2016

Dimension Reduction and Variable Selection

Hossein Moradi Rekabdarkolaee
Virginia Commonwealth University

Dimension Reduction and Variable Selection

Hossein Moradi Rekabdarkolaee

Dissertation submitted to the Faculty of the

Virginia Commonwealth University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Systems Modeling and Analysis

Qin Wang, Chair

Edward L. Boone

David J. Edwards

Jenise L. Swall

Ali Arab

December 1, 2016

Richmond, Virginia

Abstract

High-dimensional data are becoming increasingly available as data collection technology advances. Over the last decade, significant developments have been taking place in high-dimensional data analysis, driven primarily by a wide range of applications in many fields such as genomics, signal processing, and environmental studies. Statistical techniques such as dimension reduction and variable selection play important roles in high dimensional data analysis. Sufficient dimension reduction provides a way to find the reduced space of the original space without a parametric model. This method has been widely applied in many scientific fields such as genetics, brain imaging analysis, econometrics, environmental sciences, etc. in recent years.

In this dissertation, we worked on three projects. The first one combines local modal regression and Minimum Average Variance Estimation (MAVE) to introduce a robust dimension reduction approach. In addition to being robust to outliers or heavy tailed distribution, our proposed method has the same convergence rate as the original MAVE. Furthermore, we combine local modal base MAVE with a L_1 penalty to select informative covariates in a regression setting. This new approach can exhaustively estimate directions in the regression mean function and select informative covariates simultaneously, while being robust to the existence of possible outliers in the dependent variable. The second project develops sparse adaptive MAVE (saMAVE). SaMAVE has advantages over adaptive LASSO because it extends adaptive LASSO to multi-dimensional and nonlinear settings, without any model assumption, and has advantages over sparse inverse dimension reduction methods in that it does not require

any particular probability distribution on $\mathbf{X}$. In addition, saMAVE can exhaustively estimate the dimensions in the conditional mean function. The third project extends the envelope method to multivariate spatial data. The envelope technique is a new version of the classical multivariate linear model. The estimator from envelope asymptotically has less variation compare to the Maximum Likelihood Estimator (MLE). The current envelope methodology is for independent observations. While the assumption of independence is convenient, this does not address the additional complication associated with a spatial correlation. This work extends the idea of the envelope method to cases where independence is an unreasonable assumption, specifically multivariate data from spatially correlated process. This novel approach provides estimates for the parameters of interest with smaller variance compared to maximum likelihood estimator while still being able to capture the spatial structure in the data.

Dedication

To my dearest Zahra, the greatest parents on earth Hamidreza and Sedigh, and the apples of my eye Maryam and Maedeh.....

Acknowledgments

"If you are not capable of progressing, be similar to an apple, so your fall raises ideas."

This dissertation would not have been possible without the encouragement, help, and support of my family and friends. My deepest gratitude to all of you.

My special thanks goes to my advisor, Dr. Qin Wang, for his constant guidance, generous help, and encouragement. I feel really lucky to be trained by such a great master in Statistics. The research work in this dissertation was driven forward by his input and timely feedback through numerous meetings, discussions, and email correspondence. I would like to thank Dr. Edward Boone for his emphasis on finding the story in each set of data rather than just focusing on the mathematics. I would also like to thank the members of my committee: Drs. David Edwards, Jenise Swall, and Ali Arab for their insightful comments and suggestions.

I would like to thank the faculty and staff in the department of Statistical Sciences and Operations Research. In particular, thanks to Drs. Paul Brooks, Edward Boone, Cheng Ly, and Angela Reynolds for recruiting me as a Ph.D. student. I am also grateful to Dr. D'Arcy P. Mays for being the most amazing chair of the department and for his help with job applications.

Again, I want to thank my parents, my wife, and my friends. I find words can hardly describe their incredible impact on my life. Without their support and love, it would be meaningless and certainly impossible for me to finish this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

1.1 Introduction

Regression analysis is a standard statistical approach to study the relationship between a univariate response variable, $y \in \mathbb{R}$, and a set of p explanatory variables $\mathbf{X} \in \mathbb{R}^p$. Two important goals for this method are: first finding relevant explanatory variables and second having high prediction accuracy (Zou, 2006). To achieve these goals, one must estimate the regression function that best describes the relationship between dependent and explanatory variables. Without any prior knowledge of the form of the relationship, the regression function is $g(x) = E(y|\mathbf{X} = \mathbf{x})$ is often estimated nonparametrically.

With the technological advancements in last decade, it is much easier to collect information on a large pool of variables. Due to the well-known *"Curse of dimensionality"* (Bellman, 1961) analyzing these types of datasets is challenging. The *curse of dimensionality* refers to various phenomena that arise when analyzing and organizing data in high-dimensional

spaces. One of the common themes of this problem is that when the dimensionality increases, the sample size required to make inferences will increase exponentially.

Since the pioneering work of Li (1991), Sufficient Dimension Reduction (SDR) has received considerable attention as an efficient tool for analyzing high dimensional data. The basic idea of SDR is to replace the original high dimensional predictor with an appropriate low dimensional projection without losing regression information (Cook, 1998). The goal of SDR is to find a subspace S of the predictor space such that

$$y \perp\!\!\!\perp \mathbf{X} | P_S \mathbf{X}, \tag{1.1}$$

where $\perp\!\!\!\perp$ denotes independence and $P_{(.)}$ represents an orthogonal projection operator with respect to the standard inner product. Thus, if $d = \dim(S)$ and $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ is a basis for S, the predictor $\mathbf{X}$ can be replaced by the linear combinations $\boldsymbol{\beta}_1^T \mathbf{X}, \ldots, \boldsymbol{\beta}_d^T \mathbf{X}$, often $d << p$, without losing regression relationship information. When the intersection of all subspaces, S, satisfying (1.1) also satisfies (1.1), it is called the central subspace (CS; Cook, 1994) and is denoted by $S_{y|\mathbf{X}}$. When our primary interest is the conditional mean function i.e. $g(x) = E[y|\mathbf{X}]$, the objective of sufficient dimension reduction is to find a d-dimensional subspace S such that

$$y \perp\!\!\!\perp E(y|\mathbf{X}) | P_S \mathbf{X}. \tag{1.2}$$

Subspaces satisfying condition (1.2) are called mean dimension reduction subspaces (Cook and Li, 2002). When the intersection of all subspaces satisfying condition (1.2) also satisfies condition (1.2), it is called the central mean subspace (CMS) and is denoted by $S_{E(y|\mathbf{X})}$. As shown in Cook (1998) and Yin *et al.* (2008), under mild conditions, the CS and the CMS exist and are unique. Knowledge of the CS or the CMS is very useful for parsimoniously

characterizing the conditional distribution of $Y|\mathbf{X}$ or $E(Y|\mathbf{X})$. In other words, SDR provides an effective starting point for the regression. Based on their results, we assume the existence of the CS and the CMS throughout the study.

1.2 Literature review

1.2.1 Dimension reduction

Principal components analysis (PCA) is a general method for the reduction of multivariate observations (Adcock, 1878) to a smaller subspace. PCA was established as the first reductive method for regression by the mid-1900s. While PCA seems to be the dominant method of dimension reduction across the applied sciences, there are many other well established and recent statistical methods that might be used to address large-p regressions, including factor analysis (Fruchter, 1954), Inverse Regression Estimation (IRE; Cook and Ni, 2005), Partial Least Squares (PLS; Wold, 1985), projection pursuit (Friedman and Stuetzle, 1981; Huber, 1985), seeded reductions (Cook *et al.*, 2007), kernel methods (Fukumizu *et al.*, 2009) and sparse methods that are based on penalization.

Cook (2007) defined sufficient reduction as follows:

Definition 1: A reduction $R : \mathbb{R}^P \to \mathbb{R}^q$, $q \leq p$ is sufficient if it satisfies one of the following three statements:

1. Inverse approach $\mathbf{X}|(Y, R(\mathbf{X})) \sim \mathbf{X}|R(\mathbf{X})$,

2. Forward approach, $Y|\mathbf{X} \sim Y|R(\mathbf{X})$,

3. Joint approach, $\mathbf{X} \perp\!\!\!\perp Y|R(\mathbf{X})$,

where $\perp\!\!\!\perp$ indicates independence, $\sim$ means identically distributed and $\mathbf{A}|\mathbf{B}$ refers to the random vector $\mathbf{A}$ given the vector $\mathbf{B}$.

If we assume $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ with $d \leq p$ is the reduction subspace then so is $\text{span}(\mathbf{B}_0, \mathbf{B})$ for any $p \times q$ matrix $\mathbf{B}_0$. If $\text{span}(\mathbf{B}_0)$ and $\text{span}(\mathbf{B}_1)$ are both dimension-reduction subspaces, then under mild conditions so is their intersection i.e. $span(\mathbf{B}_0) \cap span(\mathbf{B}_1)$ (Cook, 1996 and 2009). Consequently, the inferential target in sufficient dimension reduction is often taken to be the central subspace $S_{Y|\mathbf{X}}$, defined as the intersection of all dimension-reduction subspaces (Cook, 1994; 1996, and 2009). The two major approaches of sufficient dimension reduction and finding the central subspace are forward approach and inverse approach (Adragni and Cook, 2009).

1.2.1.1 Inverse approach

Inverse regression deals with the (inverse) conditional distribution of $\mathbf{X}|Y$ i.e. $F_{\mathbf{X}|Y}$. Inverse reduction based sufficient dimension reduction methods provide estimates of the minimal sufficient linear reduction. Sliced Inverse Regression (SIR; Li, 1991) and Sliced Average Variance Estimation (SAVE; Cook and Weisberg, 1991) were the first methods proposed for dimension reduction using inverse regression. These methods can estimate the central subspace under two key conditions: linearity and constant covariance. Both SIR and SAVE provide $\sqrt{n}$ consistent estimators of central subspace under certain regularity conditions, but by itself consistency does not guarantee good performance in practice.

SIR has difficulty finding directions that are associated with certain types of nonlinear trends in $E(Y|\mathbf{X})$. For instance, SIR misses the directions when the dependence between y

and $\mathbf{X}$ is symmetric. SAVE was developed to address and solve this limitation, but its ability to find linear trends is generally less than SIR's (Adragni and Cook, 2009). Several methods have been developed in an effort to improve on the estimates of the central subspace provided by SIR and SAVE. Cook and Ni (2005) developed an asymptotically optimal method of estimating $S_{Y|\mathbf{X}}$ called Inverse Regression Estimation (IRE). Ye and Weiss (2003) and Zhu *et al.* (2007) attempted to combine the advantages of SIR and SAVE by using linear combinations of them. Cook and Forzani (2009) used a likelihood-based objective function to develop a method called Likelihood Acquired Directions (LAD) that is based on the same population foundations as SIR and SAVE. These methods have been developed and studied mostly in regressions where $p << n$, although there are some results for other settings (Li, 2007; Li and Yin, 2008). SIR, SAVE, IRE and LAD come with a range of inference capabilities, including methods for estimating d and tests of conditional independence hypotheses such as Y is independent of $\mathbf{X}_1$ given $\mathbf{X}_2$, where we have partitioned $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T$. Some other examples of this approach are Principal Hessian Directions (PHD; Li, 1992) and Contour Regression (CR; Li and *et al.*, 2005).

1.2.1.2 Forward approach

Forward regression methods study the conditional distribution of $Y|\mathbf{X}$, $F_{Y|\mathbf{X}}$. Examples for this approach are Ordinary Least Squares (OLS; Li and Duan, 1985), Average Direction Estimation (ADE; Hardle and Stoker, 1989; Samarov, 1993), Structure Adaptive Method (SAM; Hristache *et al.*, 2001), Minimum Average Variance Estimation (MAVE; Xia *et al.*, 2002), Fourier methods (FM; Zhu and Zeng 2006), and Sliced Regression (SR; Wang and Xia, 2008).

Minimum Average Variance Estimator (MAVE) is one of the most popular methods for dimension reduction in forward regression setting. MAVE is an adaptive approach based on semiparametric models, which combines the projection pursuit regression and the local linear smoothing nicely. The regression-type model of interest in MAVE can be written as

$$y = g(\mathbf{B}_0^T\mathbf{X}) + \epsilon, \tag{1.3}$$

where $g(\cdot)$ is an unknown smooth link function, $\mathbf{B}_0 = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D)$ is an orthogonal matrix ($\mathbf{B}_0^T\mathbf{B}_0 = \mathbf{I}_D$) with the structural dimension $D < p$ and $E(\epsilon|\mathbf{X}) = 0$. Following the idea of local linear smoothing, Xia *et al.* (2002) proposed MAVE such that the parameter $\mathbf{B}_0$ can be estimated by minimizing the following objective function

$$\mathrm{E}\{y - \mathrm{E}(y|\mathbf{B}^T\mathbf{X})\}^2 = \mathrm{E}\{y - g(\mathbf{B}^T\mathbf{X})\}^2 = \mathrm{E}\{\sigma^2_{\mathbf{B}}(\mathbf{B}^T\mathbf{X})\}, \tag{1.4}$$

where $\sigma^2_{\mathbf{B}}(\mathbf{B}^T\mathbf{X}) = \mathrm{E}[\{y - g(\mathbf{B}^T\mathbf{X})\}^2 \mid \mathbf{B}^T\mathbf{X}]$ is the conditional variance and $\mathbf{B}^T\mathbf{B} = \mathbf{I}_d$ for identifiability. Let $\{(y_i, \mathbf{X}_i), 1, \ldots, n\}$ be a random sample from $(y, \mathbf{X})$ according to (3.1). For any given $\mathbf{X}_0$ and $\mathbf{X}_i$ close to $\mathbf{X}_0$, a local linear approximation gives

$$\begin{aligned} y_i - g(\mathbf{B}^T\mathbf{X}_i) &\approx y_i - g(\mathbf{B}^T\mathbf{X}_0) - \{\bigtriangledown g(\mathbf{B}^T\mathbf{X}_0)\}^T\mathbf{B}^T(\mathbf{X}_i - \mathbf{X}_0) \\ &\equiv y_i - a - \mathbf{b}^T\mathbf{B}^T(\mathbf{X}_i - \mathbf{X}_0), \end{aligned} \tag{1.5}$$

where $\bigtriangledown$ means the first derivative. Thus,

$$\sigma^2_{\mathbf{B}}(\mathbf{B}^T\mathbf{X}_0) = \min_{a,\mathbf{b}} \sum_{i=1}^{n} \{y_i - a - \mathbf{b}^T\mathbf{B}^T(\mathbf{X}_i - \mathbf{X}_0)\}^2 w_{i0}, \tag{1.6}$$

where $w_{i0} \geq 0$ are kernel weights with $\sum_{i=1}^{n} w_{i0} = 1$. From (1.4) and (1.6), an estimate of the $\mathbf{B}$ is the solution of

$$\min_{\mathbf{B},a_j,\mathbf{b}_j;j=1,\ldots,n} \sum_{j=1}^{n}\sum_{i=1}^{n} \left[y_i - \left\{a_j + \mathbf{b}_j^T\mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_j\right)\right\}\right]^2 w_{ij}. \tag{1.7}$$

where $\boldsymbol{\theta}$ =$\{\mathbf{B}$,$(a_j$,$\mathbf{b}_j), j = 1, 2, \ldots, n\}$, $a_j \in R$, $\mathbf{b}_j \in R^d$ and $\mathbf{B}^T\mathbf{B} = \mathbf{I}_d$. Furthermore, w_{ij} are kernel weights defined as a function of the distance between $\mathbf{X}_i$ and $\mathbf{X}_j$, i.e. $w_{ij} = \frac{K_h(\mathbf{X}_i - \mathbf{X}_j)}{\sum_{l=1}^{n} K_h(\mathbf{X}_l - \mathbf{X}_j)}$ where $K_h(\cdot)$ is a multidimensional kernel function and h refers to the bandwidth. This minimization can be solved iteratively with respect to $\{(a_j, \mathbf{b}_j), j = 1, \ldots, n\}$ and $\mathbf{B}$ separately. The estimation of MAVE is very efficient since only two quadratic programming problems are involved and both have explicit solutions. To improve the estimation accuracy, a lower dimensional kernel weight $\tilde{w}_{ij}$ as a function of $\tilde{\mathbf{B}}^T(\mathbf{X}_i - \mathbf{X}_j)$ can be used after an initial estimate $\tilde{\mathbf{B}}$ was obtained. The use of a smaller bandwidth in the refined procedure can also improve the consistency rate. More details can be found in the Xia *et al.* (2002) and the references therein.

