mode.

For instance, let $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ be a 3-way tensor. The Tucker decomposition of $\mathcal{X}$ is given by

$$\mathcal{X} = [\![\mathcal{G}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

where $\mathcal{G} \in \mathbb{R}^{J_1 \times J_2 \times J_3}$ and $\mathbf{A} \in \mathbb{R}^{I_1 \times J_1}, \mathbf{B} \in \mathbb{R}^{I_2 \times J_2}$ and $\mathbf{C} \in \mathbb{R}^{I_3 \times J_3}$. $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are factor matrices and $\mathcal{G}$ is the core matrix such that

$$[\![\mathcal{G}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} = \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \sum_{j_3=1}^{J_3} g_{j_1 j_2 j_3} a_{j_1} \circ b_{j_2} \circ c_{j_3}.$$

Here, factor matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ have J_1, J_2 and J_3 many columns respectively. We can also write the decomposition above elementwise as follows.

$$x_{i_1 i_2 i_3} \approx \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \sum_{j_3=1}^{J_3} g_{j_1 j_2 j_3} a_{i_1 j_1} \circ b_{i_2 j_2} \circ c_{i_3 j_3}$$

The following image is from Kolda and Bader (2009) which makes visualizing the Tucker decomposition easier.

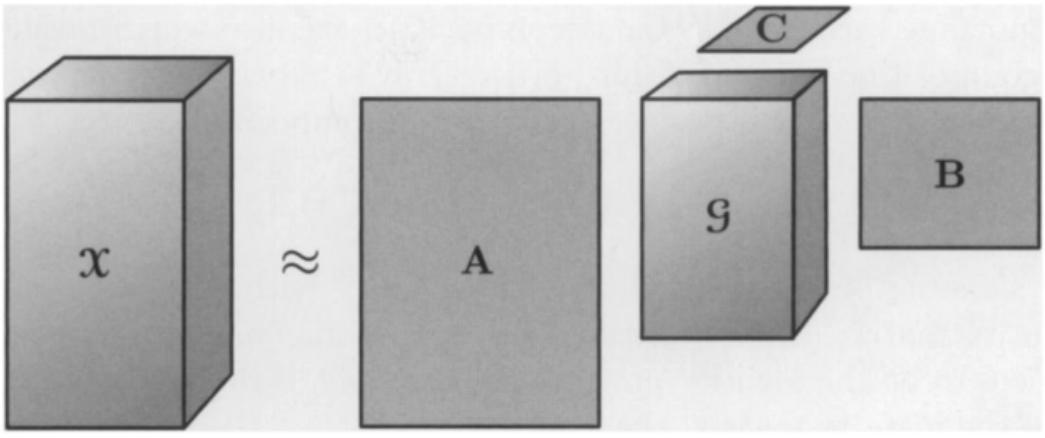

Figure 1.3: Tucker decomposition of a three-way tensor $\mathcal{X}$

1.1.3 More General Notes on Tensors

Zhong et al. (2015) introduces a new model aiming to achieve tensor dimension reduction by generalizing the tensor classification method proposed in Zhong et al. (2015). This method assumes non-linear dependence between the Y vector and the projection of all $\mathcal{X}$ predictors, and follows an estimation method based on sequential iteration.

Additionally, one very important study on tensor predictors is suggested by Zhou et al. (2013). As indicated in the name of article, the authors propose a new tensor regression method which can deal with neuroimaging data where clinical response is represented by a response vector and images are represented by a tensor-valued predictor. The method developed is based on generalized linear model (GLM); therefore, it can be used for both discrete and continuous response vector. It reduces the dimension by a low-rank approximation being imposed to $\mathcal{X}$. Also, as regularization is an important part of the research when $p >> n$, the authors discuss the tensor regression model with and without the sparsity regularization.

Finally, we will investigate the tensor envelope PLS method as well as its population interpretation introduced in Zhang and Li (2017). The goal of this study is not only to estimate the tensor coefficient parameter and predict response vector, but also to reduce the dimensionality of the tensor predictor. The authors first discuss two important vector-PLS algorithms, namely the SIMPLS algorithm and the non-iterative PLS (Wold) algorithm. The objective of both algorithms is to reduce the dimension of the vector predictor $\mathbf{X} \in \mathbb{R}^p$ to a latent vector, which is the linear combination of $\mathbf{X}$, $\mathbf{T} = \mathbf{W}^T\mathbf{X} \in \mathbb{R}^d$ with a lower dimension $d \leqslant p$. These two algorithms have different approaches for getting columns of $\mathbf{W}$, w_s, which causes the interpretation of the estimates of the two algorithms to be different. Upon discussing these algorithms with vector predictors, the authors extend both algorithms in a way that they can handle tensor predictor $\mathcal{X}$. As noted in Zhang and Li (2017), although the Wold PLS was already extended to $\mathcal{X}$ in a number of papers including Zhao et al. (2012), the tensor version of the SIMPLS algorithm is being established for the tensor predictors for the first time.

For the tensor regression model where $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_m}$, the latent tensor is given by the Tucker decomposition of $\mathcal{X}$ as

$$\mathcal{X} = [\![\mathcal{T}; \mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_m]\!]. \tag{1.1.2}$$

Using Tucker operator properties given in Kolda (2006), for orthonormal factor matrices $\mathbf{W}$s, equation (1.1.2) becomes

$$\mathcal{T} = [\![\mathcal{X}; \mathbf{W}_1^T, \mathbf{W}_2^T, \ldots, \mathbf{W}_m^T]\!]. \tag{1.1.3}$$

One important assumption made in this new algorithm is called *separable Kronecker covariance structure*, which is

$$\Sigma_X = cov\{\mathcal{X}^v\} = \Sigma_m \otimes \cdots \otimes \Sigma_1.$$

As for the population interpretation, Zhang and Li (2017) represents the tensor linear model as

$$Y_i = \langle \mathcal{B}_{::i}, \mathcal{X} \rangle + \epsilon_i, \quad i = 1, 2, \ldots, r \tag{1.1.4}$$

where $\mathcal{B}_{::i} \in \mathbb{R}^{p_1 \times \cdots \times p_m}$ is the sub-tensor of the regression coefficient tensor $\mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_m \times r}$.

The model (1.1.4) can also be written in terms of mode-$(m+1)$ matricization of $\mathcal{B}$ as follows.

$$Y = B_{(m+1)}\mathcal{X}^v + \epsilon. \tag{1.1.5}$$

It is worth noting that the dimension of $\mathcal{X}$ reduces from $\prod_{i=1}^m p_i$ to $\sum_{i=1}^m p_i$ when $\mathcal{X}^v$ is used.

The setting we are going to be focusing on in the proposed method also has a tensor predictor and vector response, so we will be utilizing the more compact form given in equation (1.1.5).

1.2 Envelope Models

In a multivariate regression setting, the response vector may not be affected by the changes made in vector X. Cook et al. (2010) propose the envelope method where the part of Y which is immaterial to the changes in the predictors is identified and removed from the analysis. This leads to a more efficient estimation as the immaterial variation which is irrelevant to the regression, yet adds up to the estimative variation, is reduced. In order to separate the two parts, a likelihood based objective function is used, and estimates are derived in Cook et al. (2010).

In order to understand the model better, we will examine the setting in more detail by formalizing the construction first. The definition of an invariant and reducing subspace from Cook et al. (2010) and Cook (2018) are as follows.

Definition 1.2.1.

A subspace $\mathcal{R} \subseteq \mathbb{R}^r$ is an invariant subspace of $M \in \mathbb{R}^{r\times r}$ if $M\mathcal{R} \subseteq \mathcal{R}$. That is, $\mathcal{R}$ is an invariant subspace of M if M maps $\mathcal{R}$ to a subset of itself.

Definition 1.2.2.

A subspace $\mathcal{R} \subseteq \mathbb{R}^r$ is a reducing subspace of $M \in \mathbb{R}^{r\times r}$ if $M\mathcal{R} \subseteq \mathcal{R}$ and $M\mathcal{R}^{\perp} \subseteq \mathcal{R}^{\perp}$. That is, $\mathcal{R}$ reduces M if M can be decomposed as $M = P_{\mathcal{R}}MP_{\mathcal{R}} + Q_{\mathcal{R}}MQ_{\mathcal{R}}$ where $P_{\mathcal{R}}$ represents projection onto subspace $\mathcal{R}$ and $Q_{\mathcal{R}} = I_r - P_{\mathcal{R}}$.

Also, by definition if $\mathcal{R}$ is a reducing subspace of $M \in \mathbb{S}^{r\times r}$, then it can be said that $\mathcal{R}$ reduces M. Here, $\mathbb{S}^{r\times r}$ denotes the class of all symmetric $r \times r$ matrices.

Moreover, as stated by Cook et al. (2010), considering the smallest reducing subspace of a matrix $M \in \mathbb{S}^{r\times r}$ is reasonable as the intersection of its two reducing subspaces is also a reducing subspaces, which leads to the following envelope definition.

Definition 1.2.3.

The M-envelope of $\mathbb{S}$, denoted as $\varepsilon_M(\mathcal{S})$, is the intersection of all reducing subspaces that contain $\mathcal{S}$, where $M \in \mathbb{S}^{r\times r}$ and $\mathbb{S} \subseteq span(M)$.

Let's now consider the multivariate linear regression model given by

$$Y = \mu_Y + \beta(X - \mu_X) + \epsilon \tag{1.2.1}$$

where response vector is $Y \in \mathbb{R}^r$, predictor vector is $X \in \mathbb{R}^p$, the unknown parameter matrix is $\beta \in \mathbb{R}^{r\times p}$, and the error vector $\epsilon \in \mathbb{R}^r \sim N(0, \Sigma)$ with unknown covariance matrix $\Sigma > 0$. Traditionally, this model is conditioned on the observed X if predictor vector X is random. Let $\mathcal{S}$ be a subspace of $\mathbb{R}^r$. Using model (1.2.1), assume that there exists an orthogonal $(\Gamma, \Gamma_0) \in \mathbb{R}^{r\times r}$ such that

(i) $span(\beta) \subseteq \mathcal{S}$

(ii) $\Gamma^T Y \perp\!\!\!\perp \Gamma_0^T Y | X$.

These conditions imply that Σ can be broken down as $\Sigma = P_\Gamma \Sigma P_\Gamma + Q_\Gamma \Sigma q_\Gamma$ which gives the key for envelope models. This is the parametric link between the two parameters β and Σ.

One reason why envelope models are very important in the literature is that in Cook et al. (2010), the authors make a connection between the covariance matrix Σ and the mean function (coefficient matrix) β by using the minimal reducing subspace of the Σ that accommodates the β. Hence, the number of parameters in the regression model decreases maximally.

Corollary 1.2.1 as illustrated in Cook et al. (2010) shows the coordinate-free form of the parametric link mentioned above which utilizes the properties of Σ.

Corollary 1.2.1.

A subspace $\mathcal{R} \subseteq \mathbb{R}^r$ *reduces* Σ *if and only if* Σ *can be written as a sum of two symmetric positive definite matrices* $\Sigma = \Sigma_1 + \Sigma_2$ *such that* $\Sigma_1\Sigma_2 = 0$ *and* $\mathcal{R} = span(\Sigma_1)$.

Then, under the multivariate linear regression model (1.2.1), let $\Gamma \in \mathbb{R}^{r\times u}$ be an orthonormal basis of the envelope $\varepsilon_\Sigma(\beta)$ of dimension u, and $\Gamma_0 \in \mathbb{R}^{r\times(r-u)}$ be a completion of Γ ; that is $(\Gamma, \Gamma_0) \in \mathbb{R}^{r\times r}$ is an orthogonal matrix. Then, there is an $\eta \in \mathbb{R}^{u\times p}$ such that $\beta = \Gamma\eta$ because $span(\beta) \subseteq \varepsilon_\Sigma(\beta) = span(\Gamma)$. Also, since the envelope $\varepsilon_\Sigma(\beta)$ is spanned by eigenvectors of the covariance matrix Σ, there exists an $\Omega \in \mathbb{S}^{u\times u}$ such that $\Omega = \Gamma^T\Sigma\Gamma$ and an $\Omega_0 \in \mathbb{S}^{(r-u)\times(r-u)}$ such that $\Omega_0 = \Gamma_0^T\Sigma\Gamma_0$. Then, using the Corollary (1.2.1), if the following conditions are satisfied, we call model (1.2.1) an envelope model of dimension u:

$$\begin{aligned} \beta &= \Gamma\eta \\ \Sigma &= \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T. \end{aligned} \tag{1.2.2}$$

Here, matrices Ω and Ω_0 carry the coordinates of $\Sigma_1 = \Gamma\Omega\Gamma^T$ and $\Sigma_2 = \Gamma_0\Omega_0\Gamma_0^T$ relative to Γ and Γ_0 respectively, and η carries the coordinates of β relative to Γ. It is important to notice that, the envelope model, which was first developed by Cook et al. (2010), makes a reduction in the y-direction only.