Wang and Yin (2008) considered sufficient dimension reduction and variable selection on the mean function $E(Y|\mathbf{X})$ only. Focusing on the central mean subspace that was introduced by Cook and Li (2002), they combined MAVE with a LASSO variable selection method (1996) and propose a new dimension reduction and variable selection method, sparse MAVE (SMAVE). SMAVE has advantages over LASSO because it extends lasso to multidimensional and nonlinear settings, without any model assumption. Furthermore, it has advantages over sparse inverse dimension reduction methods introduced by Li (2007) in that is it does not require any particular distribution on $\mathbf{X}$ and it can exhaustively estimate the dimensions in the conditional mean function.

Despite its popularity in dimension reduction, MAVE is not robust under heavy tailed error distributions and/or outliers because it uses least squares criterion. Čížek and Härdle (2006) gave a comprehensive study of the sensitivity of MAVE to outliers and proposed a robust enhancement to MAVE by replacing the local least squares with local L- or M-estimation. Yao and Wang (2013) extended the robust estimation to variable selection and proposed a robust sparse MAVE. Wang and Yao (2012) introduced an adaptive estimation for MAVE (aMAVE) which combines the kernel density estimation and MAVE that can adapt to different error distributions.

1.2.2 Variable selection

When the number of covariates is large, it is reasonable to expect only some of the explanatory variables to be relevant to predict the response variable (Yao and Wang, 2013). SDR provides a way to find sufficient dimensions without the need for a parametric model. However, each reduced variable is a linear combination of all of the original variables. Therefore, these reduced variables are difficult to interpret. As a result, variable selection is very important not only for better model interpretation, but also for higher prediction accuracy (Zou, 2006). Traditionally, variable selection is performed using an information criterion such as Akaike information criterion (AIC., Akike, 1973), Bayesian information criterion (BIC., Schwarz, 1978), etc. These criterion measure the quality of a statistical model by penalizing the model if unimportant variables are added. However, these traditional variable selection methods may suffer from instability with respect to small changes in the data set because of their inherent discreteness (Breiman, 1995). In order to solve this problem, a number of regularization approaches such as Nonnegative Garrote (Breiman, 1995 and 1996), Least

Absolute Shrinkage and Selection Operator (LASSO; Tibshirani, 1996), Smoothly Clipped Absolute Deviation (SCAD; Fan and Li, 2001), Least Angle Regression (LARS; Efron *et al.*, 2004), and Elastic Net (Zou and Hastie, 2005) were proposed to automatically select informative variables through continuous shrinkage.

Breiman (1995) introduced regression models under direct influence of non-negative garrote which is a combination between the ridge regression and best subset selection method. Instead of using normal equations, the following loss function was proposed to estimate the regression coefficients:

$$\begin{aligned} &\min \ \sum_{i=1}^{n}\left(y_i - \sum_j c_j \mathbf{x}_{ij}\hat{\beta}_j\right)^2 \\ &\textit{subject to} \\ &c_j \geq 0, \ \sum_j c_j \leq t. \end{aligned} \tag{1.8}$$

where $\hat{\beta}_j$s are the regression coefficients. Least Absolute Shrinkage and Selection Operator (LASSO) was introduced by Tibshirani (1996)

$$\min \ \sum_{i=1}^{n}\left(y_i - \sum_j \mathbf{x}_{ij}\beta_j\right)^2 + \lambda \sum_j |\beta_j|_1, \tag{1.9}$$

where $|\cdot|_1$ is the L_1 norm. LASSO solution can be viewed as the Bayesian *maximum posteriori estimation* when parameters are apriori independent from each other and each parameter has a double exponential prior distribution. LASSO continuously shrinks the coefficients toward 0 as λ increases and some coefficients are shrunk to exactly 0 if λ is large enough. Asymptotic performance for LASSO-type estimators was study by Knight and Fu (2000). At the time that LASSO was introduced, there was not a lot of interest in using this method because the computational resources were lacking compared to today and large

data problems (in n, p or both) were rare (Tibshirani, 2011). LASSO is useful for fitting a wide variety of models such as regression (Tibshirani, 1996), generalized linear model (Guisan *et al.*, 2002; Van de Geer, 2008; Zhang and Huang, 2008), classification (Ghosh and Chinnaiyan, 2005), spatial filtering (Seya *et al.*, 2015), etc. Newly developed computational algorithms allow application of these models to large data sets, exploiting sparsity for both statistical and computation gains (Wang *et al.*, 2007). Meinshausen and Buhlmann (2006) showed variable selection with LASSO can be consistent if the underlying model satisfies some conditions. They also showed the conflict of optimal prediction and consistent variable selection in LASSO. Meinshausen and Buhlmann (2006) proved the optimal λ for prediction gives inconsistent variable selection results. This conflict can become more understandable by considering an orthogonal design model (Leng *et al.*, 2006). Zou (2006) proved selecting variables via LASSO could be inconsistent and proposed an adaptive method to achieve consistent estimation. The adaptive LASSO solves the following minimization problem:

$$\min \sum_{i=1}^{N} \left(y_i - \sum_j \mathbf{x}_{ij} \hat{\beta}_j \right)^2 + \lambda \sum_j w_j |\beta_j|, \tag{1.10}$$

where w_j are known weights.

Fan and Li (2001) introduced another method for variable selection called Smoothly Clipped Absolute Deviation (SCAD). SCAD penalties are non-convex and this non-convexity is necessary for unbiasedness of estimated coefficients. Fan and Li (2001) and Fan and Peng (2004) introduced oracle procedure for variable selection. The oracle property means that the penalized estimator is asymptotically equivalent to the Oracle estimator that is the ideal estimator obtained only with independent variables without penalization. Fan and Li (2001) and Fan and Peng (2004) argued any good procedure should have oracle properties. Candes

(2006) gave a comprehensive summary on how to perform statistical estimation via oracle inequalities. Zou (2006) showed if the weights are data-dependent, then the adaptive LASSO can have the oracle properties.

The L_1 loss does not distinguish the source of coefficients and treats all the coefficients equally, no matter whether they correspond to the same variable or different variables, or they are more likely to be relevant or irrelevant. Furthermore, L_1 loss is efficient when we have error with heavy tailed distribution and/or outliers, but it loses its efficiency when the data are normally distributed. Zhang *et al.* (2008), proposed a new technique for more effective variable selection in Multicategory Support Vector Machine (MSVM) using L_∞ loss instead of L_1. In contrast to the L_1 loss, which imposes a penalty on the sum of absolute values of all coefficients, MSVM penalizes the sup-norm of the coefficients associated with each variable. Moreover, MSVM studied if the sup-norm approach encourages more sparse solutions than the L_1, and identifies important variables more precisely. This is because with a sup-norm, a noise variable will be removed if and only if all corresponding estimated coefficients are 0. On the other hand, unlike the L_1 penalty, if a variable is important sup-norm penalty does not put any additional penalties on the other coefficients.

Another method for the variable selection in linear model is Dantsing selector (Candes and Tao, 2007). This method establishes optimal L_2 norm properties under a sparsity scenario when number of covariates, p, is much larger than sample size, n. Bickel *et al.* (2009) showed that under a sparsity scenario, LASSO and Dantzing selector have similar behavior for both linear and nonparametric regression models, for L_2 prediction loss and for L_p loss in the coefficients. Koltchinskii (2009) studied the sparsity and oracle properties of Dantzing selector and found general oracle inequalities for the Dantzig selector.

Usually error with heavy tailed distribution and/or outliers cause difficulties in statistical analysis. A common method is to use a robust regression with a L_1 norm weight (Fan *et al.*, 2014). Fan *et al.*, (2014) introduced Weighted Robust LASSO (WR-LASSO) which combines the penalized quantile regression with the weighted L_1 penalty for robust regularization by considering the following weighted L_1-regularized quantile regression:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\sum_i^n \rho_\tau(y_i-\mathbf{x}_i^T\boldsymbol{\beta})+n\lambda_n|\mathbf{d}\circ\boldsymbol{\beta}|_1\right\},$$

where $\rho_\tau(u)=u(\tau-1\{u\leq 0\})$ is the quantile loss function, $\mathbf{d}=(d_1,\ldots,d_p)^T$ is the vector of nonnegative weights, $\circ$ is the Hadamard product, that is, the component-wise product of two vectors, and $\lambda_n\geq 0$ is a regularization parameter. The weights are used to reduce the bias induced by the L_1 penalty. Flexibility of the choice of the weights provides flexibility in shrinkage estimation of the regression coefficient. WR-LASSO is very similar to the folded-concave penalized quantile-regression which was introduced by Zou (2008) and Wang *et al.* (2012). The main difference between WR-LASSO and other variable selection methods is that it avoids the non-convex optimization problem. Fan *et al.* (2014) establish conditions on the error distribution in order to successfully recover the true underlying sparse model with an asymptotic probability of one. The required condition for this model is weaker than the sub-Gaussian assumption in Bradic *et al.* (2011). A random variable $x\in\mathbb{R}$ is subgaussian if for some $b>0$ and every $t\in\mathbb{R}$: $E[e^{tx}]\leq e^{\frac{b^2t^2}{2}}$. The only conditions imposed are that the density function of error should have the Lipschitz property in a neighborhood around 0. This includes a large class of heavy-tailed distributions such as the stable distributions, including the Cauchy distribution. It also covers the double exponential distribution whose density function is non-differentiable at the origin.

Due to the penalized nature of the penalized least-square estimator, the resulting estimation of WR-LASSO is biased. In order to reduce the bias, the weights need to be chosen adaptively according to the magnitudes of the unknown true regression coefficients, which makes the bias reduction infeasible for practical applications. To make the bias reduction feasible, Fan *et al.* (2014) introduce the adaptive robust LASSO (AR-LASSO). AR-LASSO first runs robust LASSO to obtain an initial estimate, and then computes the weight vector of the weighted L_1 penalty according to a decreasing function of the magnitude of the initial estimate. After that, adaptive robust LASSO runs weighted robust LASSO with the new computed weights. Fan *et al.* (2014) showed the oracle property of AR-LASSO with no assumptions on the distribution of the error and established the asymptotic normality of the AR-LASSO.

Most of the previous mentioned methods are model-based. Sufficient dimension reduction provides a way to select informative predictors without assuming a model. Recently, Ni, Cook and Tsai (2005) and Li and Nachtsheim (2006) combined the sliced inverse regression estimation and the shrinkage variable selection procedure LASSO to produce sparse dimension reduction directions. Based on these two pioneering works, Li (2007) successfully transformed a common eigen-decomposition problem in the inverse dimension reduction methods into a regression-type optimization problem, and proposed a unified estimation strategy combining dimension reduction and variable selection. Wang and Yin (2008) combined MAVE with LASSO and propose sparse MAVE (SMAVE). Yao and Wang (2013) extended the robust estimation to variable selection and proposed a robust sparse MAVE. Wang *et al.* (2013) proposed penalized MAVE which combined the MAVE and regularization in variable selection.

1.3 The Envelope Approach

In many research areas such as health science (Lave and Seskin, 1973; Liang *et al.* 1992), epidemiology (Lekkou *et al.* 2014), business (Cooper *et al.* 2003), etc. it is common to observe multiple outcomes. The traditional multivariate linear regression has proved to be useful in these cases to understand the relationships between response variables and regressors. Mathematically, this model is typically given as:

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}, \tag{1.11}$$

where $\mathbf{Y} \in \mathbb{R}^r$ denotes the response vector, $\mathbf{X} \in \mathbb{R}^p$ is a vector predictor, $\boldsymbol{\alpha} \in \mathbb{R}^r$ denotes vector of intercept coefficients, $\boldsymbol{\beta} \in \mathbb{R}^{(r\times p)}$ is the matrix of regression coefficients, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ is an error vector with $\boldsymbol{\Sigma} \geq 0$ being an unknown covariance matrix (Christensen, 2001). In order to completely specify a multivariate linear model, there are r unknown parameters to specify the intercept, $p \times r$ unknown parameters to specify the matrix of regression coefficients, and $\frac{r(r+1)}{2}$ unknown parameters to specify an unstructured covariance matrix. Therefore, one must estimate $r + pr + \frac{r(r+1)}{2}$ parameters which can be large when either r or p or both are large. Therefore, one need a large number of samples to be able to estimate the parameters. The large number of parameters also leads to other problems such as identifiability of the model, instability of the model, and the computational expense to estimate the parameters.

One unique case that may arise in multivariate regression is when some of the regression coefficients are zero for all predictors on a few of the response variables. This means those

responses do not depend on any of the predictors. Mathematically, this model is given as:

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\beta}^* \end{pmatrix} \mathbf{X} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{pmatrix},$$

which means the distribution of $\mathbf{Y}_1$ does not depend on any of the predictors in $\mathbf{X}$. Extending this setting, there are cases where the distribution of some linear combinations of the response vector $\mathbf{Y}$ do not depend on any of the predictors in $\mathbf{X}$ which are called *immaterial* to the regression. The other linear combinations of $\mathbf{Y}$ which their distribution depend on $\mathbf{X}$ are called *material* to the regression. Based on this idea, Cook *et al.* (2010) proposed the *envelope* method as a new version of the classical multivariate linear model. This approach separates the $\mathbf{Y}$ into material and immaterial, thereby allowing gains in efficiency by reducing the variance of the estimate of the parameters of interest compared to the maximum likelihood estimate (Cook *et al.*, 2010). The envelope approach constructs a link between the mean function and covariance matrix using a minimal reducing subspace such that the resulting number of parameters will maximally reduce. Cook *et al.* (2010) showed that the envelope estimator has asymptotically less variation compared to the standard maximum likelihood estimator (MLE).

For model (1.11), suppose that we can find an orthogonal matrix $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{r\times r}$ that satisfies the following two conditions: (i) $span(\boldsymbol{\beta}) \subseteq span(\boldsymbol{\Gamma})$, and (ii) $\boldsymbol{\Gamma}^T\mathbf{Y}$ is conditionally independent of $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ given $\mathbf{X}$. Together, these conditions imply that $\boldsymbol{\Gamma}_0^T\mathbf{Y}$ is marginally independent of $\mathbf{X}$ and conditionally independent of $\mathbf{X}$ given $\boldsymbol{\Gamma}_0^T\mathbf{Y}$. In this setting, we can write $\boldsymbol{\Sigma}$ as follows

$$\boldsymbol{\Sigma} = \mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{P}_{\boldsymbol{\Gamma}} + \mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{Q}_{\boldsymbol{\Gamma}}, \tag{1.12}$$

where $\mathbf{P}_{\boldsymbol{\Gamma}}$ is the projection onto $span(\boldsymbol{\Gamma})$ and $\mathbf{Q}_{\boldsymbol{\Gamma}} = \mathbf{I}_r - \mathbf{P}_{\boldsymbol{\Gamma}}$. Cook *et al.* (2010) used this idea to construct the unique smallest subspace $span(\boldsymbol{\Gamma})$ that satisfies (1.12) and contains $span(\boldsymbol{\beta})$. Therefore, the goal is to find a subspace $\boldsymbol{S} \subseteq \mathbb{R}^r$ such that

$$\mathbf{Q}_{\boldsymbol{S}}\mathbf{Y}|\mathbf{X} \sim \mathbf{Q}_{\boldsymbol{S}}\mathbf{Y}, \tag{1.13a}$$

$$\mathbf{Q}_{\boldsymbol{S}}\mathbf{Y} \perp\!\!\!\perp \mathbf{P}_{\boldsymbol{S}}\mathbf{Y}|\mathbf{X}, \tag{1.13b}$$

where $\mathbf{P}_{(\cdot)}$ represents an orthogonal projection operator with respect to the standard inner product and $\mathbf{Q}_{(\cdot)} = \mathbf{I}_r - \mathbf{P}_{(\cdot)}$. This minimal subspace is called the $\boldsymbol{\Sigma}$-envelope of $span(\boldsymbol{\beta})$ in full and the envelope for brevity. Figure 1.1 provides a graphical display of the envelope model. In both panels, the two ellipses represent two normal populations and we want to find if there is a difference between two populations. The left panel shows the analysis under the standard model (OLS) and the right panel shows the analysis using envelope model. The red and green curves in the left panel stand for the two projected distributions from the two populations. As it can be seen from the right panel, since there is considerable overlap between the two projected distributions, we need a large sample size to infer that these populations are different under the OLS. While using envelope, the left panel, it is obvious that the same inference can be done using much smaller sample size.

Following the envelope idea, equation (4.1) can be rewritten as follows

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X} + \boldsymbol{\epsilon}, \tag{1.14}$$

where $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$, $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$, and $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$ where $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$, and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(r-u)\times(r-u)}$ are unknown positive definite matrices where $0 < u \leq r$ is the dimension of the envelope subspace. Here, one only has to estimate $r + pu + \frac{r(r+1)}{2}$ parameters. The difference

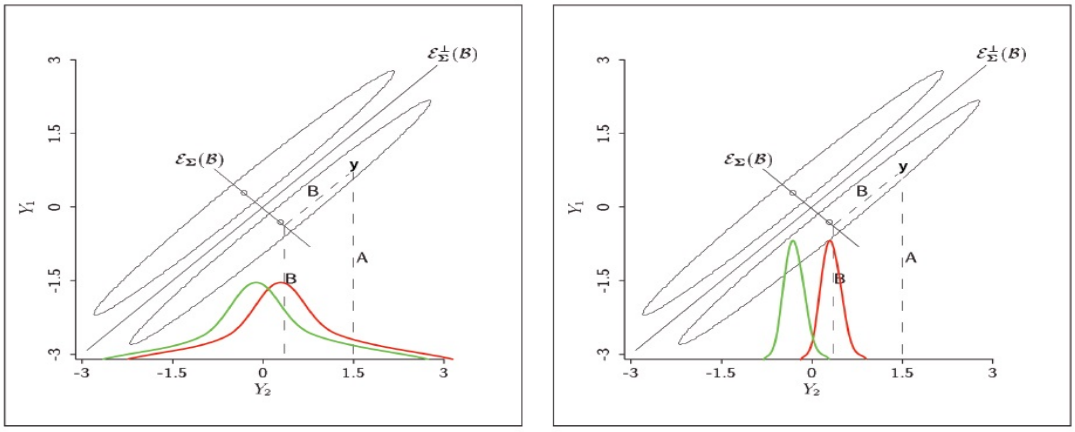

Figure 1.1: *A graphical display of the envelope model.*

in the number of parameters between the envelope and classical multivariate regression is $p(r-u)$ parameters.

Su and Cook (2011) proposed the *partial envelope method* for situations where a set of predictors are of special interest. The goal of the partial envelope is to improve the efficiency of the estimated coefficients corresponding to these particular predictors by partitioning the predictors space into two subspaces where one contains the set of the predictors of interest and another contains the remaining predictors. Since the span of any subsets of the predictor space is often a proper subset of span of the original predictor space, the partial envelope approach leads to a gain in the efficiency of the estimates. Cook,*et al.*, (2013) used the envelope to study predictor reduction in multivariate linear regression and established a connection between the envelope and partial least squares regression. Su and Cook (2013) then adapted the envelope for the estimation of multivariate means with heteroscedastic errors by freeing the constant covariance structure assumption in the original envelope method and proposing a more general covariance structure. Furthermore, Su and Cook (2012) introduced a different type of envelope construction, called *inner envelope*, that can produce efficiency gains when the ordinary envelope offers no gains. This method partitions response space into three sub-

spaces instead of partitioning into two subspaces. The distribution of the first two partition can depend on $\mathbf{X}$ while the distribution of the third projection is independent from $\mathbf{X}$. This generalization comes from the relaxation of the assumptions (1.13a) and (1.13b). Cook *et al.* (2013) built a connection between envelope methodology and partial least squares (PLS), allowing PLS to be addressed in a traditional likelihood-based framework called the *Scaled Predictor Envelopes* (SPE). This approach incorporates predictor scaling into PLS-type applications. Cook and Zhang (2015a) proposed a more general definition of the envelope and adapted the envelope methods to weighted least squares, generalized linear models, and Cox regression. Cook and Zhang (2015c) introduced the envelope for simultaneously reducing the predictors and responses in multivariate linear regression. Cook *et al.*, (2015) combined the idea of envelopes and reduced-rank regression and introduced the reduced-rank envelope model. The total number of parameters for this model is no more than either of the reduced-rank regression or envelope regression. Cook and Zhang (2015b) provide a MATLAB package which implements different proposed frameworks for envelope estimators. Cook and Zhang (2016) proposed a more efficient algorithm for envelope methods.

1.4 Spatial statistics

Classical statistics always assumes observations on a phenomena are taken under identical conditions and each observation is independent. Independence is very convenient assumption but data that involves dependency are more realistic (Cressi, 2015). Sometimes the dependency in the observed value on the phenomena is related to their spatial location i.e. the dependency is such that observations that are closer to each other in space are more similar than observations far apart in space. In the literature, this type of dependency is often called

spatial autocorrelation and the set of statistical techniques that are used to analyze this type of data is called spatial statistics. Spatial data appears in a wide range of application domains including contexts that have a very large data sets, such as, computer-aided design (CAD), robotics, environmental science, and image processing (Rigaux *et al.*, 2001). This section introduces some of the basic components of spatial statistics methodology.

1.4.1 Spatial data

Spatial data are a type of data that the correlation between the data depends on their location in the study area. Mathematically this type of data is denoted by $\{Y(s); s \in \mathbb{R}^d$ where d is usually equals 2 (Banerjee *et al.*, 2014). The set of the data is called a field and each of the complete set of samples is called a realization of the field.

1.4.2 Random field

Random fields are a standard framework in which to understand spatial data. A *random field* is a set of random variables such as $Y(\cdot) = \{Y(\mathbf{s}) : \mathbf{s} \in \mathbf{D}\}$ where $\mathbf{D}$ is the domain of the random field and s denotes the location of the study area. The domain of a random field can be continuous i.e. $\mathbf{D}$ can be a subset of Euclidean space $\mathbb{R}^d;\ d \geq 1$. The domain also can be discrete which mean $\mathbf{D}$ is a subset of $\mathbb{Z}^d$ (Adler and Taylor, 2009). Mathematically the mean, variance, variogram, covariance, and correlation function of a random field $Y(\cdot)$

are defined as follows (Khoshnevisan, 2002):

Mean: $\mu(s) = E[Y(s)]$,

Variance: $Var(Y(s)) = \sigma^2(s) = E[Y(s) - E(Y(s))]^2$,

Variogram: $2\gamma(s_i, s_j) = Var(s_i - s_j)$,

Covariance function: $Cov(s_i, s_j) = C(s_i, s_j) = E\left[(Y(s_i) - E(Y(s_i)))\,(Y(s_j) - E(Y(s_j)))\right]$,

Correlation function: $Corr(s_i, s_j) = \rho(s_i, s_j) = \dfrac{C(s_i, s_j)}{\sigma(s_i)\sigma(s_j)}$.

$\mu(\cdot)$ shows large scale variation or trend. Usually researchers use a linear model like $\beta\mathbf{X}$ for $\mu(\cdot)$. If the variance of difference in responses is only a function of the difference in their locations, then this function is called variogram.

Two important characteristics of a random field are their continuity and differentiability (Stein, 2012). A random field $\{Y(s) : s \in D\}$ is *continuous in mean square* if

$$\lim_{||s_i - s_j|| \to 0} E[Y(s_i) - E(Y(s_j))]^2 = 0, \tag{1.15}$$

where $||\cdot||$ denotes Euclidean distance. Since

$$E[Y(s_i) - E(Y(s_j))]^2 = C(s_i, s_i) + C(s_j, s_j) - 2C(s_i, s_j) + (\mu(s_i) - \mu(s_j))^2,$$

therefore it can be concluded that $Y(\cdot)$ is *second order continuous in mean square* if and only if it has continuous mean and covariance function. A random field $\{Y(s) : s \in \mathbf{D}\}$ is differentiable in mean square if there exists a random field $Y'(\cdot)$ such that for every $s \in D$,

we have

$$\lim_{||h||\to 0} E\left[\frac{Y(s+h)-Y(s)}{||h||} - Y'(s)\right]^2 = 0.$$

Stein (2012) proved the following Theorem that shows a relationship between differentiability of a random field and its covariance function.

Theorem: A second order stationary random field $Y(\cdot)$ is differentiable of order q in mean square if $C(h)$ is differentiable of order $2q$ at $h = 0$.

The order of differentiability of a random field is also called its order of smoothness. Every random field can be divided to the summation of two components: large scale variation and small scale variation, i.e.

$$Y(s) = \mu(s) + \delta(s),$$

where $\mu(\cdot)$ shows large scale variation or trend and $\delta(\cdot)$ shows small scale variation or error. In spatial statistics, usually stationarity and isotropy are assumed to simplify the problem. A random field $Y(\cdot)$ is *second order stationary* if it has a constant mean and its covariance is only a function of the difference of the locations, i.e.

$$Cov(Y(s_i), Y(s_j)) = C(s_i - s_j);\ s_i, s_j \in \mathbf{D}. \tag{1.16}$$

A random field is intrinsic stationary if in addition to constant mean, variance of $Y(s_i) - Y(s_j)$ is only a function of the distance of the locations, i.e.

$$Var(Y(s_i) - Y(s_j)) = 2\gamma(s_i - s_j);\ s_i, s_j \in \mathbf{D}. \tag{1.17}$$

A second order stationary random field is intrinsic stationary but the reverse relation is not necessary true. If $C(\cdot)$ or $\gamma(\cdot)$ in (1.16) and (1.17) are only a function of the distance and they are the same in every direction, then $Y(\cdot)$ is said to be an *isotropic random field*.

Many of the phenomena in nature or a transformation of their distribution follows normal distribution. A Gaussian spatial process (random field), $\{Y(s) : s \in R^2\}$, is a stochastic process (random field) where for any collection of locations $s_1, \ldots, s_n$, the joint distribution of $S = \{Y(s_1), \ldots, Y(s_n)\}$ is multivariate Gaussian. Any process of this kind is completely specified by its mean function, $\mu(s) = E[Y(s)]$, and its covariance function, $\gamma(s_i, s_j) = Cov(Y(s_i), Y(s_j))$. The following theorem is very important for analyzing a Gaussian random field.

Theorem (Mardia *et al.*, 1979): If Y_1 and Y_2 follows a joint multivariate normal distribution then $Y_1|Y_2$ has a normal distribution with following mean and variance:

$$\begin{aligned} E(Y_1|Y_2) &= E(Y_1) + Cov(Y_1, Y_2)Var(Y_2)^{-1}(Y_2 - E(Y_2)) \\ Var(Y_1|Y_2) &= Var(Y_1) - Cov(Y_1, Y_2)Var(Y_2)^{-1}Cov(Y_2, Y_1), \end{aligned} \tag{1.18}$$

where

$$Cov(Y_1, Y_2) = E\left[(Y_1 - E(Y_1))(Y_2 - E(Y_2))^T\right].$$

One can make predictions using this theorem if the random field is Gaussian.

1.4.3 Variogram

In spatial statistics, the *variogram* is a function describing the spatial dependence of a spatial random field. The variogram is defined as the variance of the difference between field values

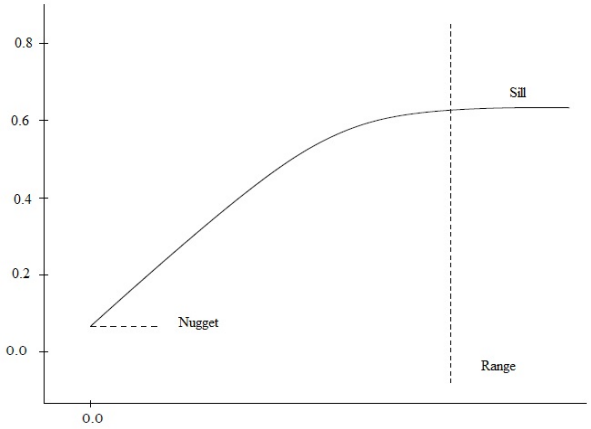

Figure 1.2: *A graphical display for the nugget, sill and range parameters.*

at two locations, s and $s+h$, across realizations of the field (Cressie, 2015):

$$2\gamma(h) = Var\left(Y(s+h) - Y(s)\right).$$

Nugget effect, sill, and range are the parameters often used to describe variograms. The Nugget effect is the jump of the variogram when h tends to zero, i.e. $h \to 0$. From a theoretical point of view, since two responses are observed at the same location, the nugget should be zero. But in application due to the sampling errors this does not happen. The sill is the limit of the variogram when h tends to infinity. The range is the distance in which the difference of the variogram from the sill becomes negligible. Figure 1.2 provides a graphical display for the nugget, sill and range parameters.

Modeling the spatial dependency structure in spatial statistics is very important and is typically done via a correlation function. The correlation function, $\rho(s_i, s_j)$, shows the similarity of variation of the obesrvations $Y(s_i)$ and $Y(s_j)$ at two sites. But estimating the correlation function solely from the data is not possible. Therefore, usually it is assumed that the form of the correlation function is a known function but it has unknown parameters

$\boldsymbol{\theta}$ where $\boldsymbol{\theta}$ control range, smoothness, and other characteristics of the correlation function. One flexible correlation function is Matern correlation function which is given by:

$$\rho(h;\boldsymbol{\theta}) = \frac{1}{2^{\theta_2-1}\Gamma(\theta_2)}\left(\frac{||h||}{\theta_1}\right)^{\theta_1}\kappa_{\theta_2}\left(\frac{||h||}{\theta_1}\right),$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)$, here $\theta_1 > 0$ is range parameter and θ_2 is smoothness parameter and κ_{θ_2} is the modified Bessel function of the second kind of order of θ_2 (Abramowitz and Stegun, 1964). Cressie (2015) provides a comprehensive review on the estimation of the $\boldsymbol{\theta}$ using Ordinary least squares (OLS), Generalized least squares (GLS), and Weighted least squares (WLS). Furthermore, Kitanidis (1983) and Mardia and Marshall (1984) proposed a likelihood-based approach to estimate the $\boldsymbol{\theta}$.

1.4.4 Multivariate spatial statistics

In spatial statistics, there exist a large number of situations that have a multivariate response. For example, if we want to study the air pollution of a city, pollutants such as ground level ozone, carbon monoxide, sulfur, etc. are measured simultaneously at each location. Observed values from these experiments produce multivariate spatial data which usually has two spatial correlations: A spatial autocorrelation which is within the observation of each variable in different locations and a cross spatial correlation which shows the relationship between two different variables in same and/or different locations. Joint modeling of both of these correlations is very important and adds significant difficulty to the data analysis.

A multivariate random field such as $\mathbf{Y}(\cdot) = \{\mathbf{Y}(s),\ s \in D\}$ is called second order

stationary if for all $s,\ h \in \mathbb{R}^d$ and for $i, j = 1, 2, \ldots, p$

$$E(\mathbf{Y}_i(s)) = m_i, \quad Cov(\mathbf{Y}_i(s), \mathbf{Y}_j(s+h)) = C_{ij}(h) \tag{1.19}$$

where $C_{ij}(h)$ is called direct variogram where $i = j$ and indirect cross variogram for $i \neq j$. Furthermore, $C(h) = [C_{ij}(h)]$ is called multivariate variogram and it should be positive definite. In other words, for every vector of locations such as $s = (s_1, \ldots, s_n)$ and every vector $a_i \in \mathbb{R}^p;\ i = 1, \ldots, n$, we have

$$Var\left(\sum_i a_i^T \mathbf{Y}(s_i)\right) = \sum_{i,j=1}^{n} a_i^T C(s_i - s_j) a_j > 0.$$

With this condition, determination of a multivariate variogram that can capture the spatial structure while not being so complicated to estimate is a difficult task. However, there exist a large number of different methods to create a joint variogram. Wackernagel (2013) provide a comprehensive discussion on creating a multivariate spatial covariance function and introduced several different methods to analyze these types of data.