1.2.1 Envelope Estimation

In order to estimate the parameters in model (1.2.2), a fast algorithm proposed by Cook et al. (2016), which doesn't require Grassmann optimization, is employed. This is a fast iterative method that computes $\arg\min L_u(\Gamma)$ where the objective function commonly used for envelope estimation is given as

$$L_u(\Gamma) = ln|\Gamma^\top\hat{V}\Gamma| + ln|\Gamma^\top(\hat{V} + \hat{U})^{-1}\Gamma|. \tag{1.2.3}$$

Here, the minimum is taken over all orthogonal $\Gamma \in \mathbb{R}^{r\times u}$. In addition, $\hat{V}$ is the maximum likelihood estimate of the covariance matrix Σ and $(\hat{V}+\hat{U})$ is the sample covariance matrix of Y. That is, $\hat{V} = S_{Y|X}$ and $\hat{V}+\hat{U} = S_Y$.

The envelope estimate of β in model (1.2.1) can be computed as $\hat{\beta} = P_{\hat{\varepsilon}}\hat{\beta}_{OLS}$ where $\hat{\varepsilon} := \hat{\varepsilon}_\Sigma(\beta) = span\{\arg\min L_u(\Gamma)\}$. This new algorithm estimates a basis Γ row by row. To get the envelope estimator, *env* function from "Renvlp" package can be used.

More recently, Forzani and Su (2019) researched a particular type of envelope model, the reduced-rank envelope, under a more general distribution family, elliptically contoured distributions. This family of distributions is explored next.

1.3 Elliptically Contoured Distributions

Even though a lot of statistical tools rest upon normality assumption, it is well known that this assumption is not always justifiable. Therefore, alternative distributions or methodologies must be considered when normality assumption is not reasonable, so that the distribution fits the data better. One choice available is the elliptically contoured distribution family, which includes normal distribution as well as many other important distributions such as Student-t, Cauchy, power exponential, etc. The elliptically contoured distribution family allows distributions to have heavier or lighter tails than the normal distribution.

Elliptical family of distributions are studied widely in the literature. for example Cook and Nachtsheim (1994) explore approaches to find Voronoi weights in order to induce elliptically contoured covariates in analysis of regression. Through examples they show that using this weighting method can improve various procedures greatly such as estimation with linear response surfaces, estimation with quadratic response surfaces etc.

Díaz-García et al. (2003) introduces several tools to perform diagnostics for the elliptical multivariate regression. They indicate that some standardized residuals are invariant in elliptical model class. The authors discuss residual analysis, local influence, and as the eliminating individual observations in elliptical multivariate regression. They show that these traditional diagnostic models are generally invariant against the distribution of the error term. That is, the known diagnostics graphics used for the normal error can be applied. They also note that the invariance property cannot be used when the parameter of interest is dispersion (covariance) matrix as its diagnostics depend upon the distribution of the error.

Frahm (2004) in their PhD thesis works on generalized elliptical distributions. It is explained that every r-dimensional elliptical random vector Z can be written as $Z =_r \mu + \mathcal{R}\Lambda U^{(k)}$ where $\mu \in \mathbb{R}^r, \Lambda \in \mathbb{R}^{r\times k}$, $U^{(k)}$ is a k-dimensional random vector which is distributed uniformly on the unit hypersphere. $\mathcal{R}$ is a non-negative random variable which is independent of $U^{(k)}$. The distribution function of $\mathcal{R}$ is called the generating distribution function, and it establishes the elliptical distribution family of Z. Let $\mathcal{R}$ belong to the Frechet distribution (Embrechts et al., 2003): $\bar{F}_{\mathcal{R}} = \lambda(z)\, z^{-\alpha}$, $\forall z > 0$ where the parameter which represents the tail index of the generating distribution function $\alpha > 0$ and λ is a slowly varying function (Resnick, 2013). Since $\bar{F}_{\mathcal{R}}$ also corresponds to the tail index of Z, the elliptically contoured distributions conserves a simple linear structure which is characteristics of the normal distribution even in the presence of heavy tails.

Frahm (2004) notes that a lot of nice properties of the Gaussian distribution also exist for the elliptically contoured distributions because the characteristic function of the normal distribution simply weakens in the elliptical distribution family. That is, assuming $t \in \mathbb{R}^d$, $\quad exp(-\frac{1}{2}t'\Sigma^{-1}t)$ becomes $g(t'\Sigma^{-1}t)$ where $g : \mathbb{R} \to [0, \infty)$ is such that $\int_0^\infty u^{\frac{d}{2}-1}g(u)du < \infty$. Here, $g(.)$ is a characteristic function.

If a random vector $Z \in \mathbb{R}^r$ follows an elliptical distribution with characteristic function, also known as density generator, i.e. $Z \sim EC_r(\mu_Z, \Sigma_Z, g)$, then the density function is given by

$$f_Z(z) = |\Sigma_Z|^{-\frac{1}{2}} g[(z-\mu_Z)^T \Sigma_Z^{-1}(z-\mu_Z)] \tag{1.3.1}$$

where $\mu_Z \in \mathbb{R}^r$ is the location parameter and $\Sigma_Z \in \mathbb{R}^{r\times r}$ is a positive definite scale matrix. (Ng et al., 1990).

Some examples of elliptically contoured distributions are listed below using the table from Galea et al. (2020) with c representing the normalizing constant.

Distribution	Notation	Generating function
Normal	$\mathcal{N}_r(\mu, \Sigma)$	$g(u) = c\, exp(-u/2),\ \ u \geqslant 0$
Student-t	$t_r(\mu, \Sigma, \nu)$	$g(u) = c(1+u/v)^{-(v+r)/2},\ \ u \geqslant 0$
Cauchy	$\mathcal{C}_r(\mu, \Sigma)$	$g(u) = c(1+u)^{-(r+1)/2},\ \ u \geqslant 0$
Power Exponential	$\mathcal{P}\varepsilon_r(\mu, \Sigma, \alpha)$	$g(u) = c\, exp(-u^\alpha/2),\ \ u \geqslant 0$

Table 1.1: Examples of Elliptical Distributions

Assume $Y|X \sim EC_r(0, \Sigma, g)$ based on the model (1.2.1) with the density function (1.3.1). Then, $Y|X \sim EC_r(\mu_{Y|X}, \Sigma, g_{Y|X})$ where $\mu_{Y|X} = \mu_Y + \beta(X - \mu_X)$. It is important to note that $E(Y|X) = \mu_{Y|X}$ and $var(Y|X) = c_X\Sigma$ where $c_X = \frac{E\left((Y-\mu_{Y|X})^\top \Sigma^{-1}(Y-\mu_{Y|X})\right)}{r}$ if conditional expectation and variance exists.

For independent samples $(X_i, Y_i), \forall i = 1, \ldots, n$ of (X, Y) and for $m_i = \big(Y_i - \mu_Y - \beta(X_i - \mu_X)\big)^\top \Sigma^{-1} \big(Y_i - \mu_Y - \beta(X_i - \mu_X)\big)$, the log-likelihood function can be obtained as

$$l = -\frac{n}{2}\, log\, |\Sigma| + \sum_{i=1}^{n} log\, g(m_i). \tag{1.3.2}$$

By posing both partial derivatives with respect to β and Σ to be 0

$$\frac{\partial l}{\partial \beta} = -\frac{1}{2}\sum_{i=1}^{n} W_i \frac{\partial m_i}{\partial \beta} = 0, \qquad \frac{\partial l}{\partial \Sigma} = -\frac{n}{2}\Sigma^{-1} - \frac{1}{2}\sum_{i=1}^{n} W_i \frac{\partial m_i}{\partial \Sigma} = 0$$

we can get maximum likelihood estimator for both parameters. Here, $W_i = -\frac{2g'(m_i)}{g(m_i)}$. More specifically, if $Y|X \sim N(0, \Sigma)$, then (1.3.2) becomes

$$l = -\frac{n}{2}\, log\, |\Sigma| - \frac{1}{2}\sum_{i=1}^{n} m_i$$

with

$$\frac{\partial l}{\partial \beta} = -\frac{1}{2}\sum_{i=1}^{n} \frac{\partial m_i}{\partial \beta} = 0, \quad \frac{\partial l}{\partial \Sigma} = -\frac{n}{2}\Sigma^{-1} - \frac{1}{2}\sum_{i=1}^{n} \frac{\partial m_i}{\partial \Sigma} = 0.$$

Then, for known $W_i > 0$ weights, the data can be updated as $(\sqrt{W_i}X_i, \sqrt{W_i}Y_i)$ and the parameter estimators can be computed as the data were distributed normally.

For instance, let $\epsilon \in \mathbb{R}^r$ in multivariate linear regression model given by (1.2.1) be distributed with $t(0, \Sigma, \nu)$. Then, the estimator of $\beta \in \mathbb{R}^{r \times p}$ can be obtained from the following algorithm.

1. Get the initial values for β and Σ.
2. Repeat the following steps until convergence
   (a) Compute the weight W_i values.
   (b) Update the data as $(\sqrt{W_i}X_i, \sqrt{W_i}Y_i)$ and estimate β and Σ as if the data is normally distributed.

Here, β can be found through the ordinary least squares estimation. To get this estimator, *lm* function can be used.

The iterative algorithm above is related to the famous EM algorithm about missing data in Dempster et al. (1977). The iterative reweighted idea has also been discussed in the regression setting in Dempster et al. (1980).

Furthermore, it is known that any marginal as well as conditional distribution function of a random vector that is elliptically contoured is also elliptically contoured. (Kelker, 1970). Frahm (2004) in their thesis generalizes elliptical distributions from only symmetrical case to allow asymmetry, and notes that the skew-elliptical distributions are also in the class of generalized elliptical distributions which is developed in this paper along with its properties. Furthermore, the author shows how the dispersion parameter Σ can be estimated robustly, and calls this estimator "spectral estimator", which is a maximum likelihood estimator. Moreover, since elliptically contoured distributions can be very useful to model financial data, the author examines the impacts of this type of data for statistical inference.

Later, Díaz-García and Gutiérrez-Jáimez (2007) discusses how exploring different kinds of residuals is very important in certain areas of statistics, for instance, for sensitivity analysis in regression. Therefore, they propose methods to find distributions of normalized residuals, standardized residuals, internally studentized residuals, and externally studentized residuals, which are also extended to a less than full rank model. It is also shown that these distributions are valid and invariant under elliptical distribution family.

In Lemonte and Patriota (2011), the authors introduce models to unify several elliptical models such as multivariate errors-in-variables models, mixed models with regressors which depend on measurement errors, nonlinear regression models etc. In the model developed, general elliptical multivariate model, the location parameter (mean vector) and the scale matrix shown to share parameters, and their maximum likelihood estimates are derived.

Moreover, diagnostic methods to determine the influential observations are investigated under perturbations, and the generalized leverage proposed by Wei et al. (1998) is reexpressed for the general elliptical multivariate model. Later, for this general model, Melo et al. (2017) derives two different adjusted likelihood ratio statistics and Melo et al. (2018) provides methods to correct and reduce the bias of the estimates.

CHAPTER 2

REWEIGHTED ENVELOPE MODEL

In the previous chapter, relevant concepts were explained and the notation that is going to be used for the rest of the dissertation is provided. In this chapter, we will elaborate the model and report comparisons of our method with others.