The Linear Coregionalization Model (LCM; Zahng, 2007, Banerjee *et al.*, 2014) is one method that has received a lot of attention in the recent literature. The most basic LCM uses the *intrinsic specification* which assumes $\mathbf{Y}(\cdot)$ can be represented as a linear combination of independent and identical random fields $\omega(s)$ i.e. $\mathbf{Y}(s) = \mathbf{A}\omega(s)$. If $\omega_k(s),\ k = 1, 2, \ldots, p$ is a stationary random field with mean 0 and variance of 1 and correlation function of $\rho(h)$ then $E(\mathbf{Y}(s)) = 0$ and its covariance matrix is

$$\mathbf{\Sigma}_{\mathbf{Y}(s_i),\mathbf{Y}(s_j)} = C(s_i - s_j) = \rho(s_i - s_j)\mathbf{A}\mathbf{A}^T,\ i, j = 1, 2, \ldots, n.$$

Suppose that $\mathbf{V} = \mathbf{A}\mathbf{A}^T$, then we have a separable covariance matrix. Intrinsic means for model specification only the first and second moment of subtracted value vector are needed. Furthermore, the first moment is zero and second moment is related to the location only through the distance vector i.e. $s_i - s_j$. In fact

$$E(\mathbf{Y}(s_i) - \mathbf{Y}(s_j)) = 0$$
$$\frac{1}{2}\boldsymbol{\Sigma}_{\mathbf{Y}(s_i)-\mathbf{Y}(s_j)} = G(s_i - s_j),$$

where $G(s_i - s_j) = C(0) - C(s_i - s_j) = \mathbf{V} - \rho(s_i - s_j)\mathbf{V} = \gamma(s_i - s_j)\mathbf{V}$ where $\gamma(\cdot)$ is a valid variogram.

The covariance matrix for observation vector using (1.16) is:

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = \mathbf{R} \otimes \mathbf{V}$$

where $\mathbf{A} = (\rho(s_i - s_j))$ and $\otimes$ shows Kronecker products. If $\mathbf{V}$ and $\mathbf{R}$ are positive definite then $\boldsymbol{\Sigma}_{\mathbf{Y}}$ is positive definite as well. Working with $\boldsymbol{\Sigma}_{\mathbf{Y}}$ is very easy in separable model because $|\boldsymbol{\Sigma}_{\mathbf{Y}}| = |\mathbf{R}|^p|\mathbf{V}|^n$ where $|\cdot|$ denotes the determinant of a matrix and $\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1} = \mathbf{R}^{-1} \otimes \mathbf{V}^{-1}$. This means for updating $\boldsymbol{\Sigma}_{\mathbf{Y}}$, instead of working with a $pn \times pn$ matrix one only needs to work with two $p \times p$ and $n \times n$ matrices. In addition, by relocating the rows of $\mathbf{Y}$ such that we have $\tilde{\mathbf{Y}} = (Y_1(s_1), \ldots, Y_1(s_n), \ldots, Y_p(s_1), \ldots, Y_p(s_n))$, we have $\boldsymbol{\Sigma}_{\mathbf{Y}} = \mathbf{T} \otimes \mathbf{R}$. A more general extension of LCM can also be defined where the covariance matrix is *indivisible*. For instance, assume $\mathbf{Y}(s) = \mathbf{A}\omega(s)$ where $\omega_j(s)$ are independent but not identically distributed. In fact, $\omega_j(s)$ are processes with mean of μ_j, variance of 1 and correlation function $\rho_j(h)$. In this case, $E(\mathbf{Y}(s)) = \mathbf{A}\boldsymbol{\mu}$ where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ and covariance matrix of $\mathbf{Y}$(s) is $\boldsymbol{\Sigma}_{\mathbf{Y}(s_i),\mathbf{Y}(s_j)} = C(s_i - s_j) = \sum_{k=1}^{p} \rho_k(s_i - s_j)\mathbf{V}_k$ where $\mathbf{V}_k = a_k a_k^T$ where a_k is k-th row of

$\mathbf{A}$ and $\sum_k \mathbf{V}_k = \mathbf{V}$. As another extension, the model can be define such that the process has a nested covariance model. In other words, $\mathbf{Y}(s) = \sum_{u=1}^{r} \mathbf{Y}^{(u)}(s) = \sum_{u=1}^{r} \mathbf{A}^{(u)}\omega^{(u)}(s)$ where $\mathbf{Y}^{(u)}(s)$ are independent intrinsic LCM with $\omega^{(u)}$ as their components with correlation function ρ_u. Covariance matrix of this model is $C(s_i - s_j) = \sum_{u=1}^{r} \rho_u(s_i - s_j)\mathbf{V}^{(u)}$ where $\mathbf{V}^{(u)} = \mathbf{A}^{(u)}(\mathbf{A}^{(u)})^T$.

Goulard and Voltz (1992) studied least squares estimators for some of the LCM parameters using the empirical multivariate variogram, under the assumption that some other parameters in the model are known. Zhang (2007) developed an EM algorithm for the maximum-likelihood estimation for the parameters in the LCM. Fasso and Finazzi (2012) extend this model to heterotopic data. Genton and Kleiber (2015) provide a comprehensive review on the main approaches to building cross-covariance models for multivariate spatial and spatiotemporal, including the linear model of coregionalization, convolution methods, the multivariate Matern and nonstationary and space-time extensions of these among others.

1.5 Summary

As discussed above, the development of sufficient dimension reduction in theory and methodology has provided a powerful tool to tackle the high dimensional data analysis. It has been widely applied into many scientific fields in recent years, such as in micro array data analysis (Bura and Pfeiffer, 2003) and gene expression data analysis (Antoniadis *et al.*, 2003). All the methods discussed above have their own advantages, and disadvantages. For instance, the inverse methods are easy to implement and have good asymptotic properties, but they requires some probabilistic assumptions. On the other hand, the forward methods do not require strong probabilistic assumptions, but are computationally more expensive than the

inverse regression approach.

In this dissertation, we study three projects. The first one, as shown in Chapter 2, we combine local modal regression and MAVE to introduce a robust dimension reduction approach. In addition to being robust to outliers or heavy tailed distributions, our proposed method has the same convergence rate as original MAVE. In addition, we combine local modal base MAVE with an L_1 penalty to select informative covariates. This new approach can exhaustively estimate directions in the regression mean function and select informative covariates simultaneously, while being robust to the existence of possible outliers in the dependent variable. The second project, which is detailed in Chapter 3, is to develop sparse adaptive MAVE (aMAVE). For this project, we combined aMAVE with adaptive LASSO (Zou, 2006). Sparse adaptive MAVE (saMAVE) has advantages over adaptive LASSO because it extends adaptive LASSO to multi-dimensional and nonlinear settings, without any model assumption, and has advantages over sparse inverse dimension reduction methods (Li, 2007) in that it does not require any particular distribution on $\mathbf{X}$ and it can exhaustively estimate the dimensions in the conditional mean function. The third project, as discussed in Chapter 4, is to extend the envelope idea to multivariate response problems with spatial correlation.

Chapter 2

Robust Estimation and Variable Selection in Sufficient Dimension Reduction

Dimension reduction and variable selection play important roles in high dimensional data analysis. Minimum Average Variance Estimation (MAVE) is an efficient approach among many others. However, because of using the least squares criterion, MAVE is not robust to outliers or errors with heavy tailed distributions. In this paper, we propose a robust extension of MAVE which can adapt to different error distributions. Our proposed estimate is shown to have the same convergence rate as the original MAVE. Furthermore, we combine our proposed method with adaptive LASSO to select the informative variables. This new approach is illustrated through simulation studies and a data analysis on air quality of Hong

Kong[1].

2.0.1 Local modal regression

Modal regression (Yao *et al.*, 2012) is an alternative approach for usual regression where instead of modeling conditional mean, it models the conditional mode of y give $\mathbf{X}$, i.e. $\text{Mode}(y|\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}$. The logic behind using local modal regression is that the conditional mode can reveal structure possibly missed by the conditional mean. For a univariate non-parametric regression model

$$y = m(x) + \epsilon, \tag{2.1}$$

the local modal regression estimates $m(x)$ and its derivative by maximizing

$$Q(\beta_0, \beta_1) \equiv \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x_i - x)\phi_{h_2}\{y_i - \beta_0 - \beta_1(x_i - x)\}, \tag{2.2}$$

where β_0 is intercept, β_1 is slope, $K_{h_1}(\text{t})$ is a weight function and $\phi_{h_2}(t) = h_2^{-1}\phi(t/h_2)$ is a kernel density function. If we treat $-\phi_{h_2}(\cdot)$ as a loss function, the resulting M-estimator is a local modal regression estimator. As mentioned in Yao et al. (2012), the bandwidth h_2 corresponds to the standard deviation in the normal density. A small h_2 results in an outlier-resistant loss function, while a large h_2 produces a loss function similar to L_2 loss. The estimator is asymptotically as efficient as the least squares estimators for normally distributed error. Since modal regression focuses on modeling the mode of $Y|\mathbf{X}$, it is robust to outliers and heavy tailed errors. When the conditional distribution of $\epsilon|X$ is symmetric about the origin, the estimates from modal regression and mean regression coincide. For

[1]Moradi Rekabdarkolaee H., Boone E. L., and Wang Q., 2016, Computational Statistics and Data Analysis.

more detail see Yao *et al.* (2012).

2.1 Local modal MAVE (lmMAVE)

Note that in (1.7) the least squares criterion is used. So, if the regression error has a heavy tail distribution or suffers from severe outliers, the finite-sample performance of the MAVE can be poor. Our suggestion is to replace the local least squares with local modal estimation. Finding $\mathbf{B}$ is equivalent to maximization of the following problem

$$Q(\boldsymbol{\theta}) = \sum_{j=1}^{n} \sum_{i=1}^{n} w_{ij} \log \phi_{h_2} \left[y_i - \left\{ a_j + \mathbf{b}_j^T \mathbf{B}^T \left(\mathbf{X}_i - \mathbf{X}_j \right) \right\} \right], \tag{2.3}$$

where $\boldsymbol{\theta} = \{\mathbf{B},\, (a_j, \mathbf{b}_j), j = 1, 2, \ldots, n\}$ and $\phi_{h_2} = h_2^{-1}\phi(t/h_2)$ is a kernel density function. For ease of computation, we use the standard normal density for $\phi(\cdot)$ throughout this paper. Note that, similar to most nonparametric regression, the choice of the kernel function is not very crucial in terms of estimation efficiency.

Given an initial value for $\boldsymbol{\theta}$ denoted by $\boldsymbol{\theta}^{(0)}$, we adopt the modal expectation-maximization (EM) algorithm proposed by Yao *et al.* (2012) to maximize (2.3). The EM algorithm is designed to find (locally) maximum likelihood parameters of a model in cases where the equations cannot be solved directly (Dempster *et al.*, 1977). The EM iteration alternates between an expectation (E) step and a maximization (M) step. At the E step, the algorithm creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters. At the M step, the algorithm computes parameters maximizing the expected log-likelihood found at the E step. The parameters that are estimated in the M step will be used in the following the E step. In our problem, the $(k+1)^{st}$ step of the EM

algorithm is as follows:

E-step: find the classification probabilities

$$p_{ij}^{(k+1)} = \frac{w_{ij}\phi_{h_2}\left[y_i - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}\left(\mathbf{X}_i - \mathbf{X}_j\right)\right\}\right]}{\sum_{l=1}^{n}\phi_{h_2}\left[y_l - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}\left(\mathbf{X}_l - \mathbf{X}_j\right)\right\}\right]}. \tag{2.4}$$

M-step: update parameter estimates of $\boldsymbol{\theta}$ by maximizing

$$\sum_{j=1}^{n}\sum_{i=1}^{n} p_{ij}^{(k+1)} \log\phi_{h_2}\left[y_i - \left\{a_j + \mathbf{b}_j^T\mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_j\right)\right\}\right]. \tag{2.5}$$

The above EM algorithm monotonically increases the local log-likelihood (2.3) after each iteration, as shown in the following theorem.

Theorem 1: Each iteration of the above E and M steps will monotonically increase the local log-likelihood (2.3), i.e. $Q(\boldsymbol{\theta}^{(k+1)}) \geq Q(\boldsymbol{\theta}^{(k)})$, for all k, where $Q(\cdot)$ is defined as in (2.3).

Theorem 2: Suppose the conditions (C1-C9) in the appendix hold. Let $\mathbf{B}$ be the direction estimated from the local modal MAVE. If $nh^p/log(n) \to \infty,\ h \to 0,\ d \geq D$ and $h_2 = h/\log(n)$ then

$$||(I - \mathbf{B}\mathbf{B}^T)\mathbf{B}_0|| = O_p(h^3 + h\delta_n + h^{-1}\delta_n^2), \tag{2.6}$$

where $\delta_n = \{log(n)/(nh^p)\}^{1/2}$. It can be seen that the local modal regression based MAVE achieves the same convergence rate as the traditional MAVE. The bandwidth condition $h_2 = h/\log(n)$ is used in Wang and Yao (2012) for the simplicity of the proof. As they also indicated, a wider range of bandwidth for h_2 can be used without changing the convergence rate but with a more complicated proof. A sketch of the proofs for the above two results is provided in the Appendix.

Remark: We focus on the robustness against the outliers in the dependent variable, but not on the leverage points. As discussed in Čížek and Härdle (2006) and Yao and Wang (2013), the leverage points have limited influence on MAVE since the estimation is based on local linear regression and high leverage points are less likely to appear in a local window determined by the bandwidth and the kernel function. Our numerical studies confirmed this conclusion (not included in the paper due to the space limitation).

To select the informative covariates robustly, an L_1 penalty can be introduced into the expression (2.3),

$$\max_{\mathbf{B},a_j,b_j,j=1,\ldots,n} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} p_{ij} \log \phi_{h_2} \left[y_i - \left\{ a_j + \mathbf{b}_j^T \mathbf{B}^T \left(\mathbf{X}_i - \mathbf{X}_j \right) \right\} \right] - \sum_{k=1}^{d} \lambda_k ||\beta_k||_1 \right). \quad (2.7)$$

where $||\cdot||_1$ represents the L_1 norm and $\{\lambda_k, k = 1, 2, \ldots, d\}$ are the non-negative regularization parameters. Adding L_1 penalty on $\mathbf{B}$ to the formula (3.5) may shrink some elements of $\mathbf{B}$ to exact zeros.

2.2 Implementation

In this section, we introduce the computation algorithm for our method and related tuning parameter selection.

2.2.1 Computation Algorithm

For a given sample $\{(y_i, \mathbf{X}_i), i = 1, 2, \ldots, n\}$,

1. Obtain an initial estimator $\{\hat{\mathbf{B}}, (\hat{a}_j, \hat{\mathbf{b}}_j), j = 1, 2, \ldots, n\}$ from the original MAVE.

2. For a given $\hat{\mathbf{B}}$, update $(a_j, \mathbf{b}_j)$ where $j = 1, 2, \ldots, n$, from the following maximization problem

$$\max_{a_j, \mathbf{b}_j, j=1,\ldots,n} \sum_{j=1}^{n} \sum_{i=1}^{n} p_{ij} \log \phi_{h_2} \left[y_i - \left\{ a_j + \mathbf{b}_j^T \mathbf{B}^T (\mathbf{X}_i - \mathbf{X}_j) \right\} \right]. \tag{2.8}$$

3. For given $(\hat{a}_j, \hat{\mathbf{b}}_j)$, $j = 1, 2, \ldots, n$, solve $\mathbf{B}$ from the following maximization problem

$$\max_{a_j, \mathbf{b}_j, j=1,\ldots,n} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} p_{ij} \log \phi_{h_2} \left[y_i - \left\{ a_j + \mathbf{b}_j^T \mathbf{B}^T (\mathbf{X}_i - \mathbf{X}_j) \right\} \right] - \sum_{k=1}^{d} \lambda_k ||\beta_k||_1 \right). \tag{2.9}$$

4. Iterate between step 2 and 3 until convergence in the estimation of $\mathbf{B}$.

The closed form solution for $\boldsymbol{\theta}^{(k+1)}$ is one of the benefits of using the Gaussian kernel function $\phi_{h_2}(\cdot)$ in Equation (2.3). More details can be found in Yao *et al.* (2012) and Wang and Yao (2012). Based on our empirical experience, the proposed algorithm is not sensitive to the initial estimator. Furthermore, the algorithm usually converges within $10 - 20$ iterations. However, one might further speed up the computation based on the one-step M-estimation as discussed in Fan and Jiang (2000), Welsh and Ronchetti (2002), and Čížek and Härdle (2006).

2.2.2 Tuning Parameter Selection

We employed a very efficient Lasso algorithm recently proposed by Friedman *et al.* (2010) to solve the L_1 regularized maximization (2.7). Cyclical coordinate descent methods were used to calculate the solution path for a large number of λ at once. We used the Matlab package glmnet in all the simulation studies. More details can be found at http://www-stat.stanford.edu/∼tibs/glmnet-matlab/. A BIC criterion was used to select the optimal λ's

in the Lasso estimation,

$$BIC_\lambda = n\ log\left(\frac{Q_\lambda(\hat{\mathbf{B}})}{n}\right) + log(n)p_\lambda,$$

where $Q_\lambda(\hat{\mathbf{B}}) = \sum_{j=1}^{n}\sum_{i=1}^{n} p_{ij} \log \phi_{h_2}\left[y_i - \left\{\hat{a}_j + \hat{\mathbf{b}}_j^T \hat{\mathbf{B}}^T(\mathbf{X}_i - \mathbf{X}_j)\right\}\right]$ is the residual sum of squares from the LASSO fit, and p_λ denotes the number of non-zero coefficients.

The estimation of the structural dimension d is another important task in sufficient dimension reduction. In this section, we adopt a robust cross-validation (CV) procedure proposed in Yao and Wang (2013) to determine the optimal dimension d. Once we have an estimated $\hat{\mathbf{B}}$ for a given dimension k, we can calculate the corresponding CV value as

$$CV_k = n^{-1}\sum_{i=1}^{n} \rho\left(y_i - \frac{\sum_{j\neq i} y_j K_h\{\hat{\mathbf{B}}^T(\mathbf{X}_j - \mathbf{X}_i)\}}{\sum_{l\neq i} K_h\{\hat{\mathbf{B}}^T(\mathbf{X}_l - \mathbf{X}_i)\}}\right), \tag{2.10}$$

where $\rho(\cdot)$ is the Tukey's bisquare loss function defined as

$$\rho(t) = \begin{cases} 1 - [1 - (t/c)^2]^3 & \text{if } |t| \leq c; \\ 1 & \text{if } |t| > c. \end{cases}$$

Then the structural dimension d can be estimated by

$$\hat{d} = \underset{0\leq k\leq p}{\operatorname{argmin}}\ CV_k.$$

The tuning constant $c = 4.685\hat{\sigma}$ is proportional to the scaled estimate of $\hat{\sigma}$, and controls the balance between robustness and the estimation efficiency. More details can be found in Yao and Wang (2013) and the references therein.

2.3 Simulation study

In this section, we carried out the simulation study to evaluate the finite sample performance of the proposed local modal MAVE (lmMAVE) and its sparse version (slmMAVE). We compare them with sliced inverse regression (SIR; Li, 1991), the traditional refined MAVE (rMAVE; Xia *et al.*, 2002), sparse MAVE (sMAVE, Wang and Yin 2008), robust MAVE (rtMAVE; Čížek and Härdle, 2006), and robust sparse MAVE (rsMAVE; Yao and Wang, 2013).

The data $\{(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_n, y_n)\}$ were generated from the model

$$Y = \frac{\beta_1^T \mathbf{X}}{0.5 + (\beta_2^T \mathbf{X} + 1.5)^2} + 0.5\epsilon, \tag{2.11}$$

where $\mathbf{X}_i \in R^{10}$ and $\beta_1 = (1, 0, \ldots, 0)^T$, $\beta_2 = (0, 1, \ldots, 0)^T$. Four error distributions of ϵ were investigated:

1. N(0,1), the standard normal errors. This density serves as a benchmark with no outliers;
2. $t_3/\sqrt{3}$, the scaled t-distribution with 3 degree of freedom;
3. $0.95N(0, 1) + 0.05N(0, 10^2)$, the standard normal errors contaminated by 5% normal errors with mean 0 and standard deviation 10;
4. 0.95N(0,1)+0.05U($-$50,50), the standard normal errors contaminated by 5% errors from a uniform distribution in between -50 and 50.

Sample size was chosen as 100, 200 and 400, and 200 data replicates were drawn in each

setup. We considered both independent and correlated cases for $\mathbf{X}$: (a) $\mathbf{X} \sim N_{10}(\mathbf{0}_{10}, \mathbf{I}_{10})$, and (b) $\mathbf{X} \sim N_{10}(\mathbf{0}_{10}, \Sigma)$, where $\Sigma = [\sigma_{ij}]$ with $\sigma_{ij} = 0.5^{|i-j|}$.To compare different estimators, we used vector correlation coefficient r^2 as in Ye and Weiss (2003). Let S(A) and S(B) denote two d-dimensional spaces where A and B are orthonormal bases, respectively. The vector correlation coefficient is defined as $r^2 = \frac{1}{d}\text{trace}(B^T A A^T B)$. To measure the effectiveness of variable selection, we employ the true positive rate (TPR): the ratio of the number of correctly identified active predictors to the number of truly active predictors, and the false positive rate (FPR): the ratio of the number of falsely identified active predictors to the number of inactive predictors. Ideally, the TPR should be close to 1 and the FPR should be close to 0 simultaneously.

Tables 2.1 and 2.2 summarize the vector correlation coefficients from independent and correlated data, respectively. Tables 2.3 and 2.4 show the results from variable selection for sMAVE, rsMAVE, and slmMAVE. We chose Gaussian kernel in the weights w_{ij} and used the so-called normal reference bandwidth as $h = 1.06n^{-0.2}\hat{\sigma}$ where $\hat{\sigma} = \min\{(q_{0.75} - q_{0.25})/1.34, \hat{\sigma}_\epsilon\}$ with $\hat{\sigma}_\epsilon$ and q_i being the standard deviation and the i-th quantile of the error ϵ, respectively. For more details, we refer readers to Silverman (1986).

From the summary of all four different error distributions, we have the following observations.

1. For the standard normal errors, local modal MAVE gave broadly comparable results as the least squares based methods. The performance of SIR was not affected by the outliers in the y-space since only the rank of the response values was used, but SIR does require some strong probabilistic assumptions on the predictor space.

2. rMAVE did show some robustness when the errors were from the scaled t-distribution,

Table 2.1: *Mean (standard deviation) of the vector correlation coefficient r^2 for independent predictors from 200 data replications*

ϵ	n	SIR	rMAVE	sMAVE	rtMAVE	rsMAVE	lmMAVE	slmMAVE
1	100	0.8566	0.9474	0.9842	0.9317	0.9683	0.9180	0.9452
		(0.079)	(0.101)	(0.152)	(0.114)	(0.148)	(0.132)	(0.124)
	200	0.9382	0.9819	0.9933	0.9760	0.9896	0.9746	0.9922
		(0.026)	(0.076)	(0.084)	(0.095)	(0.083)	(0.057)	(0.037)
	400	0.9725	0.9922	0.9973	0.9900	0.9956	0.9906	0.9992
		(0.011)	(0.019)	(0.009)	(0.030)	(0.008)	(0.028)	(0.008)
2	100	0.8644	0.9474	0.9830	0.9366	0.9645	0.8910	0.9300
		(0.072)	(0.115)	(0.154)	(0.121)	(0.142)	(0.148)	(0.142)
	200	0.9347	0.9819	0.9937	0.9788	0.9895	0.9649	0.9751
		(0.028)	(0.087)	(0.097)	(0.089)	(0.067)	(0.074)	(0.073)
	400	0.9689	0.9921	0.9976	0.9912	0.9956	0.9884	0.9943
		(0.012)	(0.038)	(0.023)	(0.042)	(0.012)	(0.027)	(0.017)
3	100	0.8197	0.7224	0.8461	0.9038	0.9437	0.9148	0.9455
		(0.088)	(0.144)	(0.251)	(0.113)	(0.163)	(0.184)	(0.182)
	200	0.9175	0.8171	0.8865	0.9332	0.9698	0.9545	0.9831
		(0.036)	(0.113)	(0.217)	(0.106)	(0.128)	(0.098)	(0.082)
	400	0.9613	0.8762	0.8974	0.9494	0.9759	0.9707	0.9984
		(0.016)	(0.100)	(0.204)	(0.059)	(0.041)	(0.039)	(0.024)
4	100	0.8026	0.5190	0.6601	0.9128	0.9393	0.9020	0.9244
		(0.090)	(0.169)	(0.276)	(0.117)	(0.184)	(0.134)	(0.167)
	200	0.9121	0.5444	0.7934	0.9353	0.9596	0.9416	0.9731
		(0.036)	(0.154)	(0.257)	(0.104)	(0.102)	(0.087)	(0.096)
	400	0.9591	0.6343	0.9075	0.9497	0.9854	0.9680	0.9940
		(0.020)	(0.128)	(0.243)	(0.068)	(0.038)	(0.041)	(0.039)

Table 2.2: *Mean (standard deviation) of the vector correlation coefficient r^2 for correlated predictors from 200 data replications*

ϵ	n	SIR	rMAVE	sMAVE	rtMAVE	rsMAVE	lmMAVE	slmMAVE
1	100	0.7398	0.6880	0.8658	0.6627	0.7731	0.7011	0.8197
		(0.118)	(0.112)	(0.158)	(0.111)	(0.148)	(0.123)	(0.146)
	200	0.9132	0.8510	0.9652	0.8347	0.9019	0.7708	0.8652
		(0.045)	(0.098)	(0.118)	(0.092)	(0.083)	(0.073)	(0.067)
	400	0.9621	0.9331	0.9814	0.9293	0.9748	0.9015	0.9726
		(0.018)	(0.073)	(0.081)	(0.083)	(0.024)	(0.071)	(0.033)
2	100	0.8045	0.6981	0.8390	0.7134	0.7969	0.6604	0.7736
		(0.098)	(0.126)	(0.184)	(0.135)	(0.128)	(0.157)	(0.137)
	200	0.8925	0.8385	0.9377	0.8794	0.9423	0.7953	0.8564
		(0.054)	(0.086)	(0.104)	(0.096)	(0.076)	(0.107)	(0.097)
	400	0.9549	0.9045	0.9699	0.9384	0.9707	0.8559	0.9684
		(0.021)	(0.056)	(0.094)	(0.061)	(0.039)	(0.046)	(0.041)
3	100	0.7385	0.5273	0.6579	0.6530	0.7648	0.6918	0.7941
		(0.113)	(0.176)	(0.196)	(0.159)	(0.134)	(0.125)	(0.127)
	200	0.8716	0.5315	0.6783	0.8091	0.8929	0.7603	0.8384
		(0.074)	(0.126)	(0.134)	(0.125)	(0.093)	(0.098)	(0.087)
	400	0.9463	0.5946	0.7790	0.9233	0.9722	0.9021	0.9644
		(0.024)	(0.106)	(0.092)	(0.094)	(0.068)	(0.053)	(0.043)
4	100	0.7169	0.3193	0.3835	0.6543	0.7471	0.6589	0.7602
		(0.123)	(0.196)	(0.184)	(0.145)	(0.123)	(0.135)	(0.137)
	200	0.8710	0.3301	0.4226	0.8165	0.8845	0.7814	0.8730
		(0.069)	(0.146)	(0.124)	(0.129)	(0.098)	(0.087)	(0.064)
	400	0.9409	0.3443	0.4528	0.9219	0.9646	0.8991	0.9552
		(0.026)	(0.136)	(0.094)	(0.085)	(0.071)	(0.072)	(0.040)

Table 2.3: *True positive rate and false positive rate for independent predictors.*

ϵ	n	sMAVE	rsMAVE	slmMAVE
1	100	(0.875, 0.175)	(0.821, 0.179)	(0.815, 0.165)
	200	(0.954, 0.074)	(0.948, 0.083)	(0.947, 0.083)
	400	(1.000, 0.038)	(1.000, 0.042)	(1.000, 0.043)
2	100	(0.883, 0.204)	(0.861, 0.187)	(0.817, 0.155)
	200	(0.956, 0.127)	(0.954, 0.092)	(0.937, 0.094)
	400	(1.000, 0.079)	(1.000, 0.089)	(0.995, 0.086)
3	100	(0.790, 0.346)	(0.842, 0.196)	(0.875, 0.135)
	200	(0.847, 0.197)	(0.946, 0.105)	(0.998, 0.101)
	400	(0.893, 0.163)	(1.000, 0.063)	(1.000, 0.065)
4	100	(0.615, 0.543)	(0.857, 0.189)	(0.878, 0.124)
	200	(0.649, 0.476)	(0.963, 0.088)	(0.925, 0.076)
	400	(0.642, 0.424)	(1.000, 0.056)	(0.995, 0.048)

Table 2.4: *True positive rate and false positive rate for correlated predictors.*

ϵ	n	sMAVE	rsMAVE	slmMAVE
1	100	(0.795, 0.143)	(0.767, 0.151)	(0.750, 0.180)
	200	(0.965, 0.114)	(0.949, 0.139)	(0.817, 0.103)
	400	(1.000, 0.050)	(1.000, 0.099)	(0.945, 0.061)
2	100	(0.899, 0.196)	(0.908, 0.216)	(0.830, 0.204)
	200	(0.943, 0.142)	(1.000, 0.125)	(0.905, 0.164)
	400	(1.000, 0.113)	(1.000, 0.122)	(0.975, 0.114)
3	100	(0.885, 0.437)	(0.952, 0.210)	(0.742, 0.234)
	200	(0.917, 0.383)	(0.985, 0.156)	(0.802, 0.147)
	400	(0.985, 0.345)	(1.000, 0.131)	(0.945, 0.098)
4	100	(0.725, 0.557)	(0.872, 0.228)	(0.802, 0.216)
	200	(0.662, 0.469)	(0.947, 0.165)	(0.875, 0.155)
	400	(0.692, 0.408)	(1.000, 0.139)	(0.967, 0.115)

as mentioned in the original MAVE paper. However, with the inclusion of larger outliers in the response as in the error distributions 3 and 4, the least squares based methods, rMAVE and sMAVE, failed to estimate the true directions and to select the informative covariates.

3. In the error distributions $2-4$, the robust estimation procedures performed almost equally as well as they did in the cases without outliers. By selecting the informative covariates, the slmMAVE outperformed the lmMAVE in terms of estimation accuracy and also eased the subsequent model building. In addition, slmMAVE provide comparable results as rsMAVE and both of these methods outperformed sMAVE, especially in the error distributions 3 and 4 where relatively large outliers appear.

4. As one reviewer pointed out, the proposed slmMAVE had lower TPR compared to rsMAVE for the correlated predictors (Table 2.4). This might be due to the use of numerical estimation (EM algorithm) in slmMAVE instead of the trimmed least squares used in rsMAVE, and also the selection of h_2.

Based on the above findings, we can conclude that the proposed lmMAVE and slmMAVE

procedures provided very consistent estimates with good direction estimation and variable selection accuracy in all error distributions considered.

2.4 Real data analysis

Air pollution is the existence of one or several pollutant elements such as dust, gases, or smoke in air that has a serious impact on the health of plants and animals (including humans). Substances that are not naturally found in the air or at greater concentrations than usual are referred to as pollutants.

The pollutant and weather data that we used in this study are the daily average levels of nitrogen dioxide (NO_2), sulphur dioxide (SO_2), respirable suspended particulates (rsp), temperature (temp) and relative humidity (hum). The data were collected daily in Hong Kong from January 1, 1994, to December 31, 1997. This data set can be found at http://www.stat.nus.edu.sg/∼staxyc/.

We are interested in studying the statistical relation between ground level Ozone (y) and the levels of other pollutants and weather conditions. In addition to the main effects, all two way interactions were also included in the model. Each variable was standardized individually for the ease of interpretation. We initially used LASSO to analyze the data and we find out that LASSO fails to capture the nonlinear behavior of the ozone. The corresponding direction estimates from both lmMAVE and slmMAVE are listed in Table 2.5.

By checking the estimated coefficients (directions), we can see all the main effects are important in first direction. Except the interaction between NO_2 and SO_2, all other interac-

Table 2.5: *The estimated directions for Hong Kong air pollution data.*

	slmMAVE		lmMAVE	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
NO_2	0.02	0	0.01	-0.04
SO_2	-0.04	0	-0.05	-0.07
rsp	0.02	0	0.01	-0.14
temp	0.11	0	0.11	-0.01
hum	-0.12	0	-0.12	0.10
$NO_2 \times SO_2$	0	0	0.02	0.08
NO_2×rsp	0.45	0	0.44	0.06
NO_2×temp	0.41	0	0.42	0.23
NO_2×hum	-0.07	0	-0.09	-0.22
SO_2×rsp	0.05	0	0.06	0.02
SO_2×temp	0.16	0	0.20	0.05
SO_2×hum	-0.63	0.77	-0.62	0.56
rsp×temp	0.36	0	0.36	0.34
rsp×hum	0.12	0	0.10	-0.07
temp×hum	-0.15	-0.64	-0.15	-0.64

tions are important, especially the interaction between those primary pollutants and weather conditions. The second direction includes the interactions between two weather conditions and the interaction between SO_2 and humidity. These findings support the chemical claim that ozone level is affected by chemical reactions between weather conditions and pollutant such as oxides of nitrogen or sulfur in the presence of sunlight. Fig 2.1 shows the scatter plots between the ground level ozone vs. the two estimated directions.