2.1 Population Model

We consider the multivariate regression model with tensor predictor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_m}$, and elliptically distributed error vector $\epsilon \sim EC_r(0, \Sigma, g)$

$$Y = B\mathcal{X}^v + \epsilon \tag{2.1.1}$$

where $B := B_{(m+1)} \in \mathbb{R}^{r \times \prod_{i=1}^{m} p_i}$ and $Y \in \mathbb{R}^r$. Without loss of generality, we will assume that $\mathcal{X}^v$ and Y are centered. Also, to keep the notation simple, we use B for the $(m+1)-$ matricized coefficient tensor, $B_{(m+1)}$.

Through envelope structure introduced in Section 1.2, a link is established between the covariance matrix $\Sigma \in \mathbb{R}^{r\times r}$ and matricized tensor coefficient B in model (2.1.1) as follows

$$B = \Gamma\eta, \quad \Sigma = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T \tag{2.1.2}$$

where $u \leqslant r$ is the dimension of Σ- envelope of $\mathbb{B} = span(B)$, denoted as $\varepsilon_\Sigma(\mathbb{B})$, $\Gamma \in \mathbb{R}^{r\times u}$ is an orthonormal basis of $\varepsilon_\Sigma(\mathbb{B})$, $\Gamma_0 \in \mathbb{R}^{r\times(r-u)}$ is a completion of Γ, $\eta \in \mathbb{R}^{u\times\prod_{i=1}^{p_m} p_i}$ and $\Omega \in \mathbb{S}^{u\times u}$, $\Omega_0 \in \mathbb{S}^{(r-u)\times(r-u)}$.

2.2 Estimation

The purpose of this section is to derive an estimate for the coefficient matrix B in model (2.1.1). Let $(\mathcal{X}_i^v, Y_i)$ for $i = 1, 2, \ldots, n$ be n independent samples, and let

$$m_i = \left(Y_i - B\,\mathcal{X}_i^v\right)^\top \Sigma^{-1} \left(Y_i - B\,\mathcal{X}_i^v\right)$$

for $\forall i$. Then, the weights can be defined as

$$W_i = -2\,\frac{g'(m_i)}{g(m_i)}. \tag{2.2.1}$$

In order to estimate the $(m+1)$- matricized coefficient tensor B assuming the $g(.)$ function is known, we follow the reweighting algorithm below for when $W_i > 0$:

1. Get the initial values for B and Σ.
2. Repeat the following steps until convergence
   (a) Compute the weight W_i values
   (b) Update the data as $(\sqrt{W_i}\mathcal{X}_i^v, \sqrt{W_i}Y_i)$ and and estimate B and Σ as if the data is normally distributed

Here, $\hat{B}$ can be computed through the response envelope estimation which is achieved following the method explained in Section 1.2.1. It is important to note that the initial values can be gotten from the envelope model as well.

Furthermore, we would like to discuss a specific distribution from the elliptical distribution family, t-distribution, to demonstrate how our method works. This distribution is considered for some of the comparisons in simulations section.

Density function for $Y|\mathcal{X}^v \sim t_r(B\,\mathcal{X}^v, \Sigma, \nu)$ is given by

$$f(y) = \frac{\Gamma(\frac{v+r}{2})}{\Gamma(\frac{v}{2})\,\nu\pi^{r/2}\,\sqrt{|\Sigma|}}\Big(1 + \frac{\big(Y - B\,\mathcal{X}^v\big)^\top\,\Sigma^{-1}\,\big(Y - B\,\mathcal{X}^v\big)}{\nu}\Big)$$

That is, $Y|X \sim EC_r(B\,\mathcal{X}^v, \Sigma, g)$ where $g(u) = c(1 + u/v)^{-(v+r)/2}$ and normalizing constant $c = \frac{\Gamma(\frac{v+r}{2})}{\Gamma(\frac{v}{2})\,\nu\pi^{r/2}}$.

From the Table 1.3, we know that for t-distribution, $g(t_i) \propto exp(1 + \frac{t_i}{\nu})^{-\frac{(\nu+r)}{2}}$. Then, as

$$\begin{aligned} g'(t_i) &= \frac{\partial g(t_i)}{\partial t_i} \\ &\propto -\Big(\frac{\nu+r}{2}\Big)\Big(\frac{1}{\nu}\Big)\Big(1 + \frac{t_i}{\nu}\Big)^{-(\nu+r)/2}(1 + \frac{t_i}{\nu})^{-1} \\ &= -\Big(\frac{\nu+r}{2}\Big)\Big(\frac{1}{\nu}\Big)\Big(1 + \frac{t_i}{\nu}\Big)^{-1}\,g(t_i) \end{aligned}$$

then, the weight W_i becomes

$$W_i = -2\,\frac{g'(m_i)}{g(m_i)} = \frac{\nu + r}{\nu + m_i}.$$

The reweighting algorithm then can be followed to get the estimator of B. Note that these weights will be called exact weights throughout this dissertation as $g(.)$ function is assumed to be known.

2.3 Selecting Envelope Dimension

It is important to note that the value of the envelope dimension, u, identifies specific envelope models. Therefore, it can be viewed as a model selection parameter. In order to select the envelope dimension u, a standard method, Bayesian information criterion (BIC), can be used. Using the likelihood function given in Section 1.3.2, the BIC can be calculated as follows when $g(.)$ function is known.

$$-2l + k\, log(n) = n\, log\, |\hat{\Sigma}| - 2\sum_{i=1}^{n} log\, \hat{g}(m_i) + k\, log(n)$$

where $\hat{g}(m_i)$ is the value of $g(m_i)$ with parameter values $\hat{B}$ and $\hat{\Sigma}$ that maximizes the likelihood function l, and k is the number of parameters of the envelope at dimension u.

2.4 Weights When Generating Function is Unknown

While computing the reweighted envelope estimator, the assumption is that the $g(.)$ function is known. However, $g(.)$ may not always be available. The weights calculated in the absence of knowledge of $g(.)$ function is called approximate weights in this dissertation. In these cases, we can make changes in the way we calculate weights in the reweighting algorithm given in Section 2.2, and get approximate weights instead of exact weights. For each $(\mathcal{X}_i^v, Y_i)$, $\forall i = 1, \ldots, n$, the approximate weight can be calculated as

$$c_X = \frac{E(Q^2)}{r} \;\; and \;\; Q^2 = \big(Y_i - B\,\mathcal{X}_i^v\big)^\top \Sigma^{-1} \big(Y_i - B\,\mathcal{X}_i^v\big)$$

$c_{X_i}^{-1/2}$ can be used to update $(\mathcal{X}_i^v, Y_i)$ in step (2b) of the algorithm.

However, if c_{X_i} is unknown, then it can be estimated as $\hat{c}_{X_i} = \big(Y_i - \hat{B}\,\mathcal{X}_i^v\big)^\top \hat{\Sigma}^{-1} \big(Y_i - \hat{B}\,\mathcal{X}_i^v\big)$ where the values $\hat{B}$ and $\hat{\Sigma}$ can be again gotten from the envelope model. Then in step (2b) of the algorithm, data can be updated as $(\hat{c}_{\mathcal{X}_i^v}^{-1/2}\mathcal{X}_i^v, \hat{c}_{X_i}^{-1/2} Y_i)$.

2.5 Numerical Studies

In this section, we first will discuss the methods that will be used to compare estimation results against the reweighted envelope estimator. Next, we will simulations with vector predictors and second-order tensor predictors (matrix), and finally we will complete the section by evaluating the proposed reweighting envelope method through image simulations.

2.5.1 Other Methods to Compare

It is known that the least squares (LS) approach leads to the best estimate when the error terms are normal in regression setting. The least square estimator of B in model (2.1.1) can be calculated as

$$\hat{B}_{LS} = \underset{B \in \mathbb{R}^{r \times p}}{\arg\min} \sum_{i=1}^{n} \big(Y_i - B\mathcal{X}_i^v\big)^\top \Sigma^{-1} \big(Y_i - B\mathcal{X}_i^v\big) \tag{2.5.1}$$

That is, the OLS estimator of B has the form of $\hat{B}_{LS} = (\mathcal{X}^{v\top}\mathcal{X}^v)^{-1}\mathcal{X}^{v\top}Y$. Nevertheless, LS estimator looses its power when the error term is not normal. A more efficient alternative, proposed by Huber (1981), that would perform better for a wide variety of error terms is the class of M-estimators. The M-estimator is a "maximum likelihood type" robust estimator. It uses a general type of loss function, typically one that puts less weight to large residuals.

The M-estimator of the coefficient matrix B can be found by

$$\hat{B}_M = \underset{B\in\mathbb{R}^{r\times p}}{\arg\min} \sum_{i=1}^{n} w_i^r \big(Y_i - B\mathcal{X}_i^v\big)^\top \Sigma^{-1} \big(Y_i - B\mathcal{X}_i^v\big) \tag{2.5.2}$$

where w_i^r are the weights that depend on the residual values. The M-estimator can be computed with an iterative weighting approach. There are different types of weightings used to achieve the robust M-estimator. The one we use, the Huber weighting, gives small weights to those observations with large residuals and gives 1 to the ones with small residuals. For instance, at iteration k, the coefficient matrix B is $B_k = (\mathcal{X}^{v\top} W_{k-1} \mathcal{X}^v)^{-1} \mathcal{X}^{v\top} W_{k-1} Y$. This process, which is similar to Newton-Raphson technique, continues until convergence. To get the M-estimators for our simulations, *rlm* function with (method="M") option from "MASS" package in R is used.

One other method that is used for comparison is basic envelope model, which is essentially the response envelope method which is discussed in detail in Section 1.2.1. To get basic envelope estimate of B, B_{env}, *env* function from "Renvlp" package is used.

A well-known method used for dimension reduction and in higher-order variate setting is partial least squares (PLS). The tensor PLS, that is proposed by Zhang and Li (2017), is also included in the comparison shown in Table (2.3).

For the linear model (1.1.4), Zhang and Li (2017) derives PLS estimators. The authors note that one way of using the envelope idea is that a part of the predictors may be immaterial to the change in the response (envelope in the x-direction), and also they are irrelevant to the other predictor. This idea is highly related to the generalized sparsity principle which can be formally represented as

$$\mathcal{X} \times_k Q_k \perp\!\!\!\perp \mathcal{X} \times_k P_k$$

$$Y \perp\!\!\!\perp \mathcal{X} \times_k Q_k | \mathcal{X} \times_k P_k$$

where $P_k \in \mathbb{R}^{p_k \times p_k}$ is projection onto a subspace $S_k \subseteq \mathbb{R}^{p_k}$ and $Q_k = I_{p_k} - P_k \in \mathbb{R}^{p_k \times p_k}$

for $k = 1, 2, \ldots, m$.

Therefore, the way that the authors used the envelope idea for the model (1.1.5), under the generalized sparsity principle given above is as follows.

$$\begin{aligned} \mathcal{B} &= [\![\Theta; \Gamma_1, \ldots, \Gamma_m, I_r]\!] \quad for some \ \ \Theta \in \mathbb{R}^{u_1 \times \cdots \times u_m} \\ \Sigma_k &= \Gamma_k \Omega_k \Gamma_k^T + \Gamma_{0k} \Omega_{0k} \Gamma_{0k}^T \ \ for \ \ k = 1, 2, \ldots, m \end{aligned} \tag{2.5.3}$$

where $\Gamma_k \in \mathbb{R}^{p_k \times u_k}, \Gamma_{0k} \in \mathbb{R}^{p_k \times (p_k - u_k)}$ such that $P_k = \Gamma_k \Gamma_k^T$, $Q_k = \Gamma_{0k} \Gamma_{0k}^T$ with $\Omega_k \in \mathbb{S}^{u_k \times u_k}$ and $\Omega_{0k} \in \mathbb{S}^{(p_k - u_k) \times (p_k - u_k)}$.

Then, the authors point out that the PLS estimation introduced through the new algorithm essentially gives an estimate to the tensor predictor envelope which captures all the relevant information of the tensor predictor, and this PLS estimator is given as

$$\hat{\mathcal{B}}_{PLS} = [\![\hat{\Psi}; \hat{\mathbf{W}}_1, \hat{\mathbf{W}}_2, \ldots, \hat{\mathbf{W}}_m, I_r]\!] \tag{2.5.4}$$

where $\hat{\Psi} = [\![\hat{C}_T; (\hat{\mathbf{W}}_1^\top \hat{\Sigma}_1 \hat{\mathbf{W}}_1)^{-1}, \ldots, (\hat{\mathbf{W}}_m^\top \hat{\Sigma}_m \hat{\mathbf{W}}_m)^{-1}, I_r]\!]$ and $\hat{C}_T = \hat{cov}(\mathcal{T}, Y)$ and $\hat{\mathbf{W}}_i$ is the estimator of $\mathbf{W}_i$ values given in equation (1.1.2).

We measure performance of the coefficient estimator $\hat{B}$ by calculating the Frobenius distance between the projection of the vectorized estimated coefficient and the projection of the vectorized true coefficient. That is,

$$\Delta = \| P_{B^v} - P_{\hat{B}^v} \|_F$$

is compared, where $P_B = B(B^\top B)^{-1} B^T$ for any B and $\| \cdot \|_F$ denotes the matrix Frobenius norm. The smaller the delta, the better the estimate.

2.5.2 Simulations

For the simulations with vector predictors and second-order tensor predictors (matrix) that were carried out in this section, the settings given below are used.

- $\Gamma \in \mathbb{R}^{r\times u}$, $\Gamma_0 \in \mathbb{R}^{r\times(r-u)}$, $\Omega \in \mathbb{S}^{u\times u}$ $\Omega_0 \in \mathbb{S}^{(r-u)\times(r-u)}$, $\eta \in \mathbb{R}^{u\times\prod_{i=1}^{pm} p_i}$ are all generated from $N(0,1)$.
- The $(m+1)$-matricized tensor predictor $B \in \mathbb{R}^{r\times\prod_{i=1}^{m} p_i}$ is calculated as $B = \Gamma\eta$, and covariance matrix $\Sigma = \Gamma\Omega\Gamma^\top + \Gamma_0\Omega_0\Gamma_0^\top$.
- For matrix predictors $X \in \mathbb{R}^{p_1\times p_2}, \forall X_{ij} \overset{iid}{\sim} N(0,1)$ where $i = 1,\ldots,p_1$ and $j = 1,\ldots,p_2$
- For vector predictors $X \in \mathbb{R}^p$, $\forall X_i \overset{iid}{\sim} N(0,1)$ where $i = 1,\ldots,p$.
- $\epsilon \sim t_r(0,\Sigma,\nu)$
- $Y = BX^v + \epsilon$

For each n, we conduct 100 repetitions. In order to compare the estimation performance of the reweighted envelope algorithm, we compute average Δ for the reweighted envelope algorithm as well as the other well-known methods, and then perform comparisons. All average distances between the true coefficient value and estimates are provided along with the standard errors below.

The first simulations carried out with a vector predictor with $p = 5$ where we utilize a sample size of $n = 200$. The response dimension is $r = 20$, and dimension of envelope is set to $u = 4$. We then compute the reweighted envelope estimator with exact weights, the OLS estimator, the basic envelope estimator, and the M-estimator for the coefficient matrix B.

		Rew Env	OLS	Basic Env	M-estimator
$\nu = 5$	$ave(\Delta)$	0.093	0.115	0.106	0.106
	SE_Δ	0.0015	0.0015	0.0019	0.0013
$\nu = 10$	$ave(\Delta)$	0.089	0.101	0.093	0.100
	SE_Δ	0.0012	0.0013	0.0011	0.0010
$\nu = 20$	$ave(\Delta)$	0.089	0.096	0.089	0.098
	SE_Δ	0.0013	0.0011	0.0012	0.0012
$\nu = 40$	$ave(\Delta)$	0.085	0.093	0.086	0.096
	SE_Δ	0.0012	0.0011	0.0012	0.0010

Table 2.1: Effect of the degrees of freedom.
$p = 5$, $r = 20$, $n = 200$, $u = 4$.

We notice that the reweighted envelope outperforms all other methods when the degrees of freedom ν is 10 or less, and it performs better as ν gets smaller as shown in the Table 2.1. As the degrees of freedom ν gets larger, t-distribution becomes less varying and gets closer to normal in which case reweighted envelope with exact weights still performs very well and gives results that are very close to basic envelope results as expected. We also observe that OLS performs the worst for smaller ν values and improves as ν gets very large, for instance when $\nu = 40$.

Both the reweighted envelope and basic envelope estimators are calculated with BIC estimated envelope dimensions. Next set of simulations provide evidence to suggest that the information criterion BIC performs reasonably well in selecting envelope dimension u.

For the second group simulations that carried out with a matrix predictor where $p_1 = 2$ and $p_2 = 2$, we again utilize a sample size of $n = 200$. ϵ has degrees of freedom $\nu = 5$. We, then, compute the reweighted envelope estimator with exact weights when true envelope dimension u is used and when BIC estimate of the u is used,

the OLS estimator, the basic envelope estimator with estimated envelope dimension (BIC), and the M-estimator for the coefficient matrix.

The comparison is provided for a number of different true envelope dimensions and response dimensions. These settings are chosen so that we can explore performances of the chosen methods.

It is clearly seen from the table that the reweighted envelope estimator with exact weights outperforms the alternatives when the estimated envelope dimension is used. In almost all other cases, again reweighted envelope with the true dimension performs really well compare to OLS, basic envelope and m-estimator. In most cases, both reweighted envelope estimators results in substantially smaller average distance compare to other methods.

		$u=2$ $r=10$	$u=2$ $r=20$	$u=4$ $r=10$	$u=4$ $r=20$
Rew Env - u_{BIC}	$ave(\Delta)$	0.140	0.117	0.103	0.090
	SE_Δ	0.0038	0.0025	0.0018	0.0012
Rew Env - u_{true}	$ave(\Delta)$	0.146	0.116	0.104	0.090
	SE_Δ	0.0065	0.0024	0.0020	0.0012
OLS	$ave(\Delta)$	0.177	0.175	0.122	0.115
	SE_Δ	0.0033	0.0026	0.0018	0.0019
Basic Env	$ave(\Delta)$	0.154	0.150	0.117	0.110
	SE_Δ	0.0038	0.0035	0.0021	0.0020
M-estimator	$ave(\Delta)$	0.162	0.158	0.109	0.103
	SE_Δ	0.0029	0.0019	0.0016	0.0013

Table 2.2: Comparison of the estimated u and true u.
$p_1 = 2, p_2 = 2$, $\nu = 5$, $n = 200$.

Next, we generated samples with various sample sizes $200, 400, 800$ and 1600. Here, $p_1 = 2, p_2 = 2, r = 10, u = 2$ and $\nu = 5$.

		Rew Env	Basic Env	OLS	M-estimator	Tensor PLS
$n = 200$	$ave(\Delta)$	0.145	0.160	0.181	0.163	1.395
	SE_Δ	0.0037	0.0037	0.0031	0.0027	0.0023
$n = 400$	$ave(\Delta)$	0.093	0.113	0.129	0.115	1.401
	SE_Δ	0.0020	0.0026	0.0022	0.0017	0.0019
$n = 800$	$ave(\Delta)$	0.065	0.077	0.085	0.079	1.391
	SE_Δ	0.0014	0.0017	0.0014	0.0012	0.0029
$n = 1600$	$ave(\Delta)$	0.046	0.058	0.065	0.058	1.394
	SE_Δ	0.0011	0.0012	0.0011	0.0010	0.0026

Table 2.3: Effect of sample size.
$p_1 = 2, p_2 = 2$, $r = 10$, $u = 2$, $\nu = 5$.

We notice that the reweighed envelope estimator outperforms all other methods for all four sample sizes, as shown in the Table 2.3. Basic envelope estimator and M-estimator give similar results to one another and performs better than the OLS estimator as expected. The tensor PLS estimator performs poorly compared to all other method which is not surprising as the goal of this method is to result in reduction of dimensionality in the predictor space which is why the setting used in Zhang and Li (2017) has the envelope structure in the predictor space for each p_i as shown in equation (2.5.3) in Section 2.5.1.

In the next set of simulations, we investigated the performance of the reweighted envelope method with exact weights versus approximate weights and compared these two to basic envelope and m-estimators. Here, $p_1 = 2, p_2 = 2, r = 20, u = 4$, and the true u is used throughout these simulations. We generated samples of size $n = 400$.

		Rew Env (Exact Weights)	Rew Env (Approx Weights)	Basic Env	M-estimator
t-dist ($\nu = 3$)	$ave(\Delta)$	0.063	0.064	0.117	0.080
	SE_Δ	0.0009	0.0008	0.0044	0.0010
t-dist ($\nu = 5$)	$ave(\Delta)$	0.064	0.064	0.079	0.075
	SE_Δ	0.0009	0.0008	0.0011	0.0009
t-dist ($\nu = 10$)	$ave(\Delta)$	0.063	0.063	0.066	0.069
	SE_Δ	0.0009	0.0009	0.0010	0.0009
Normal Dist.	$ave(\Delta)$	0.062	0.063	0.059	0.065
	SE_Δ	0.0009	0.0009	0.0008	0.0007

Table 2.4: Comparison of exact weights and approximate weights.
$p_1 = 2, p_2 = 2$, $r = 20$, $u = 4$, $n = 400$.

As shown in Table 2.4, reweighted envelope estimator with approximate weights leads to very similar/ almost the same results as the exact weights. In order make this comparison, errors from two different elliptically contoured distributions (t-distribution and normal distribution) are generated. Reweighted envelope with exact and approximate weights perform the best when errors are from t-distribution with mean 0 and a small degrees of freedom ($\nu = 3$) as well as a large one ($\nu = 10$). When error is generated from multivariate normal distribution, it is observed that the basic envelope estimator gives better results as expected, yet both reweighted envelope estimators perform quite well.

2.5.3 Image Simulations

In this set of simulations, we choose the coefficient matrix in such a way that it takes values of 1 and -1 in which -1 values take a particular shape like a square and a cross. We considered a model where $p_1 = 2, p_2 = 2, r = 9$. The sample generated has

a size $n = 200$.

In the first image, the square, degrees of freedom is set to $\nu = 3$. The reweighted envelope estimator is calculated with exact weights and BIC estimate is used for the dimension of envelope. After 50 iterations, we observe the following values:

	Rew Env	Basic Env	OLS
$ave(\Delta)$	0.057	0.071	0.080
SE_Δ	0.0022	0.0036	0.0032

Table 2.5: Comparison for square image

In the second image, the cross, degrees of freedom is set to $\nu = 2$. The reweighted envelope estimator is again calculated with exact weights and BIC estimate is used for the dimension of envelope. After 50 iterations, the following values are observed:

	Rew Env	Basic Env	OLS
$ave(\Delta)$	0.055	0.156	0.161
SE_Δ	0.0024	0.0170	0.0184

Table 2.6: Comparison for the cross image

Three methods are considered: reweighted model, basic envelope and OLS. As it is seen in both Table 2.5 and Table 2.6, as well as the figures Figure 2.1 and Figure 2.2, the reweighted envelope outperforms the alternatives as it leads to the smallest average Δ in the square case, and substantially small average Δ in the cross case. The images estimated with reweighted envelope captures the true shape very precisely.