2.5 Theoretical results

2.5.1 Regularity conditions

The following technical conditions are imposed in this section:

(C1) $\{(\mathbf{X}_i, y_i), i = 1, \ldots, n\}$ are i.i.d. samples from the joint density $f_{\mathbf{X},y}(\mathbf{x}, y)$.

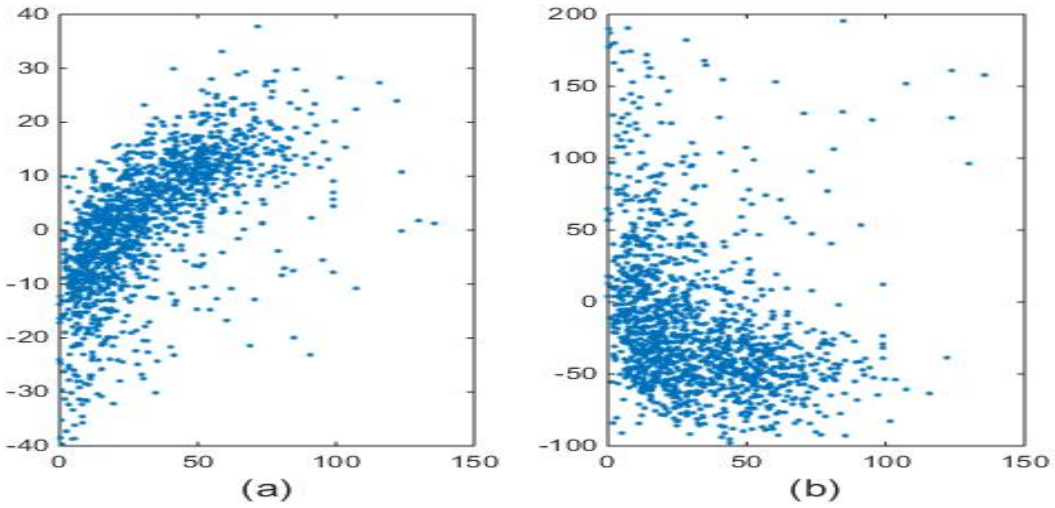

Figure 2.1: *Ground level ozone plotted against (a) the first direction (b) the second direction from slmMAVE.*

(C2) $\{\epsilon_i\}$ are i.i.d. with $E(\epsilon_i) = 0$, $E(|\epsilon_i|^3) < \infty$. The probability density function of ϵ, $f_\epsilon(\cdot)$, is symmetric about the origin.

(C3) $\{\mathbf{X}_i\}$ and $\{\epsilon_i\}$ are mutually independent. Additionally, the predictor $\mathbf{X}$ has a bounded support.

(C4) The kernel density function $\phi_{h_2(\cdot)}$ has bounded continuous derivatives up to order 4. Let $\ell(\cdot) = \log \phi_{h_2}(\cdot)$. Assume $\ell'''(\cdot)$ is bounded and $E\{\ell'(\epsilon)^2 + |\ell''(\epsilon)| + |\ell'''(\epsilon)|\} < \infty$.

(C5) $E|y|^k < \infty$ and $E||\mathbf{X}||^k < \infty$ for all $k > 0$.

(C6) The density function $f_y(\cdot)$ of y has bounded derivative and is bounded away from 0 on a compact support.

(C7) $g(\cdot)$ has bounded, continuous 3rd derivatives.

(C8) $E(\mathbf{X}|y)$ and $E(\mathbf{X}\mathbf{X}^T|y)$ have bounded, continuous 3rd derivatives.

(C9) $K(\cdot)$ is a spherical symmetric density function with a bounded derivative and support. All the moments of $K(\cdot)$ exist and $\int UU^T K(U)dU = I$.

The above conditions are imposed to facilitate the proof and most of them are similar to Xia *et al.* (2002) and Wang and Yao (2012). They are not the weakest possible conditions. We require the error density function $f_\epsilon(\cdot)$ being symmetric so that the proposed method targets on the central mean subspace. which makes the comparison with original MAVE meaningful.

2.5.2 Proof of Theorem 1

Note that

$$
\begin{aligned}
Q(\boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k)}) &= \sum_{j=1}^{n} \log \left\{ \frac{\sum_{i=1}^{n} \phi_{h_2}\left[y_i - \left\{a_j^{(k+1)} + \mathbf{b}_j^{(k+1)^T}\mathbf{B}^{(k+1)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\}\right]}{\sum_{l=1}^{n} \phi_{h_2}\left[y_l - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_l - \mathbf{X}_j)\right\}\right]} w_{ij} \right\} \\
&= \sum_{j=1}^{n} \log \{ \sum_{i=1}^{n} \left(\frac{\phi_{h_2}\left[y_i - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\}\right]}{\sum_{l=1}^{n} \phi_{h_2}\left[y_l - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_l - \mathbf{X}_j)\right\}\right]} \right) \\
&\times \left(\frac{\phi_{h_2}\left[y_i - \left\{a_j^{(k+1)} + \mathbf{b}_j^{(k+1)^T}\mathbf{B}^{(k+1)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\}\right]}{\phi_{h_2}\left[y_i - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\}\right]} w_{ij} \right) \} \\
&= \sum_{j=1}^{n} \log \left\{ \sum_{i=1}^{n} p_{ij}^{(k+1)} \frac{\phi_{h_2}\left[y_i - \left\{a_j^{(k+1)} + \mathbf{b}_j^{(k+1)^T}\mathbf{B}^{(k+1)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\}\right]}{\phi_{h_2}\left[y_i - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\}\right]} \right\},
\end{aligned}
$$

where

$$
p_{ij}^{(k+1)} = \frac{w_{ij}\phi_{h_2}\left[y_i - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\}\right]}{\sum_{l=1}^{n} \phi_{h_2}\left[y_l - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_l - \mathbf{X}_j)\right\}\right]}.
$$

From the Jensen's inequality, we have

$$Q(\boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k)}) \geq \sum_{j=1}^{n} \left[\sum_{i=1}^{n} p_{ij}^{(k+1)} \log \left\{ \frac{\phi_{h_2}\left[y_i - \left\{ a_j^{(k+1)} + \mathbf{b}_j^{(k+1)^T} \mathbf{B}^{(k+1)^T} \left(\mathbf{X}_i - \mathbf{X}_j \right) \right\} \right]}{\phi_{h_2}\left[y_i - \left\{ a_j^{(k)} + \mathbf{b}_j^{(k)^T} \mathbf{B}^{(k)^T} \left(\mathbf{X}_i - \mathbf{X}_j \right) \right\} \right]} \right\} \right].$$

Based on the property of M-step of (2.5), we have $Q(\boldsymbol{\theta}^{(k+1)}) - Q(\boldsymbol{\theta}^{(k)}) \geq 0$

2.5.3 Proof of Theorem 2

The $\hat{\boldsymbol{\theta}} = \{\hat{a}_j, \hat{\mathbf{b}}_j, j = 1, 2, \ldots, n, \hat{\mathbf{B}}\}$ is obtained by maximizing the following objective function

$$\max_{\mathbf{B}, a_j, \mathbf{b}_j, j=1,\ldots,n} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \log \phi_{h_2} \left[y_i - \left\{ a_j + \mathbf{b}_j^T \mathbf{B}^T \left(\mathbf{X}_i - \mathbf{X}_j \right) \right\} \right] w_{ij} \right), \tag{2.12}$$

where $\phi_{h_2} = h_2^{-1}\phi(t/h_2)$. Let $\mathbf{G}$ denote the gradient of $g(\cdot)$ with respect to its arguments, therefore $\mathbf{G}_k(u_1, \ldots, u_D) = \partial g(u_1, \ldots, u_D)/\partial u_k$, $\mathbf{G}^2_{k,l}(u_1, \ldots, u_D) = \partial^2 g(u_1, \ldots, u_D)/(\partial u_k \partial u_l)$, and $\mathbf{G}^3_{k,l,m}(u_1, \ldots, u_D) = \partial^3 g(u_1, \ldots, u_D)/(\partial u_k \partial u_l \partial u_m)$, where $1 \leq k, l, m \leq D$. Let $\mathbf{B}_0 = (\boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0D})$, then based on the Taylor expansion of $g(\mathbf{B}_0^T \mathbf{X}_i)$ for $\mathbf{X}_i$ close to $\mathbf{x}$, we have

$$g(\mathbf{B}_0^T \mathbf{X}_i) = g(\mathbf{B}_0^T \mathbf{x}) + (\mathbf{X}_i - \mathbf{x})^T \mathbf{B}_0 \mathbf{G}(\mathbf{B}_0^T \mathbf{x}) + h^2 \mathbf{A}_{h,i}(\mathbf{x}) + R_i(\mathbf{x}), \tag{2.13}$$

where

$$\mathbf{A}_{h,i}(\mathbf{x}) = \frac{1}{2}\sum_{k,l=1}^{D} \mathbf{G}^2_{k,l}(\mathbf{B}_0^T\mathbf{x}) \left\{\boldsymbol{\beta}_{0k}^T\left(\frac{\mathbf{X}_i-\mathbf{x}}{h}\right)\right\}\left\{\boldsymbol{\beta}_{0l}^T\left(\frac{\mathbf{X}_i-\mathbf{x}}{h}\right)\right\} + \frac{h}{6}\sum_{k,l,m=1}^{D} \mathbf{G}^3_{k,l,m}(\mathbf{B}_0^T\mathbf{x}) \left\{\boldsymbol{\beta}_{0k}^T\left(\frac{\mathbf{X}_i-\mathbf{x}}{h}\right)\right\}\left\{\boldsymbol{\beta}_{0l}^T\left(\frac{\mathbf{X}_i-\mathbf{x}}{h}\right)\right\}\left\{\boldsymbol{\beta}_{0m}^T\left(\frac{\mathbf{X}_i-\mathbf{x}}{h}\right)\right\},$$

and $R_i(\mathbf{x})$ is defined as the remainder.

Since $y_i = g(\mathbf{B}_0^T\mathbf{X}_i) + \epsilon_i$, we have

$$y_i = \mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})\boldsymbol{\Psi}(\mathbf{x},h) + (\mathbf{X}_i-\mathbf{x})^T(I-\mathbf{B}\mathbf{B}^T)\mathbf{B}_0\mathbf{G}(\mathbf{B}_0^T\mathbf{x}) + h^2\mathbf{A}_{h,i}(\mathbf{x}) + R_i(\mathbf{x}) + \epsilon_i, \quad (2.14)$$

where, $\boldsymbol{\Psi}(\mathbf{x},h) = \begin{pmatrix} g(\mathbf{B}_0^T\mathbf{x}) \\ \mathbf{B}^T\mathbf{B}_0\mathbf{G}(\mathbf{B}_0^T\mathbf{x})h \end{pmatrix}$ and $\mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x}) = (1, (\mathbf{X}_i-\mathbf{x})^T\mathbf{B}/h)^T$. Let $K_{h,i}(\mathbf{x}) = K_h(\mathbf{X}_i-\mathbf{x})$ and considering the local modal likelihood based on local linear kernel smooth, we have

$$S_n(\mathbf{B},\mathbf{x}) = \sum_{i=1}^{n} \ell\left\{y_i - \mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})\boldsymbol{\Omega}\right\} K_{h,i}(\mathbf{x}), \quad (2.15)$$

where $\ell(\cdot) = \log\phi_{h_2}(\cdot)$ and $\boldsymbol{\Omega} = \begin{pmatrix} a(\mathbf{x}) \\ \mathbf{b}(\mathbf{x})h \end{pmatrix}$. Let $r_i = y_i - \mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})\boldsymbol{\Psi}(\mathbf{x},h)$, then, based on the Taylor expansion of $S_n(\mathbf{B},\mathbf{x})$ close to $\boldsymbol{\Psi}$ and the bounded third order derivative of $\ell(\cdot)$, we have

$$\begin{aligned}\hat{\boldsymbol{\Omega}} = \boldsymbol{\Psi}(\mathbf{x},h) &+ T_n^{-1}(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n} K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\ell'(\epsilon_i) \\ &+ T_n^{-1}(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n} K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\ell''(\epsilon_i)(r_i-\epsilon_i),\end{aligned} \quad (2.16)$$

where

$$T_n(\mathbf{B},\mathbf{x}) = n^{-1}\sum_{i=1}^{n} K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})\ell''(\epsilon_i)(1+O_p(h^2+\delta_n)), \tag{2.17}$$

with $\delta_n = (nh^p)^{-1/2}(log\ n)^{1/2}$.

Let

$$\begin{aligned}
\Xi_{n,i}(\mathbf{B},\mathbf{x}) &= \mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})T_n^{-1}(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n} K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\ell'(\epsilon_i),\\
P_{n,i}(\mathbf{B},\mathbf{x}) &= \mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})T_n^{-1}(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n} \ell''(\epsilon_i)K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\mathbf{A}_{h,i}(\mathbf{x}),\\
R_{n,i}(\mathbf{B},\mathbf{x}) &= \mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})T_n^{-1}(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n} \ell''(\epsilon_i)K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})R_i(\mathbf{x}),\\
\gamma_{n,i}(\mathbf{B},\mathbf{x}) &= (\mathbf{X}_i-\mathbf{x}_i)^T - \mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})T_n^{-1}(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n} \ell''(\epsilon_i)K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})(\mathbf{X}_i-\mathbf{x})^T.
\end{aligned}$$

By replacing $\hat{\mathbf{\Omega}}$ in $S_n(\mathbf{B},\mathbf{x})$, we have

$$\begin{aligned}
\sum_{j=1}^{n}\frac{S_n(\mathbf{B},\mathbf{X}_j)}{n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{X}_j)} &= \sum_{j=1}^{n}\sum_{i=1}^{n}\xi_{h,i}(\mathbf{X}_j)(\ell(\gamma_{n,i}(\mathbf{B},\mathbf{x})(I-\mathbf{B}\mathbf{B}^T)\mathbf{B}_0\mathbf{G}(\mathbf{B}_0^T\mathbf{X}_j)+\Delta_{ij}(\mathbf{B})\\
&= \sum_{j=1}^{n}\sum_{i=1}^{n}\ell\{\gamma_{n,i}(\mathbf{B},\mathbf{X}_j)(I-\mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0k}\mathbf{G}_k(\mathbf{B}_0^T\mathbf{X}_j)\\
&\quad + \sum_{l\neq k}\gamma_{n,i}(\mathbf{B},\mathbf{X}_j)(I-\mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0l}\mathbf{G}_l(\mathbf{B}_0^T\mathbf{X}_j)+\Delta_{ij}(\mathbf{B})\}\xi_{h,i}(\mathbf{X}_j).
\end{aligned} \tag{2.18}$$

where $\xi_{h,i}(\mathbf{x}) = \frac{K_{h,i}(\mathbf{x})}{n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{x})}$, and

$$\Delta_{ij}(\mathbf{B}) = \epsilon_i + h^2\mathbf{A}_{h,i}(\mathbf{X}_j) + R_i(\mathbf{X}_j) - P_{n,i}(\mathbf{B},\mathbf{X}_j)h^2 - \Xi_{n,i}(\mathbf{B},\mathbf{X}_j) - R_{n,i}(\mathbf{B},\mathbf{X}_j).$$