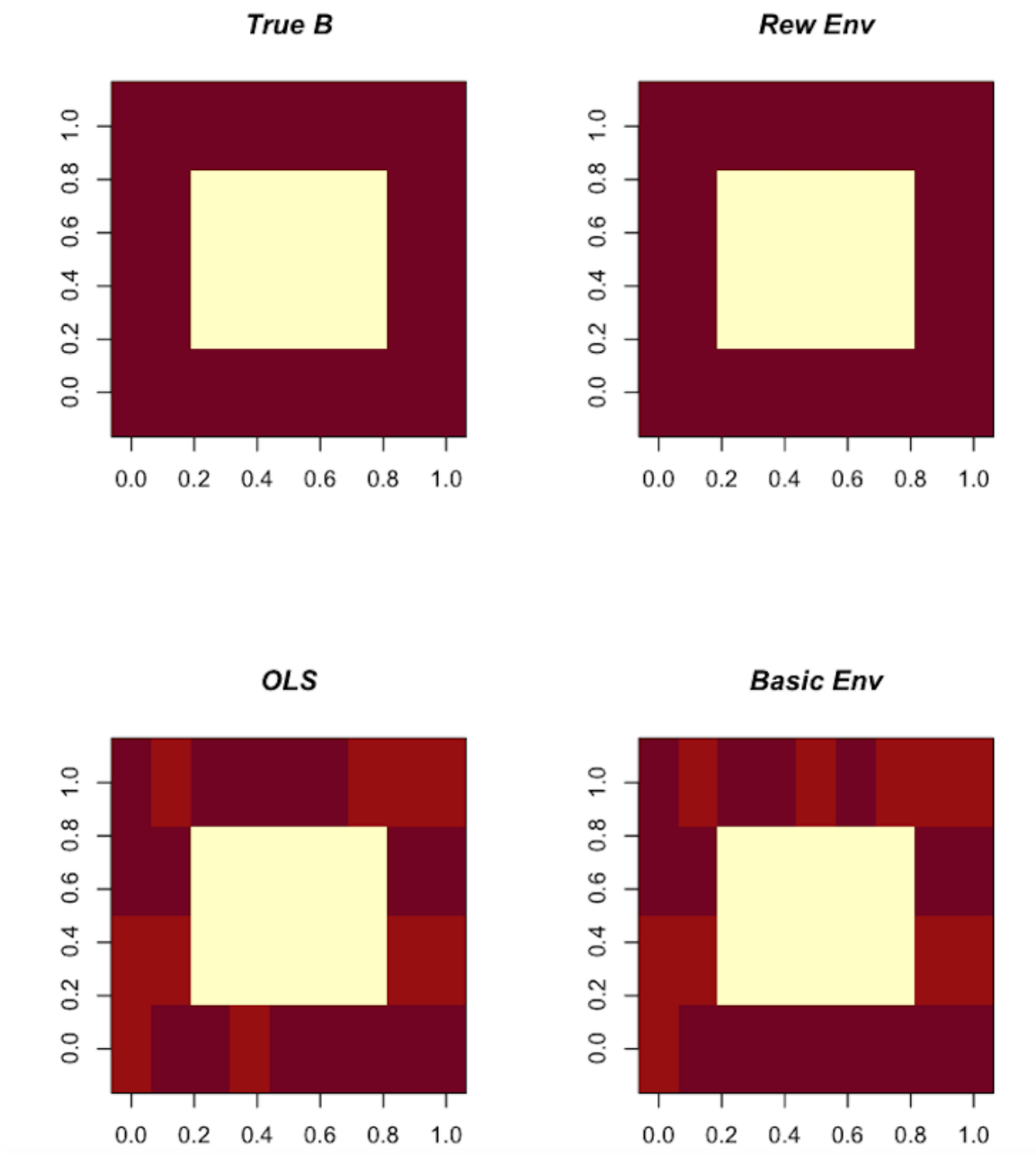

Figure 2.1: Square Image Example

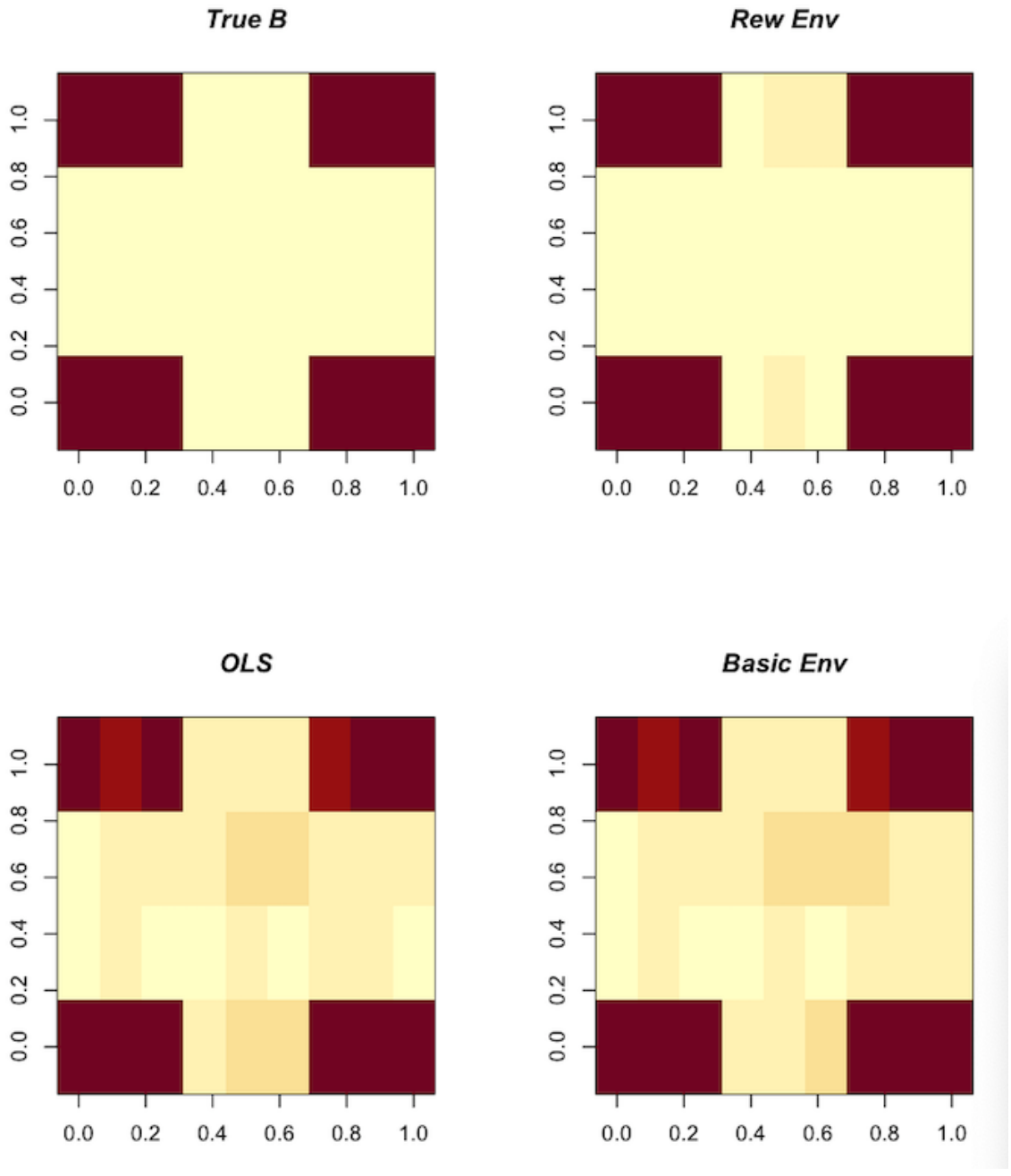

Figure 2.2: Cross Image Example

CHAPTER 3

ENVELOPES WITH IGNORABLE MISSING DATA IN ELLIPTICAL MULTIVARIATE LINEAR REGRESSION

Missing data problems emerge in most statistical applications. Missingness in data may be caused by different reasons including but not limited to cases where a subject refuses to respond to certain questions in a survey setting or when a patient drops out from a study etc.

Extension of envelope methods to cases where missingness is present in data has not been examined thoroughly. As explained in great detail in the first chapter, envelope models can enhance the estimation efficiency significantly; however, the existing envelope methods can be conducted only if we have the data fully observed to solve for the orthonormal basis Γ of the envelope (Cook and Zhang, 2016). Since missing data, especially in the big data era, is nearly inevitable, it is important to explore how envelope estimation can be improved when data is incomplete and

distributed with elliptical distribution.

In this chapter, we will investigate how multiple imputation can be used to improve regression coefficient estimation when missing data is involved. We will focus on improving the estimation once the multiple imputation fills the missing observations, and the way we achieve better parameter estimation is through using the reweighting idea that is proposed in Chapter 2. We will use the "Amelia" package (Honaker et al., 2011) to employ the multiple imputation in the simulations studies as well as the real data example. First, we will review different missingness structures.

3.1 Missing Data and Multiple Imputation

Little and Rubin (2019) defines missing data as the observations that are unobserved and if they were observed, these values would be meaningful for the statistical analyses. Missingness mechanism describes the association between the variable values in a data matrix and missingness.

Let $D \in \mathbb{R}^{n \times p}$ be the data matrix with parameters θ and let $R = (r_{ij})$, the missingness indicator matrix with parameters ϕ be defined as

$$r_{ij} = \begin{cases} 1, & \text{if } d_{ij} \text{ is missing} \\ 0, & \text{if } d_{ij} \text{ is observed.} \end{cases} \tag{3.1.1}$$

That is, $D_{obs} = \{d_{ij} | r_{ij} = 0\}$ and $D_{mis} = \{d_{ij} | r_{ij} = 1\}$ represents observed part and missing part of the data respectively. That is, $D = (D_{obs}, D_{mis})$. Two sets of parameters are considered: the parameter of interest θ and the nuisance parameters ϕ (Zhang, 2003).

Then, joint probability of the missingness indicator R and the data matrix D can be written as

$$P(D, R|\theta, \phi) = P(D|\theta)P(R|\phi, D) \tag{3.1.2}$$

where the marginal distribution of the data is denoted by $P(D|\theta)$ and the conditional distribution of the missingness indicator is denoted by $P(R|\phi, D)$.

3.1.1 Missingness Mechanism

It is crucial to examine underlying missingness mechanisms because assuming that missingness is random without any investigation and ignoring the missing values can lead to invalid and biased results. Three missingness mechanisms that will be defined here are missing completely at random, missing at random (Rubin, 1976) and missing not at random.

- The missingness mechanism is said to be Missing Completely at Random (MCAR) if the missingness is independent of both the observed and the missing values. That is, if

  $$P(R|D_{obs}, D_{mis}, \phi) = P(R|\phi)$$

  where $P(.)$ is the probability distribution function. It is important to note that this is the best possible case as the D_{obs} can be viewed as a random sample of the complete data matrix D, yet the most restrictive one at the same time. An example of MCAR would be that the survey results got lost in the mail.

- The missingness mechanism is said to be Missing at Random (MAR) if the missingness is independent of the missing values (depends only on observed

values). That is, if

$$P(R|D_{obs}, D_{mis}, \phi) = P(R|D_{obs}, \phi).$$

MAR is less constraining than MCAR because the missing values can depend on the response values through the observed portion of the data. An example of MAR would be that if students don't score well in midterms, they are more likely to drop out from a class, and have a missing value for their final exams.

- The missingness mechanism is said to be Missing Not at Random (MNAR) if the missingness depends both the missing values and the observed values. An example of MNAR would be that of people refusing to answer an income related question because their income is unusually high.

As discussed in Chapter 6 of Little and Rubin (2019), for both Bayesian and likelihood based approaches, MAR is a sufficient condition to make the inferences valid without requiring modeling the missingness mechanism. Furthermore, θ and ϕ are called distinct,

- from a Bayesian perspective if the joint prior distribution of θ and ϕ can be written as a product of independent marginal prior distributions, and
- from a frequentist perspective if θ and ϕ has a joint parameter space that factorizes into a θ-space and a ϕ-space. (Rubin, 1976; Rubin, 1987; Little and Rubin, 2019)

If θ and ϕ are distinct and missingness has a mechanism of MCAR or MAR, then missingness is said to be ignorable. That is, missingness mechanism can be ignored while making statistical inferences. Therefore, we assume MAR and ignorable missingness structure throughout our simulation studies in this dissertation.

As Rubin and Little (2019) explained, focusing the analysis only on the observed portion of the data, also known as complete case analysis, would results in inefficient and possibly biased conclusions. Thus, we will base our approach in multiple imputation, one of the best missing data handling methods, which is explored in the next section.

3.1.2 Multiple Imputation

Multiple imputation idea first was proposed by Rubin (1977). In recent years, multiple imputation has been a popular approach while dealing with missingness in data because it accounts for the uncertainties associated with imputations unlike many other methods that can be used in the presence of missingness.

Multiple imputation consists of three steps, which are the imputation step, the analysis step and the pooling step. In the imputation step, each missing value is filled with $M(M > 1)$ independent draws from a predictive distribution given observed values resulting in M "complete" data sets. Next, in the analysis step, each imputed data set is analyzed separately. In this step, standard analysis methods can be applied. Finally in step three, the pooling step, M analysis results are combined in such a way that uncertainties are accounted for by keeping between and within imputation variances of the pooled parameter estimates. Consequently, in the end the uncertainties are correctly accounted for in the final inferences.