Following Xia *et al.* (2002), we have

$$(I - \mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0k} = \nu_{k,k}^{-1} \sum_{j=1}^{n}\sum_{i=1}^{n} \left\{ \ell' \left(\sum_{l\neq k} \gamma_{n,i}(\mathbf{B}, \mathbf{X}_j)(I - \mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0l}\mathbf{G}_l(\mathbf{B}_0^T\mathbf{X}_j) + \Delta_{ij}(\mathbf{B}) \right) \xi_{h,i}(\mathbf{X}_j)\mathbf{G}_k(\mathbf{B}_0^T\mathbf{X}_j)\gamma_{n,i}^T(\mathbf{B}, \mathbf{X}_j) \right\}, \tag{2.19}$$

where

$$\begin{aligned} n^{-2}\nu_{k,l} &= -n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n} \ell'' \left(\sum_{l\neq k} \gamma_{n,i}(\mathbf{B}, \mathbf{X}_j)(I - \mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0l}\mathbf{G}_l(\mathbf{B}_0^T\mathbf{X}_j) + \Delta_{ij}(\mathbf{B}) \right) \\ &\quad \times \xi_{h,i}(\mathbf{X}_j)\mathbf{G}_k(\mathbf{B}_0^T\mathbf{X}_j)\mathbf{G}_l(\mathbf{B}_0^T\mathbf{X}_j)\gamma_{n,i}^T(\mathbf{B}, \mathbf{X}_j)\gamma_{n,i}(\mathbf{B}, \mathbf{X}_j) \\ &= -h^2 n^{-1}\sum_{j=1}^{n} \mathbf{G}_k(\mathbf{B}_0^T\mathbf{X}_j)\mathbf{G}_l(\mathbf{B}_0^T\mathbf{X}_j)(I - \mathbf{B}\mathbf{B}^T)E\{\ell''(\epsilon)\} + O_p(h^3 + h\delta_n). \end{aligned} \tag{2.20}$$

We have

$$\begin{aligned} &\ell' \left(\sum_{l\neq k} \gamma_{n,i}(I - \mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0l}\mathbf{G}_l + \Delta_{ij}(\mathbf{B}) \right) \\ =&\ell'(\epsilon_i) + \ell''(\epsilon_i)\sum_{l\neq k} \gamma_{n,i}(\mathbf{B}, \mathbf{X}_j)(I - \mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0l}\mathbf{G}_l(\mathbf{B}_0^T\mathbf{X}_j) + \ell''(\epsilon_i)\left\{\Delta_{ij}(\mathbf{B}_0) - \epsilon_i\right\}, \end{aligned}$$

therefore,

$$\begin{aligned} &-(I - \mathbf{B}\mathbf{B}^T)\sum_{l=1}^{D}\boldsymbol{\beta}_{0l} \left\{ n^{-1}h^2\sum_{j=1}^{n} \mathbf{G}_k(\mathbf{B}_0^T\mathbf{X}_j)\mathbf{G}_l(\mathbf{B}_0^T\mathbf{X}_j)E\{\ell''(\epsilon)\} + O_p(h^3 + h\delta_n) \right\} \\ &= n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n} \ell'(\epsilon_i)\xi_{h,i}(\mathbf{X}_j)\mathbf{G}_k(\mathbf{B}_0^T\mathbf{X}_j)\gamma_{n,i}^T(\mathbf{B}, \mathbf{X}_j) \\ &+ n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n} \ell''(\epsilon_i)\xi_{h,i}(\mathbf{X}_j)\mathbf{G}_k(\mathbf{B}_0^T\mathbf{X}_j)\gamma_{n,i}^T(\mathbf{B}, \mathbf{X}_j)\{\Delta_{ij}(\mathbf{B}_0) - \epsilon_i\} = A_1 + A_2. \end{aligned} \tag{2.21}$$

Now, we are going to check the order of A_1 and A_2, respectively. For the order of A_1, note that,

$$\begin{aligned}
n^{-1}\sum_{i=1}^{n} K_{h,i}(\mathbf{x})\gamma_{n,i}^{T}(\mathbf{B},\mathbf{x})\ell'(\epsilon_i) &= n^{-1}\sum_{i=1}^{n} K_{h,i}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\ell'(\epsilon_i) \\
&- n^{-1}\sum_{i=1}^{n} \ell''(\epsilon_i)K_{h,i}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\mathbf{X}_{h,i}^{T}(\mathbf{B},\mathbf{x})T_n^{-1}(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n} K_{h,i}(\mathbf{x})\ell'(\epsilon_i)\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x}) \\
&= (I-\mathbf{B}\mathbf{B}^T)n^{-1}\left\{\sum_{i=1}^{n} K_{h,i}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\ell'(\epsilon_i) - \frac{h^2}{\phi_{h_2}(\mathbf{x})}\Delta\phi_{h_2}(\mathbf{x})\sum_{i=1}^{n} K_{h,i}(\mathbf{x})\ell'(\epsilon_i)\right\} + O_p(h^3\delta_n + h\delta_n^2) \\
&= O_p(h^3\delta_n + h\delta_n^2),
\end{aligned} \tag{2.22}$$

Next, we will check the order of A_2. First, we have,

$$\begin{aligned}
&n^{-1}\sum_{i=1}^{n} \ell''(\epsilon_i)K_{h,i}(\mathbf{x})\gamma_{n,i}^{T}(\mathbf{B},\mathbf{x})\Xi_{n,i}(\mathbf{B},\mathbf{x}) \\
&= \left\{T_n^{-1}(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n} \ell''(\epsilon_i)K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}^{T}(\mathbf{X}_i-\mathbf{x})^T\right\}^T \times n^{-1}\sum_{i=1}^{n}\mathbf{X}_{h,i}^{T}\ell'(\epsilon_i)O_p(a_n) \\
&= O_p(h^3\delta_n + h\delta_n^2),
\end{aligned}$$

and

$$n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n} \ell''(\epsilon_i)\xi_{h,i}(\mathbf{X}_j)\mathbf{G}_k(\mathbf{B}_0^T\mathbf{X}_j)\gamma_{n,i}^{T}(\mathbf{B},\mathbf{X}_j)\Xi_{n,i}(\mathbf{B},\mathbf{X}_j) = O_p(h^3\delta_n + h\delta_n^2). \tag{2.23}$$

Secondly,

$$n^{-2}h^2\sum_{j=1}^{n}\sum_{i=1}^{n}\ell''(\epsilon_i)\xi_{h,i}(\mathbf{X}_j)\mathbf{G}_k(\mathbf{B}_0^T\mathbf{X}_j)\gamma_{n,i}^T(\mathbf{B},\mathbf{X}_j)[\mathbf{A}_{h,i}(\mathbf{x})-P_{n,i}(\mathbf{B},\mathbf{x})]$$
$$=(I-\mathbf{B}\mathbf{B}^T)\mathbf{B}_0 n^{-1}\sum_{j=1}^{n}\phi_{h_2}^{-1}(\mathbf{X}_j)\left\{\bar{\mathbf{G}}(\mathbf{X}_j)-\frac{1}{2}\sum_{l=1}^{D}\mathbf{G}_{l,l}^2(\mathbf{B}^T\mathbf{X}_j)\mathbf{B}_0\bigtriangledown\phi_{h_2}(\mathbf{x})\right\}E\{\ell''(\epsilon)\}h^4$$
$$+O_p(h^5+h^3\delta_n). \tag{2.24}$$

where $\bar{\mathbf{G}}(\mathbf{B}_0^T\mathbf{x})=\tilde{\mathbf{G}}^2(\mathbf{B}_0^T\mathbf{x})\mathbf{B}_0^T\bigtriangledown\phi_{h_2}(\mathbf{x})+\tilde{\mathbf{G}}^3(\mathbf{B}_0^T\mathbf{x})$ with $\tilde{\mathbf{G}}^2(\mathbf{B}_0^T\mathbf{x})$ being $D\times D$ matrix of the upper left part of $\sum_{m,l=1}^{D}\left\{\mathbf{G}_{m,l}^2(\mathbf{B}_0^T\mathbf{x})\int K(U)UU^Tu_lu_mdU\right\}$, $\kappa_4=\int u^4K(u)du$ and

$$\mathbf{B}_0\tilde{\mathbf{G}}^3(\mathbf{B}_0^T\mathbf{x})=\frac{1}{6}\left\{\sum_{l=1}^{D}\mathbf{G}_{l,l,l}^3(\mathbf{B}_0^T\mathbf{x})\kappa_4\boldsymbol{\beta}_{0l}+\sum_{m\neq l}\mathbf{G}_{m,m,l}^3(\mathbf{B}_0^T\mathbf{x})\boldsymbol{\beta}_{0l}\right\}.$$

Similarly, we have

$$n^{-1}\sum_{i=1}^{n}\ell''(\epsilon_i)K_{h,i}(\mathbf{x})\gamma_{n,i}^T(\mathbf{B},\mathbf{x})[R_i(\mathbf{x})-R_{n,i}(\mathbf{B},\mathbf{x})]=O_p(h^5).$$

Hence,

$$A_2=(I-\mathbf{B}\mathbf{B}^T)\mathbf{B}_0 n^{-1}\sum_{j=1}^{n}\phi_{h_2}^{-1}(\mathbf{X}_j)\left\{\bar{\mathbf{G}}(\mathbf{X}_j)-\frac{1}{2}\sum_{l=1}^{D}\mathbf{G}_{l,l}^2(\mathbf{B}^T\mathbf{X}_j)\mathbf{B}_0\bigtriangledown\phi_{h_2}(\mathbf{x})\right\}E\{\ell''(\epsilon)\}h^4$$
$$+O_p(h^5+h^3\delta_n+h\delta_n^2). \tag{2.25}$$

Form (2.22) and (2.25), for $k = 1, \ldots, D$, we have

$$
\begin{aligned}
&-(I - \mathbf{B}\mathbf{B}^T) \sum_{l=1}^{D} \beta_{0l} \left\{ n^{-1} h^2 \sum_{j=1}^{n} \mathbf{G}_k(\mathbf{B}_0^T \mathbf{X}_j) \mathbf{G}_l(\mathbf{B}_0^T \mathbf{X}_j) E\{\ell''(\epsilon)\} + O_p(h^3 + h\delta_n) \right\} \\
&\quad = (I - \mathbf{B}\mathbf{B}^T) \mathbf{B}_0 n^{-1} \sum_{j=1}^{n} \phi_{h_2}^{-1}(\mathbf{X}_j) \left\{ \bar{\mathbf{G}}(\mathbf{X}_j) - \frac{1}{2} \sum_{l=1}^{D} \mathbf{G}_{l,l}^2(\mathbf{B}^T \mathbf{X}_j) \mathbf{B}_0 \Delta \phi_{h_2}(\mathbf{x}) \right\} E\{\ell''(\epsilon)\} h^4 \\
&\quad + O_p(h^5 + h^3 \delta_n + h\delta_n^2),
\end{aligned}
$$

therefore,

$$
(I - \mathbf{B}\mathbf{B}^T) \mathbf{B}_0 n^{-1} \sum_{j=1}^{n} \mathbf{G}(\mathbf{B}_0^T \mathbf{X}_j) \mathbf{G}^T(\mathbf{B}_0^T \mathbf{X}_j) E\{\ell''(\epsilon)\} = O_p(h^3 + h\delta_n + h^{-1}\delta_n^2).
$$

Since $n^{-1} \sum_{j=1}^{n} \mathbf{G}(\mathbf{B}_0^T \mathbf{X}_j) \mathbf{G}^T(\mathbf{B}_0^T \mathbf{X}_j) = O_p(1)$, we have

$$
||(I - \mathbf{B}\mathbf{B}^T) \mathbf{B}_0|| = O_p(h^3 + h\delta_n + h^{-1}\delta_n^2).
$$

Chapter 3

Sparse Adaptive MAVE

Due to the explosion of massive data in the last decades, high dimensional data analysis has attracted considerable attention in many scientific fields. There exists a large number of model-based variable selection approaches in literature, such as C_p, AIC, BIC, etc. These criteria measure the quality of a statistical model by penalizing the model complexity if non-informative variables were added. To deal with the instability that effects these traditional measures (Breiman, 1996), several regularization methods, such as Nonnegative Garrote (Breiman, 1995), LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), Lars (Efron *et al.*, 2004) and Elastic Net (Zou and Hastie, 2005), were proposed to automatically select informative variables through continuous shrinkage. However, due to the so-called "curse of dimensionality" it is very difficult or even infeasible to formulate and validate a parametric model with a large number of variables.

Li *et al.* (2005) proposed a model-free variable selection method through sufficient dimension reduction (SDR). The basic idea of SDR is to replace the original high dimensional predictor with an appropriate low dimensional projection without losing regression informa-

tion (Li 1991; Cook 1998). By incorporating shrinkage estimation into SDR, we can achieve dimension reduction and variable selection simultaneously without assuming any particular model. Along this line, many methods have been proposed recently, such as Ni, Cook and Tsai (2005), Li and Nachtsheim (2006), Li (2007), Zhou and He (2008), Wang and Yin (2008), and Bondell and Li (2009).

Minimum average variance estimation (MAVE; Xia *et al.*, 2002) is a popular SDR method for both dimension reduction and nonparametric function estimation. Wang and Yao (2012) introduced an adaptive estimation for MAVE (aMAVE), which combines the kernel density estimation and MAVE such that aMAVE can be adaptive to different error distributions. Although aMAVE is efficient under non-normal error distributions, each reduced variable is still a linear combination of all the original predictors. In this work, we combine aMAVE with shrinkage estimation to propose a new variable selection method, sparse aMAVE (saMAVE). SaMAVE extends shrinkage estimation to multi-dimensional and nonlinear settings, without any particular model assumption.

3.1 A Brief review of adaptive MAVE

The regression-type model of interest in adaptive MAVE can be written as

$$y = g(\mathbf{B}_0^T \mathbf{X}) + \epsilon, \tag{3.1}$$

where $g(\cdot)$ is an unknown smooth link function, $\mathbf{B}_0 = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ is an orthogonal matrix ($\mathbf{B}_0^T \mathbf{B}_0 = \mathbf{I}_d$) with $d \leq p$ and $E(\epsilon|\mathbf{X}) = 0$. Here, $\mathbf{B}_0$ forms a basis of the central mean subspace in SDR literature and d is called the structural dimension. In most real world

applications, the error may not be normally distributed. Therefore, it will be logical to treat the error density as another unknown parameter similar to the link function. Under this setting, Wang and Yao (2012) proposed the adaptive MAVE (aMAVE) to estimate the $\mathbf{B}_0$ assuming d is known. Let $f_\epsilon(\epsilon)$ be the density function of ϵ. If f_ϵ is known, the direction $\mathbf{B}_0$ can be estimated via

$$\max_{\substack{\mathbf{B}:\mathbf{B}^T\mathbf{B}=\mathbf{I}_d \\ a_j,\mathbf{b}_j;j=1,\ldots,n}} \sum_{j=1}^{n}\sum_{i=1}^{n} \log f_\epsilon \left[y_i - \left\{a_j + \mathbf{b}_j^T\mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_j\right)\right\}\right] w_{ij}, \tag{3.2}$$

where w_{ij} are some weights defined as a function of the distance between $\mathbf{X}_i$ and $\mathbf{X}_j$. In practice, f_ϵ is usually unknown but can be estimated by

$$\tilde{f}_\epsilon(\epsilon) = \frac{1}{2}K_{h_1}(\epsilon - \tilde{\epsilon}_i), \tag{3.3}$$

where $K_{h_1}(\nu) = h_1^{-1}K(\nu/h_1)$ with $K(\nu)$ being a kernel function and h_1 being the bandwidth. Thus, based on some initial residual estimate $\{\tilde{\epsilon}_i, i = 1,\ldots,n\}$, aMAVE maximizes the following objective function

$$\sum_{j=1}^{n}\sum_{i=1}^{n} \log\left(\sum_{l=1}^{n} K_{h_1}\left[y_i - \left\{a_j + \mathbf{b}_j^T\mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_j\right)\right\} - \tilde{\epsilon}_l\right]\right) w_{ij}. \tag{3.4}$$

Wang and Yao (2012) proposed an efficient EM algorithm to estimate $\mathbf{B}$. The choice of a Gaussian kernel for $K(\cdot)$ gives the nice quadratic function from (3.4) such that the least squares based MAVE estimation can be adopted here. More details can be found in Wang and Yao (2012) and the references therein.