It is important to note that when data suffers from missing values, the full data can be expressed as the joint probability distribution $P(D, R|\theta, \phi) = P(D_{obs}, D_{mis}, R|\theta, \phi)$. Due to the fact that D_{mis} unknown, it is not possible to assess this joint distribution; therefore, the likelihood function of the observed data will be the evaluated portion. Let $L(.)$ denote the likelihood function,

$$L(\theta, \phi|D_{obs}, R) \propto P(D_{obs}, R|\theta, \phi) \tag{3.1.3}$$

and

$$\begin{aligned} P(D_{obs}, R|\theta, \phi) &= \int P(D_{obs}, D_{mis}, R|\theta, \phi) dD_{mis} \\ &= \int P(R|D_{obs}, D_{mis}, \phi) P(D_{obs}, D_{mis}|\theta) dD_{mis} \end{aligned} \tag{3.1.4}$$

where for missing at random (MAR)

$$P(D_{obs}, R|\theta, \phi) = P(R|D_{obs}, \phi) P(D_{obs}|\theta). \tag{3.1.5}$$

Since we assume ignorable missingness, through the fact that the parameter of interest θ and the nuisance parameter ϕ being distinct, we can base our likelihood-based inferences on the $P(D_{obs}|\theta)$ part alone ignoring $P(R|D_{obs}, \phi)$. In other words, for the inferences on θ, instead of joint observed data probability distribution $P(D_{obs}, R|\theta, \phi)$, the marginal observed data probability distribution $P(D_{obs}|\theta)$ can be evaluated.

In this chapter, we will again consider the model (2.1.1):

$$Y = B\mathcal{X}^v + \epsilon$$

with $\epsilon \sim EC_r(0, \Sigma, g)$ where $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_m}$, $B := B_{(m+1)} \in \mathbb{R}^{r \times \prod_{i=1}^{m} p_i}$ and $Y \in \mathbb{R}^r$. Our focus is going to be on improving the estimation of B in the presence of missing values in response. First, multiple imputation will be applied to data $D = (Y, \mathcal{X})$, then the estimated coefficient matrices $\hat{B}^{(i)}$s will be computed from M imputed data sets, and finally these M estimates of B will be pooled as

$$\overline{B} = \frac{1}{M} \sum_{i=1}^{M} \hat{B}^{(i)}. \tag{3.1.6}$$

Let $\hat{D}^{(j)} = (D_{obs}, D_{mis}^{(j)})$ with $j = 1, 2, \ldots, M$. Here, within and between imputation variance of the $\overline{B}$ can be calculated as $\frac{1}{M} \sum_{j=1}^{M} \hat{D}^{(j)}$ and $\frac{1}{M-1} \sum_{j=1}^{M} (\hat{B}^{(i)} - \overline{B})^2$ respectively.

Throughout the numerical studies in this chapter, "Amelia" package in R is used to get multiply imputed data sets. Therefore, we will explore this package next.

Amelia package (Version II) written by Honaker et al. (2011) is a great tool to apply multiple imputation in R software. "Amelia" uses a bootstrapping approach to conduct the multiple imputation (MI). The idea is that the algorithm, called expectation-maximization with bootstrapping (EMB), takes multiple bootstrapped samples of the observed data, and then uses expectation-maximization (EM) to fill in the missing values.

In essence, EM algorithm is an iterative process which starts with estimating initial parameters upon filling missing values with initial values such as the mean value of the data, and then the missing values are updated by using the predicted parameter values and then parameters are estimated again, and so on until convergence.

Upon imputing the missing values with EMB algorithm, we then analyze data and pool the estimated coefficients. Below is an image from Honaker et al. (2011) to help visualize the process we will be using.

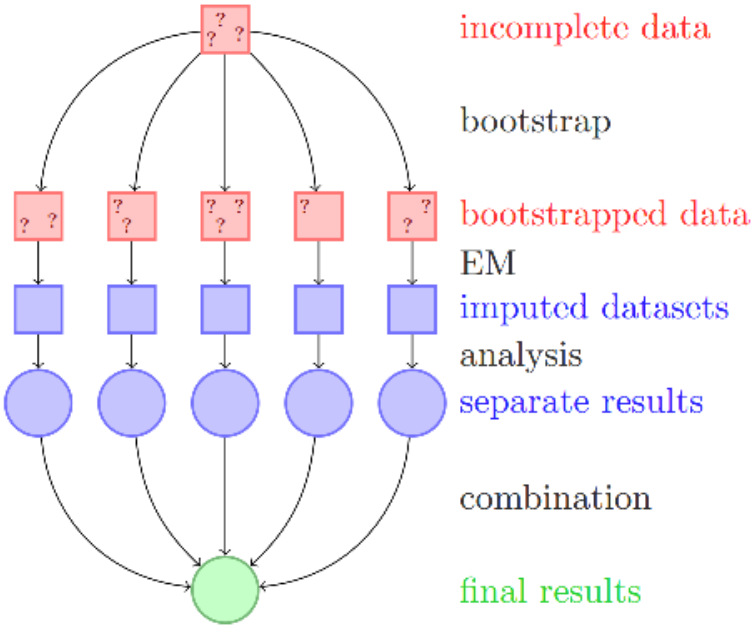

Figure 3.1: Schema of our Multiple Imputation Approach

One assumption that is made by Honaker et al. (2011) is that the data $D = (D_{obs}, D_{mis})$ has multivariate normal distribution with parameter vector θ. Even though this is a restrictive assumption, because the EMB algorithm is really fast and easy to use, and still performs relatively well even when the assumption is violated, we will use the "Amelia" package, specifically *amelia* function, and focus on improving the analysis part which comes after obtaining M imputed data sets.

As mentioned before, for the inferences on θ, the marginal observed data probability distribution $P(D_{obs}|\theta)$ can be used. The observed data likelihood function then is proportional to this marginal observed data probability distribution. That is,

$$L(\theta|D_{obs}) \propto P(D_{obs}|\theta).$$

Following this, let $\pi(\theta)$ be the prior distribution of parameter of interest θ,

$$P(\theta|D_{obs}) \propto L(\theta|D_{obs}) \times \pi(\theta).$$

Our focus is on the inference of complete data because we investigate an approach that uses imputation to deal with the missing values. Posterior predictive distribution of the missing data D_{mis} given the observed data D_{obs}, proposed by Rubin (1987), is defined as

$$P(D_{mis}|D_{obs}) = \int P(D_{mis}|D_{obs}, \theta)P(\theta|D_{obs})d\theta. \tag{3.1.7}$$

Because of the integration used in the definition of the posterior predictive distribution in (3.1.7) of D_{mis} given D_{obs}, this distribution cannot generally be written in a closed form. That is, taking random samples from such a distribution is challenging and non-trivial. As done by Honaker et al. (2011), when assuming multivariate distribution for the data D, we know that predictive distribution of D_{mis} given D_{obs}, θ would also be normal. That is, if we can draw samples from $P(\theta|D_{obs})$ to get values for

θ, then the draws from $P(D_{mis}|D_{obs}, \theta)$ are essentially the draws from $P(D_{mis}|D_{obs})$.

It is computationally challenging to take sample draws from the posterior distribution $P(\theta|D_{obs})$. In order to take random draws from this posterior, EMB algorithm incorporates the standard EM algorithm as well as the bootstrap method. Honaker and King (2010) elaborates how EMB algorithm operates. Moreover, more details about the EM algorithm can be found in the appendix.

3.2 Analysis of Imputed Data Sets

Upon obtaining data with missingness, we first use multiple imputation through bootstrapped EM algorithm which is briefly discussed in the previous section. Next, for each one of the M imputed data sets, we estimate the coefficient B following the algorithm given in Section 2.2. Next, these M estimates $\hat{B}$s are pooled together as $\hat{B}_p = \frac{1}{M}\sum_1^M \hat{B}$.

3.3 Numerical Studies

In this section, we will go over a number of numerical studies. In the first part, we will consider synthetic data examples with vector and tensor predictors under a number of different conditions. Following that, we will examine image simulations where the coefficient matrix is set in a way that it takes values of 1 and -1 where -1 values take a certain shape. Finally, we will discuss a real data application with the cattle data from Kenward (1987). We artificially introduce missingness into the predictor and the response, and examine the performance of the reweighted multiple imputation on this data set and compare it to some other popular methods that are used to deal with missing data. We used $M = 5$ for any study conducted in this section.

3.3.1 Other Methods to Compare

- Full Data Analysis: In the full data analysis, we focus on the data as if there is no missingness involved. Then, to analyze this full data set, the reweighting algorithm proposed in Chapter 2 is used. This is the best a model can get, so we will compare the other methods to the full data analysis, and decide the best one based on their closeness to the results of the full model.

- Complete Case Analysis: In complete case analysis, we essentially remove all observations that has missingness, and analyze the observed portion of the data following the algorithm proposed in this dissertation.

- PEMM Approach: As discussed in Section 1.2.1, to minimize the objective function given in equation (1.2.3), we can use $\hat{V} = S_{Y|X}$ and $\hat{V} + \hat{U} = S_Y$ where $S = \begin{pmatrix} S_Y & S_{YX} \\ S_{XY} & S_X \end{pmatrix}$ is the covariance matrix of (Y_i, X_i), and MLE of S when the data is fully observed However, when data has missingness, we need to make adjustments in the way the MLE of S is calculated. We will use penalized EM algorithm idea that incorporates missingness mechanism, proposed by Chen et al. (2014). This approach will still provide us with $\sqrt{n}$-consistent estimates for $V = Var(Y_i|X_i) = \Sigma$ and $V + U = Var(Y_i)$. Further details can be found in Chen et al. (2014). To get maximum likelihood estimates for V and $V + U$, we use *PEMM_fun* function from "PEMM" package in R software.

  The way we use this approach is that, we first compute $\hat{S}$ through applying PEMM algorithm on (Y_{obs}, X). Then, we get $\hat{V} = \hat{S}_Y - \hat{S}_{YX}\hat{S}_X^{-1}\hat{S}_{YX}^\top$ and $\hat{U} = \hat{S}_{YX}\hat{S}_X^{-1}\hat{S}_{YX}^\top$ from *PEMM_fun* function in R. Finally, *envMU* function from "Renvlp" package is used to get envelope estimates for the coefficient matrix and error covariance matrix.

- Mean-Imputation: The missing values are imputed with the mean of the observed data and once the data has no missing values, then it is analyzed using the reweighting algorithm.

3.3.2 Simulations

In this section, we conducted simulations for both vector predictors and second-order tensor predictors (matrix predictors). The settings used is as follows. $\Gamma \in \mathbb{R}^{r\times u}$, $\Gamma_0 \in \mathbb{R}^{r\times(r-u)}$, $\Omega \in \mathbb{S}^{u\times u}$ $\Omega_0 \in \mathbb{S}^{(r-u)\times(r-u)}$, $\eta \in \mathbb{R}^{u\times\prod_{i=1}^{pm} p_i}$ are all generated from $N(0,1)$. Also, the $(m+1)$-matricized tensor predictor $B \in \mathbb{R}^{r\times\prod_{i=1}^{m} p_i}$ is calculated as $B = \Gamma\eta$, and covariance matrix $\Sigma = \Gamma\Omega\Gamma^\top + \Gamma_0\Omega_0\Gamma_0^\top$. Then, we use the model $Y = BX^v + \epsilon$ with $\epsilon \sim t_r(0,\Sigma,\nu)$. Moreover, vector predictors $X \in \mathbb{R}^p$ are generated with $\forall X_i \overset{iid}{\sim} N(0,1)$ where $i = 1,\ldots,p$ and similarly matrix predictors $X \in \mathbb{R}^{p_1\times p_2}$ are generated with $\forall X_{ij} \overset{iid}{\sim} N(0,1)$ where $i = 1,\ldots,p_1$ and $j = 1,\ldots,p_2$. 100 repetitions were run for each setting, and in order to compare performances, we compared average Δ where $\Delta = \|P_{B^v} - P_{\hat{B}^v}\|_F$. For tensor predictor studies below, around 30% of the response variables have missing values.

Five missingness mechanisms for the response vector is chosen as follows. $\text{logit}P(R_{Y_{i,2}} = 1, R_{Y_{i,4}} = 1|x^v_{i,1}, y_{i,8}, y_{i,9}) = 2 - x^v_{i,1} - y_{i,8} - 3y_{i,9}$, $\text{logit}P(R_{Y_{i,3}} = 1, R_{Y_{i,9}} = 1|x^v_{i,2}, y_{i,4}, y_{i,6}) = 2 - x^v_{i,2} - 3y_{i,4} + 3y_{i,6}$, $\text{logit}P(R_{Y_{i,7}} = 1, R_{Y_{i,8}} = 1|y_{i,1}, y_{i,2}, y_{i,3}) = 2 - 2y_{i,1} + y_{i,2} - 3y_{i,3}$, $\text{logit}P(R_{Y_{i,1}} = 1, R_{Y_{i,10}} = 1|x^v_{i,1}, x^v_{i,2}) = 2 - x^v_{i,1} - x^v_{i,2}$, $\text{logit}P(R_{Y_{i,5}} = 1, R_{Y_{i,6}} = 1|y_{i,1}, y_{i,10}, x^v_{i,1}, x^v_{i,2}) = 1 - y_{i,1} - y_{i,10} - x^v_{i,1} - x^v_{i,2}$. For each subject i, one of these missingness mechanism is randomly selected and then the missingness indicator $R = (R_{Y_{i,1}}, \ldots, R_{Y_{i,r}})$ is generated.

We first start with the comparison between reweighted multiple imputation method, full data analysis, complete case analysis and the PEMM approach for when the predictor is a vector. Average Δ is compared when error term has t-distribution with two different degrees of freedoms and when it is normally distributed.

		Full	Reweighted MI	Complete Case	PEMM
t-dist ($\nu = 5$)	$ave(\Delta)$	0.155	0.160	0.192	0.233
	SE_Δ	0.0034	0.0035	0.0048	0.0054
t-dist ($\nu = 10$)	$ave(\Delta)$	0.152	0.153	0.192	0.202
	SE_Δ	0.0037	0.0036	0.0043	0.0051
Normal Dist.	$ave(\Delta)$	0.141	0.142	0.165	0.172
	SE_Δ	0.0030	0.0031	0.0035	0.0035

Table 3.1: Comparison for vector response.
$p = 3$, $r = 10$, $u = 2$, $n = 200$.

We notice in Table 3.1 that the reweighted MI is the method that gets closest to the best possible analysis, full data analysis, and it performs better as ν gets larger and the distribution finally becomes normal. In the standard normal case, we observe that reweighted multiple imputation performs well as expected because while imputing the missing values it assumes normality. Similarly, since PEMM approach considers parameter estimation under normality assumption.

Next group of simulations are carried out with a second order tensor (matrix) predictor. Here, we investigate how reweighted multiple imputation compares to the alternative methods and the full data analysis when envelope has different dimensions.

In Table 3.2, it is shown that for both envelope dimensions $u = 4$ and $u = 1$, reweighted multiple imputation performs almost as well as the full data analysis. A popular approach, complete case analysis performs worse than our proposed estimation method which then is followed by the PEMM approach with an even larger distance from the full data analysis. It is important to note that the PEMM approach assumes normality, therefore it's poor performance is not surprising when ε has an elliptical distribution.

Moreover, as the envelope dimension gets larger, the amount of relevant information captured by the envelope is larger. Thus, the fact that we observe that all methods result in smaller average Δ values, as u gets larger, is expected.

		Full	Reweighted MI	Complete Case	PEMM
$u = 4$	$ave(\Delta)$	0.066	0.067	0.084	0.103
	SE_Δ	0.0007	0.0008	0.0009	0.0012
$u = 1$	$ave(\Delta)$	0.070	0.072	0.085	0.105
	SE_Δ	0.0014	0.0016	0.0013	0.0021

Table 3.2: Effect of envelope dimension when response is missing.
$p_1 = 3, p_2 = 3$, $r = 20$, $v = 5$, $n = 400$.

Next, we evaluate three methods and compare them to full data analysis for two different sample sizes. Table 3.3 shows that reweighted MI outperforms both the complete case analysis and the PEMM approach, and it performs almost as well as the full data analysis. As sample size gets larger, t-distribution becomes less variable, and it more and more resembles the normal distribution. That is why, the amount of improvement we observe in all cases is expected.

		Full	Reweighted MI	Complete Case	PEMM
$n = 400$	$ave(\Delta)$	0.066	0.067	0.084	0.103
	SE_Δ	0.0007	0.0008	0.0009	0.0012
$n = 600$	$ave(\Delta)$	0.054	0.054	0.067	0.082
	SE_Δ	0.0006	0.0005	0.0008	0.0010

Table 3.3: Effect of sample size when response is missing.
$p_1 = 3, p_2 = 3$, $r = 20$, $v = 5$, $u = 4$.

Finally, we conducted simulation studies again with tensor predictors to see how each method compares to the full data analysis when error term has t-distribution with different degrees of freedoms.

		Full	Reweighted MI	Complete Case	PEMM
$\nu = 5$	$ave(\Delta)$	0.066	0.067	0.084	0.103
	SE_Δ	0.0007	0.0008	0.0009	0.0012
$\nu = 10$	$ave(\Delta)$	0.065	0.065	0.082	0.086
	SE_Δ	0.0008	0.0008	0.0009	0.0011

Table 3.4: Effect of degrees of freedom when response is missing.
$p_1 = 3, p_2 = 3$, $r = 20$, $n = 400$, $u = 4$.

In Table 3.4, we observe similar results when we compare three methods against the full data analysis for different degrees of freedoms. Reweighted multiple imputation outperforms both alternative methods, and it gets very close to the full data analysis when we have a smaller degrees of freedom chosen, and gives the same average distance for larger degrees of freedom values like $\nu = 10$.

3.3.3 Image Simulations

In this set of simulations, we considered a model where $p_1 = 2, p_2 = 2, r = 9$. We choose the coefficient matrix in such a way that it takes values of 1 and -1 in which -1 values take a certain shape like a square and a cross. The sample generated has a size $n = 200$. We again compared average Δ where $\Delta = \|P_{B^v} - P_{\hat{B}^v}\|_F$.

Similar to Section 3.3.2, five missingness mechanisms for the response vector was chosen, and for each subject i, one of these missingness mechanism selected and then the missingness indicator $R = (R_{Y_{i,1}}, \ldots, R_{Y_{i,r}})$ is generated. Following shows the missingness mechanisms that were used. $\text{logit}P(R_{Y_{i,2}} = 1, R_{Y_{i,3}} = 1, R_{Y_{i,4}} =$

$1|x^v_{i,1}, y_{i,8}, y_{i,9}) = 2 - x^v_{i,1} - y_{i,8} - 3y_{i,9}$, $\text{logit}P(R_{Y_{i,9}} = 1|x^v_{i,2}, y_{i,4}, y_{i,6}) = 2 - x^v_{i,2} - 3y_{i,4} + y_{i,6}$, $\text{logit}P(R_{Y_{i,7}} = 1, R_{Y_{i,8}} = 1|y_{i,1}, y_{i,2}, y_{i,3}) = 2 - 2y_{i,1} + y_{i,2} - 3y_{i,3}$, $\text{logit}P(R_{Y_{i,1}} = 1|x^v_{i,1}, x^v_{i,2}) = 2 - x^v_{i,1} - x^v_{i,2}$, $\text{logit}P(R_{Y_{i,5}} = 1, R_{Y_{i,6}} = 1|y_{i,1}, y_{i,10}, x^v_{i,1}, x^v_{i,2}) = 1 - y_{i,1} + y_{i,9} - x^v_{i,1} - x^v_{i,2}$.

In the first image, the square, degrees of freedom is set to $\nu = 3$. After 100 iterations, we observe the following values given in Table 3.5.

	Full	Reweighted MI	Complete Case
$ave(\Delta)$	0.0504	0.0546	0.0662
SE_Δ	0.0014	0.0013	0.0016

Table 3.5: Comparison for square image with missing response and degrees of freedom $\nu = 3$

In the second image, the cross, degrees of freedom is set to $\nu = 5$. After 100 iterations, we observe the following values given in Table 3.6.

	Full	Reweighted MI	Complete Case
$ave(\Delta)$	0.0489	0.0489	0.0626
SE_Δ	0.0012	0.0012	0.0015

Table 3.6: Comparison for cross image with missing response and degrees of freedom $\nu = 5$

The reweighted multiple imputation and complete case analysis are compared to the full data analysis. As it can be seen in both Table 3.5 and Table 3.6, as well as the figures Figure 3.2 and Figure 3.3, the reweighted multiple imputation outperforms the alternative method, complete case analysis as it leads to an average Δ that is closer to the Δ of the full data analysis in the square case, and an average Δ that

is the same in the cross case. The images estimated when the reweighted multiple imputation is used captures the true shape very precisely.

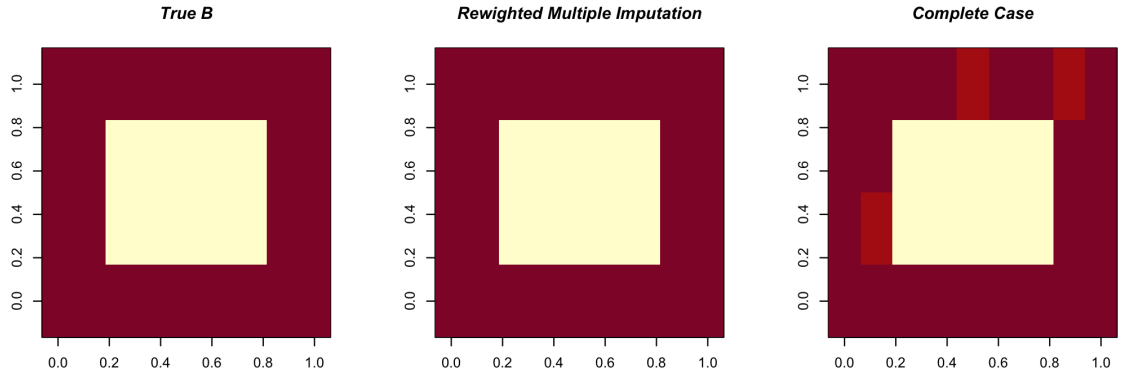

Figure 3.2: Square Image Example with Missing Response

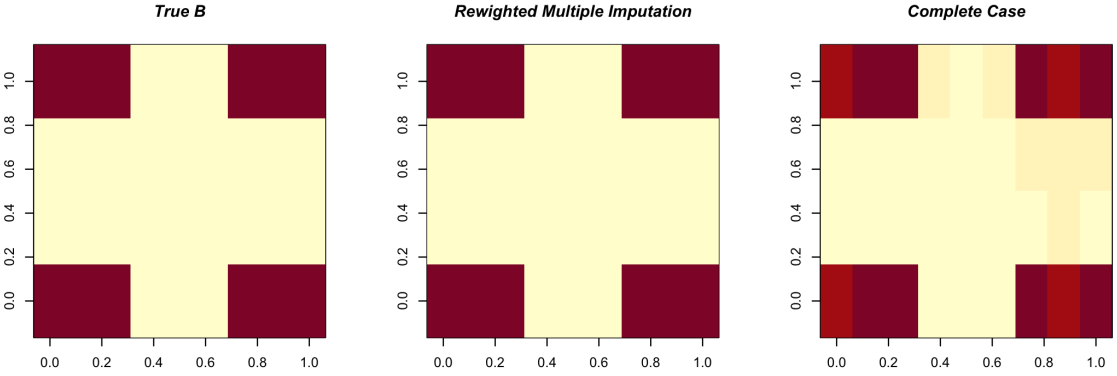

Figure 3.3: Cross Image Example with Missing Response

3.4 Analysis of Cattle Data in Presence of Missingness

To show the superior performance of the proposed estimation approach with multiple imputation discussed in Section 3.1.2, we constructed a simulation study based on a real data set called the Cattle data (Kenward, 1987).

Outside the grazing season, animals are mostly deprived of nutrients, so their immune systems are weaker. Once the grazing season starts in spring, the cattle may ingest roundworm which then causes them to develop diseases. Therefore, to control such diseases 60 animals were randomly assigned to one of the two treatment groups of size 30. The weight of every animal was recorded 11 times throughout the season.

We are interested in estimating the treatment effect on the animals' weight: $\beta = E(Y_i|X_i = 1) - E(Y_i|X_i = 0)$. True parameters μ, β and Σ are computed from the complete cattle data with envelope estimation. Then, data is generated as $Y_i = \mu + \beta X_i + \epsilon$ with $\epsilon \overset{iid}{\sim} t_r(0, \Sigma, \nu)$ where $r = 11$ and degrees of freedom $\nu = 5$. We artificially introduce missingness into the data using Bernoulli distribution with a certain probability p_{miss}. Also, we consider an envelope dimension of $u = 1$ for both missing predictor and missing response cases.

3.4.1 Missing Predictor

First, we introduced missingness into the predictor X_i values for $i = 1, \ldots, 100$ where $X_i = 1$ for treatment group and $X_i = 0$ for control group.

The way the *amelia* function handles categorical variables is that $(p-1)$ binary variables are substituted for a $p-$category categorical variable to fix this variable in a way that the imputations fall into one of the original categories which are 0 and 1 in our case. The *amelia* then pretends these $(p-1)$ dummy variables are like any other variables, so they are imputed as continuous variables. Finally, these imputed values are scaled into probabilities for each category and choosing one of the p categories, the function returns the original $p-$category variable. In our case, the predictor variable is the treatment and in takes only two values, 0 or 1. Therefore, the *amelia* function will substitute this variable with another binary variable, and impute it as a continuous variable and scale these imputed values into probabilities. After that, it randomly selects one category of the two, say the treated $(X = 0)$, scales the probabilities for

this category based on the original categories of the original categorical variable X and returns the reconstructed version of the selected category.

The goal in this section is to show the performance improvement in estimating the coefficient matrix when the reweighted multiple imputation approach is applied compare to the other methods. We first introduce missingness into the predictor with $R_{ij} \overset{iid}{\sim} Bernoulli(p_{miss})$ where $p_{miss} = 0.1$. We report the comparison of the average Δ where $\Delta = \|P_B - P_{\hat{B}}\|_F$ below.

	Full	Reweighted MI	Complete Case	Mean Imputation
$ave(\Delta)$	0.292	0.294	0.316	0.320
SE_Δ	0.0083	0.0092	0.0094	0.0148

Table 3.7: Comparison for Cattle data analysis with missing predictor

Table 3.7 shows that the reweighted MI performs very similar to the full model, and its performance is much better than both the complete case analysis and the mean imputation which is another popular imputation technique.

3.4.2 Missing Response

In this section, we changed the missingness set up and introduced missingness only into the response variable, weight $Y_i \in \mathbb{R}^r$ where $r = 11$, and $i = 1, \ldots, 100$. We consider an envelope dimension of $u = 1$. Missing at random missingness mechanism is created for the response vector. In order to apply that, one of the following is randomly selected: $(Y_{i,2}, Y_{i,4})$, $(Y_{i,3}, Y_{i,9}, Y_{i,11})$, $(Y_{i,7}, Y_{i,8})$, $(Y_{i,1}, Y_{i,10})$, $(Y_{i,5}, Y_{i,6})$ and missingness is assigned for each subject i for $i = 1, \ldots, n$ for the selected part of the response with $Bernoulli(p_{miss})$ where $p_{miss} = 0.3$. The response vector ended up having around 5% to 10% missingness. We compared average Δ where $\Delta = \|P_B - P_{\hat{B}}\|_F$.

	Full	Reweighted MI	Complete Case	Mean Imputation
$ave(\Delta)$	0.292	0.294	0.383	0.430
SE_Δ	0.0083	0.0080	0.0193	0.0303

Table 3.8: Comparison for Cattle data analysis with missing response

The reweighted multiple imputation remains the best method in terms of closeness to the full data analysis, which is followed by complete case and mean imputation methods. However, the difference between reweighted MI and the two alternatives is substantial.

The details of how multiple imputation in this section is executed through bootstrapped EM algorithm are provided in the appendix.

CHAPTER 4

DISCUSSION

First of all, this dissertation offers a method to estimate regression tensor coefficient in presence of a tensor predictor. The proposed estimation method allows the error to have a more general distribution family than Normal distribution; namely elliptically contoured distributions. The underlying envelope structure links the coefficient matrix and the error covariance matrix. This approach is based upon envelopes proposed originally by Cook et al. (2010) as discussed earlier in this dissertation. The proposed estimation method focuses on the model (1.1.5) and suggests that using the iterative reweighting algorithm, regression coefficients can be estimated more precisely and efficiently compare to other available methods.

Secondly, we suggest that the reweighting algorithm can be applied to multiple imputation, a popular approach that can be used in the presence of missing values, to impute missing values while accounting for uncertainties. In this dissertation, multiple imputation is applied through the bootstrapped EM algorithm. Considering the same multivariate linear regression model with tensor predictors as in the first part of this dissertation, we allowed error to have a more varying distribution than the standard normal. We showed that even though the way missing data was imputed was not adjusted for elliptical distributions or the envelope setting, reweighting the

imputed data while estimating the regression coefficient clearly improved the resulting estimate and made it perform very close the the best case possible, which we call full data analysis through the Chapter 3. This framework can also be further enhanced by adjusting the imputation step for elliptical distributions under the envelope structure. This part is left as a topic of future research.

BIBLIOGRAPHY

[1] Chen, L. S., Prentice, R. L., and Wang, P. (2014). A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. *Biometrics*, *70*(2), 312-322.

[2] Cook, R. D., Helland, I. S., and Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(5), 851–877.

[3] Cook, R. D. (2018). *An introduction to envelopes: dimension reduction for efficient estimation in multivariate statistics*. New York: Wiley.

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

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

[6] Cook, R. D., and Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association*, *89*(426), 592–599.

[7] Cook, R. D., and Zhang, X. (2015). Simultaneous envelopes for multivariate linear regression. *Technometrics*, *57*(1), 11–25.

[8] Cook, R. D., and Zhang, X. (2016). Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics*, *25*(1), 284–300.

[9] De Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *18*(3), 251–263.

[10] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

[11] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1980). Iteratively reweighted least squares for linear regression when errors are normal/ independent distributed. *Multivariate Analysis V*, 35–57.

[12] Díaz-García, José A and Galea Rojas, Manuel and Leiva-Sánchez, Víctor (2003). Influence diagnostics for elliptical multivariate linear regression models. *Communications in Statistics-Theory and Methods*, *32*(3), 625–641.

[13] Díaz-García, José A and Gutiérrez-Jáimez, Ramón (2007). The distribution of residuals from a general elliptical linear model. *Journal of Statistical Planning and Inference*, *137*(7), 2347–2354.

[14] Embrechts, P and Klüppelberg, C and Mikosch, T (2003). Measuring extremal events for insurance and finance.

[15] Fang, K. T., Kotz, S., and Ng, K. W. (2018). *Symmetric multivariate and related distributions.* Chapman and Hall/CRC

[16] Forzani, L. and Su, Z. (2019). Envelopes for elliptical multivariate linear regression. *Statistica Sinica*, *31*(2021), 301–332.

[17] Frahm, G. (2004). *Generalized elliptical distributions: theory and applications* (Doctoral dissertation, Universität zu Köln).

[18] Galea, M., Riquelme, M., and Paula, G. A. (2000). Diagnostic methods in elliptical linear regression models. *Brazilian Journal of Probability and Statistics*, 167–184.

[19] Honaker, J., and King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, *54*(2), 561–581.

[20] Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7), 1–47.

[21] Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

[22] Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. Sankhyā: *The Indian Journal of Statistics, Series A*, 419–430.

[23] Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *36*(3), 296–308.

[24] Kolda, T. G. (2006). *Multilinear operators for higher-order decompositions*. Technical report, Sandia National Laboratories.

[25] Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*(3), 455–500.

[26] Lemonte, A. J. and Patriota, A. G. (2011). Multivariate elliptical models with general parameterization. *Statistical Methodology*, *8*(4), 389–400.

[27] Little, R. J., and Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley Sons.

[28] McLachlan, G. J. and Krishnan, T. (2007). *The EM Algorithm and Extensions* (Vol. 382). John Wiley Sons.

[29] Melo, T. F., Ferrari, S. L., and Patriota, A. G. (2017). Improved hypothesis testing in a general multivariate elliptical model. *Journal of Statistical Computation and Simulation*, *87*(7), 1416–1428.

[30] Melo, T. F., Ferrari, S. L., and Patriota, A. G. (2018). Improved estimation in a general multivariate elliptical model. *Brazilian Journal of Probability and Statistics*, *32*(1), 44–68.

[31] Resnick, S. I. (2013). *Extreme values, regular variation and point processes.* Springer.

[32] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

[33] Rubin. D. B. (1977). The design of a general and flexible system for handling non-response in sample surveys. Manuscript prepared for the US. Social Security Administration, July 1, 1977.

[34] Rubin, D. B. (1987). *Multiple Imputation for Survey Nonresponse.* John Wiley Sons. http://doi.org/10.1002/9780470316696

[35] Tucker, L. R. (1963). Implications of factor analysis of three-way matrices for measurement of change. *Problems in Measuring Change*, *15*, 122–137.

[36] Wei, B.-C., Hu, Y.-Q., Fung, W.-K. (1998). Generalized leverage and its applications. *Scandinavian Journal of Statistics*, *25*(1), 25–37.

[37] Zhang, P. (2003). Multiple imputation: Theory and Method. *International Statistical Review*, *71*(3), 581–592.

[38] Zhang, X. and Li, L. (2017). Tensor envelope partial least-squares regression. *Technometrics*, *59*(4), 426–436.

[39] Zhao, Q., Caiafa, C. F., Mandic, D. P., Chao, Z. C., Nagasaka, Y., Fujii, N., Zhang, L., and Cichocki, A. (2012). Higher order partial least squares (hopls):

A Generalized multilinear regression method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(7), 1660–1673.

[40] Zhong, W., Xing, X., and Suslick, K. (2015). Tensor sufficient dimension reduction. *Wiley Interdisciplinary Reviews: Computational Statistics*, *7*(3), 178–184.

[41] Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, *108*(502), 540–552.

APPENDIX

Bootstrapped EM Algorithm used in Section 3.4.2

We followed the algorithm give by McLachlan and Krishnan (2007) to execute EM algorithm to impute missing values as part of the multiple imputation. This particular approach benefits from the simplicity of the least squares method. Steps that are followed in Section 3.4 are as follows.

Step 1: Bootstrap (re-sample with replacement) the data with missing values.

Step 2: Select initial values for each missing values. Mean of the observed data can be used as the initial value.

Step 3: Compute least square estimates based on the completed data.

Step 4: Get predictions with these LS estimates for the missing data above.

Step 5: Go to step 3, and continue until missing values or the residual sum of squares converged.

Details of EM Algorithm used in "Amelia"

Having data suffer from missingness, we would need to use observed data likelihood rather than the full data likelihood, which then requires handling multivariate inte-

grals. The integrals in the likelihood function of observed data can get really complex. Therefore, EM algorithm is a great option to compute maximum likelihood estimates. On the other hand, a closed form of the likelihood can be derived when data D follows normal distribution and make the estimation computationally more manageable. Thus, throughout the Chapter 3, we assumed that the data was normally distributed.

In the most general sense, EM algorithm would apply as follows. Let θ denote parameter vector of the data matrix D. Then, $l(\theta|D) = logL(\theta|D)$ is the log-likelihood function. To get the maximum likelihood estimate $\hat{\theta}$, the $E - step$ and $M - step$ are iterated until convergence is achieved.

$E - Step$: In this step, the missing values are filled in using the current initial estimate of the parameter. Let $\theta^{(t)}$ be the initial value of θ. Then, the expectation of likelihood can be evaluated as

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E\{l(\theta|D)|D_{obs}, \theta^{(t)}\} \\ &= \int l(\theta|D) f(D_{mis}|D_{obs}, \theta^{(t)}) dD_{mis} \end{aligned}$$

$M - Step$: This is the maximization step in which a new parameter estimate is obtained by maximizing the expectation in the $E - step$ that will then be used as the updated expectation. In this step, the expected log likelihood function computed in the $E - step$ is maximized and $\theta^{(t+1)}$ is computed.

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geqslant Q(\theta|\theta^{(t)})$$

Nevertheless, getting parameter estimates based on envelope model would require reparameterization of the covariance matrix Σ which would make envelope estimation through EM-algorithm complicated. Therefore, we will leave such an adjustment of envelope model parameters for future research.