3.2 Sparse adaptive MAVE (saMAVE)

To select the informative covariates, a bridge penalty can be added to the expression (3.4) following the suggestion in Wang *et al.* (2013). That is, we maximize the following objective function

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\log\left(\sum_{l=1}^{n}K_{h_1}\left[y_i-\left\{a_j+\mathbf{b}_j^T\mathbf{B}^T\left(\mathbf{X}_i-\mathbf{X}_j\right)\right\}-\tilde{\epsilon}_l\right]\right)w_{ij}-\lambda_n\sum_{k=1}^{p}||\beta_k||_1^{\gamma}, \tag{3.5}$$

where β_k is the kth row of $\mathbf{B}$, $||\cdot||_1$ represents the L_1 norm, λ_n is the nonnegative regularization parameter, and $\gamma\in(0,1)$. By adopting a bridge penalty for the L_1 norms of the rows of $\mathbf{B}$, it is possible to carry out variable screening and element screening simultaneously. More details can be found in Wang *et al.* (2013).

3.2.1 Computation Algorithm

For a given sample $\{(y_i,\mathbf{x}_i), i=1,2,\ldots,n\}$,

1. obtain an initial estimator $\{\hat{\mathbf{B}},(\hat{a}_j,\hat{\mathbf{b}}_j), j=1,2,\ldots,n\}$. This initial estimate can be obtained from the traditional MAVE method;

2. for a given $\hat{\mathbf{B}}$ and corresponding $\{\tilde{\epsilon}_i, i=1,\ldots,n\}$, update $\{(a_j,\mathbf{b}_j), j=1,2,\ldots,n\}$ from the following maximization problem

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\log\left(\sum_{l=1}^{n}K_{h_1}\left[y_i-\left\{a_j+\mathbf{b}_j^T\hat{\mathbf{B}}^T\left(\mathbf{X}_i-\mathbf{X}_j\right)\right\}-\tilde{\epsilon}_l\right]\right)w_{ij}; \tag{3.6}$$

3. for given $\{(\hat{a}_j,\hat{\mathbf{b}}_j),\ j=1,2,\ldots,n\}$, solve $\mathbf{B}$ from the following constrained maximiza-

tion problem

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\log\left(\sum_{l=1}^{n}K_{h_1}\left[y_i-\left\{a_j+\mathbf{b}_j^T\mathbf{B}^T\left(\mathbf{X}_i-\mathbf{X}_j\right)\right\}-\tilde{\epsilon}_l\right]\right)w_{ij}-\lambda_n\sum_{k=1}^{p}||\beta_k||_1^{\gamma}, \quad (3.7)$$

4. iterate between the previous step 2 and 3 until convergence in the estimation of $\mathbf{B}$.

Through the EM algorithm in aMAVE (Wang and Yao, 2012) and the equivalent form of the above group bridge penalty (Wang *et al.*, 2013), we can adopt the least squares based MAVE estimation and the LASSO algorithm to implement our proposed method.

3.2.2 Tuning Parameter Selection

We employed a very efficient Lasso algorithm recently proposed by Friedman et al. (2010) to solve the L_1 regularized maximization (3.5). Cyclical coordinate descent methods were used to calculate the solution path for a large number of λ at once. We used the Matlab package glmnet in all the simulation studies. More details can be found at http://www-stat.stanford.edu/∼tibs/glmnet-matlab/. A BIC criterion was used to select the optimal λ in the Lasso estimation,

$$BIC_\lambda = n\ \log\left(\frac{\mathbf{Q}_\lambda}{n}\right) - log(n)p_\lambda \quad (3.8)$$

where

$$\mathbf{Q}_\lambda = \sum_{j=1}^{n}\sum_{i=1}^{n}\log\left(\sum_{l=1}^{n}K_{h_1}\left[y_i-\left\{\hat{a}_j+\hat{\mathbf{b}}_j^T\hat{\mathbf{B}}^T(\mathbf{X}_i-\mathbf{X}_j)\right\}-\tilde{\epsilon}_l\right]\right)w_{ij}$$

is similar to the residual sum of squares from the Lasso fit, and p_λ denotes the number of non-zero coefficients in $\hat{\mathbf{B}}$.

Suppose, without loss of generality, assume that only the first $q < p$ predictor are relevant to the response variable and let $A_1 = \{1, 2, \ldots, q\}$ denotes the relevant predictors and $A_2 = \{q+1, q+2, \ldots, p\}$ denotes the irrelevant predictors. Therefore, we can define $\boldsymbol{\beta}_{A_1} = \boldsymbol{\beta}_k I(k \in A_1)$ and $\boldsymbol{\beta}_{A_2} = \boldsymbol{\beta}_k I(k \in A_2)$ for $k = 1, 2, \ldots, p$ where $I(\cdot)$ shows the indicator function.

In order to prove theorem 1, we need the following condition to hold:

$$\tilde{\mathbf{B}} = \mathbf{B}_0 + O_p\left(\frac{1}{\sqrt{n}}\right), \tag{3.9}$$

where $\tilde{\mathbf{B}}$ is an initial estimate for $\mathbf{B}_0$.

Theorem : Assume $\gamma \in (0, 1)$, $d \leq 3$, and the regularity condition in the appendix condition holds. Then we have the followings:

(i) If $\lambda_n n^{1/2} = O(1)$, then there exists a local maximizer $\hat{\mathbf{B}}$ for $Q(\theta)$ such that

$$||\hat{\mathbf{B}} - \mathbf{B}_0||_2 = O_p\left(\frac{1}{\sqrt{n}}\right), \tag{3.10}$$

where $||\cdot||_2$ represents the L_2 norm.

(ii) If $\lambda_n n^{1/2} = O(1)$ and $\lambda_n n^{-\gamma/2} \to \infty$, then $P(\hat{\beta}_{A_2} = 0) \to 1$ as $n \to \infty$.

Proof can be found in the Appendix.

3.3 Simulation study

In this section, we carried out the simulation study to evaluate the finite sample performance of the proposed sparse adaptive MAVE (saMAVE) and to compare it with the traditional refined MAVE (rMAVE; Xial *et al.*, 2002), sparse MAVE (sMAVE; Wang and Yin, 2008), and adaptive MAVE (aMAVE; Yao and Wang, 2012).

The data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ were generated from the model

$$Y = \frac{\beta_1^T X_1}{0.5 + (\beta_2^T X_2 + 1.5)^2} + 0.5\epsilon, \tag{3.11}$$

where $X_i \in R^p$ where p=5 and 10 and $\beta_1 = (1, 0, \ldots, 0)^T$, $\beta_2 = (0, 1, \ldots, 0)^T$. Four error distributions of ϵ were investigated:

1. N(0,1), the standard normal errors. This density serves as a benchmark with no outliers;
2. $t_3/\sqrt{3}$, the scaled t-distribution with 3 degree of freedom;
3. $0.5N(-1, 0.5^2) + 0.5N(1, 0.5^2)$;
4. $0.3N(-1.4, 1) + 0.7N(0.6, 0.4^2)$;

Sample size was chosen as $n = 50, 100, 200$ and we drew 500 data replicates in each case. We considered both independent and correlated cases for $\mathbf{X}$: (a) $\mathbf{X} \sim N_p(\mathbf{0}_p, \mathbf{I}_p)$ and (b) $\mathbf{X} \sim N_p(\mathbf{0}_p, \Sigma)$, where $\Sigma = [\sigma_{i,j}]$ where $\sigma_{i,j} = 0.5^{|i-j|}$. In order to compare different estimators, we used vector correlation coefficient r^2 defined by Ye and Weiss (2003). Consider S(A) and S(B) denote two d-dimensional spaces where A and B are orthonormal bases of the two

spaces, respectively. The vector correlation coefficient is defined as $r^2 = \frac{1}{2}trace(B^T AA^T B)$. Furthermore, we employ the true positive rate (TPR): the ratio of the number of correctly identified active predictors to the number of truly active predictors, and the false positive rate (FPR): the ratio of the number of falsely identified active predictors to the number of inactive predictors, for measuring the performance in selecting active variables. Table 3.1 to 3.4 represent the results of these simulation for i.i.d and correlated data, respectively. In our simulation, we used our bandwidth as $h = 1.06n^{-0.2}\sigma$ where $\sigma = \min\{(q_{0.75} - q_{0.25})/1.34, \sigma_\epsilon\}$ where q_i is i-th quantile of the distribution of ϵ and σ_ϵ shows the standard deviation of the error. More details can be found in Wang *et al.* (2007) and references therein.

Table 3.1: *Estimation accuracy comparison based on the vector correlation coefficient defined as* $r^2 = \frac{1}{2}tr(B^T AA^T B)$ *for independent predictors when p=5.*

ϵ	n	Estimation accuracy based on mean				TPR & FPR	
		rMave	SMAVE	aMAVE	saMAVE	sMAVE	saMAVE
1	50	0.8064	0.9076	0.8015	0.9035	(0.923 , 0.094)	(0.756 , 0.105)
	100	0.9397	0.9850	0.9418	0.9699	(0.940 , 0.074)	(0.889 , 0.037)
	200	0.9822	0.9951	0.9816	0.9898	(0.940 , 0.068)	(0.900 , 0.037)
2	50	0.8387	0.9236	0.8396	0.9138	(0.926 , 0.094)	(0.799 , 0.104)
	100	0.9344	0.9816	0.9356	0.9623	(0.938 , 0.075)	(0.881 , 0.049)
	200	0.9781	0.9947	0.9765	0.9818	(0.940 , 0.070)	(0.920 , 0.059)
3	50	0.9549	0.9786	0.9546	0.9861	(0.939 , 0.115)	(0.861 , 0.055)
	100	0.9873	0.9930	0.9871	0.9970	(0.940 , 0.099)	(0.906 , 0.042)
	200	0.9941	0.9962	0.9939	0.9914	(0.940 , 0.089)	(0.924 , 0.046)
4	50	0.9477	0.9720	0.9471	0.9782	(0.938 , 0.110)	(0.868 , 0.061)
	100	0.9851	0.9921	0.9849	0.9847	(0.940 , 0.095)	(0.913 , 0.040)
	200	0.9938	0.9963	0.9937	0.9954	(0.940 , 0.085)	(0.934 , 0.040)

Tables 3.1 to 3.4 report the estimation accuracy comparison based on the average r^2, TPR, and FPR for simulated model with different combinations of sample size and dimensions and various error distributions f_ϵ, respectively. The first four horizontal blocks in these tables show the results under measures r^2 and it shows that saMAVE are consistently better than rMAVE, sMAVE, and aMAVE. High TPR and low FPR again illustrate the accuracy of saMAVE. TPR and FPR are not defined for rMAVE and aMAVE since they do not select

Table 3.2: *Estimation accuracy comparison based on the vector correlation coefficient defined as* $r^2 = \frac{1}{2}tr(B^T AA^T B)$ *for correlated predictors when p=5.*

ϵ	n	Estimation accuracy based on mean				TPR & FPR	
		rMave	SMAVE	aMAVE	saMAVE	sMAVE	saMAVE
1	50	0.7232	0.8158	0.7178	0.8361	(0.930 , 0.147)	(0.774 , 0.104)
	100	0.8492	0.9417	0.8504	0.9439	(0.940 , 0.142)	(0.909 , 0.088)
	200	0.9331	0.9741	0.9308	0.9724	(0.940 , 0.166)	(0.958 , 0.108)
2	50	0.7454	0.8362	0.7446	0.8366	(0.929 , 0.150)	(0.832 , 0.123)
	100	0.8562	0.9278	0.8540	0.9078	(0.940 , 0.154)	(0.894 , 0.120)
	200	0.9201	0.9667	0.9191	0.9593	(0.940 , 0.169)	(0.961 , 0.106)
3	50	0.8947	0.9353	0.8964	0.9466	(0.939 , 0.177)	(0.922 , 0.136)
	100	0.9495	0.9621	0.9495	0.9823	(0.940 , 0.197)	(0.963 , 0.169)
	200	0.9587	0.9634	0.9583	0.9809	(0.940 , 0.222)	(0.980 , 0.180)
4	50	0.8777	0.9227	0.8781	0.9454	(0.939 , 0.175)	(0.925 , 0.127)
	100	0.9448	0.9622	0.9444	0.9809	(0.940 , 0.190)	(0.970 , 0.162)
	200	0.9578	0.9637	0.9575	0.9810	(0.940 , 0.223)	(0.979 , 0.174)

Table 3.3: *Estimation accuracy comparison based on the vector correlation coefficient defined as* $r^2 = \frac{1}{2}tr(B^T AA^T B)$ *for independent predictors when p=10.*

ϵ	n	Estimation accuracy based on mean				TPR & FPR	
		rMave	SMAVE	aMAVE	saMAVE	sMAVE	saMAVE
1	50	0.5541	0.7687	0.5534	0.7350	(0.885 , 0.190)	(0.618 , 0.087)
	100	0.7931	0.9661	0.7897	0.9211	(0.949 , 0.093)	(0.738 , 0.048)
	200	0.9389	0.9937	0.9367	0.9746	(0.960 , 0.083)	(0.886 , 0.014)
2	50	0.6089	0.7915	0.6064	0.7617	(0.893 , 0.207)	(0.658 , 0.084)
	100	0.8210	0.9583	0.8079	0.8780	(0.953 , 0.106)	(0.740 , 0.066)
	200	0.9301	0.9915	0.9313	0.9607	(0.959 , 0.086)	(0.857 , 0.025)
3	50	0.8332	0.9311	0.8292	0.9448	(0.957 , 0.275)	(0.767 , 0.033)
	100	0.9557	0.9842	0.9560	0.9881	(0.960 , 0.238)	(0.853 , 0.020)
	200	0.9834	0.9930	0.9831	0.9925	(0.960 , 0.236)	(0.914 , 0.017)
4	50	0.7950	0.9312	0.7929	0.9298	(0.956 , 0.249)	(0.754 , 0.035)
	100	0.9464	0.9840	0.9470	0.9806	(0.960 , 0.217)	(0.863 , 0.030)
	200	0.9815	0.9935	0.9814	0.9926	(0.960 , 0.215)	(0.909 , 0.013)

Table 3.4: *Estimation accuracy comparison based on the vector correlation coefficient defined as* $r^2 = \frac{1}{2}tr(B^T AA^T B)$ *for correlated predictors when p=10.*

ϵ	n	Estimation accuracy based on mean				TPR & FPR	
		rMave	SMAVE	aMAVE	saMAVE	sMAVE	saMAVE
1	50	0.5128	0.6503	0.5058	0.6940	(0.814 , 0.172)	(0.626 , 0.100)
	100	0.6767	0.8264	0.6798	0.8026	(0.878 , 0.161)	(0.775 , 0.097)
	200	0.8452	0.9654	0.8414	0.9308	(0.880 , 0.145)	(0.921 , 0.106)
2	50	0.5458	0.6860	0.5486	0.7303	(0.841 , 0.183)	(0.656 , 0.104)
	100	0.6786	0.8338	0.6730	0.7831	(0.875 , 0.160)	(0.754 , 0.109)
	200	0.8370	0.9482	0.8369	0.8998	(0.878 , 0.144)	(0.907 , 0.112)
3	50	0.7180	0.8231	0.7133	0.8782	(0.879 , 0.228)	(0.808 , 0.104)
	100	0.8877	0.9541	0.8834	0.9520	(0.880 , 0.221)	(0.942 , 0.128)
	200	0.9440	0.9666	0.9429	0.9829	(0.880 , 0.192)	(0.978 , 0.105)
4	50	0.6895	0.8071	0.6857	0.8683	(0.879 , 0.219)	(0.797 , 0.096)
	100	0.8631	0.9481	0.8666	0.9453	(0.880 , 0.206)	(0.915 , 0.129)
	200	0.9409	0.9701	0.9401	0.9877	(0.880 , 0.177)	(0.971 , 0.073)

individual variables. As expected, results for all methods improve as n increases from 50 to 100 to 200. From the summary of these tables, we can see that the proposed saMAVE is comparable to the rMAVE, sMAVE, and aMAVE for normal errors but more efficient than them when the error is non-normal.

3.4 Real data analysis

This data set concerns the salary of 263 baseball hitters in 1987 and their performance. The question of interest is "Are they paid based on their performance?". This data set was analyzed by many statisticians. Chaudhuri *et al.* (1994) proposed a piecewise polynomial regression tree (SUPPORT) approach. Li *et al.* (2000) proposed a dimension reduction based regression tree, PHDRT, and identified several outliers. Xia *et al.* (2002) applied MAVE to find the low dimensional projection and chose a partially linear model to fit the data. All previous studies suggested using different models to fit different parts of the data. Wang and Yao (2012) used adaptive MAVE to analyze this data set. Similar to other authors, we

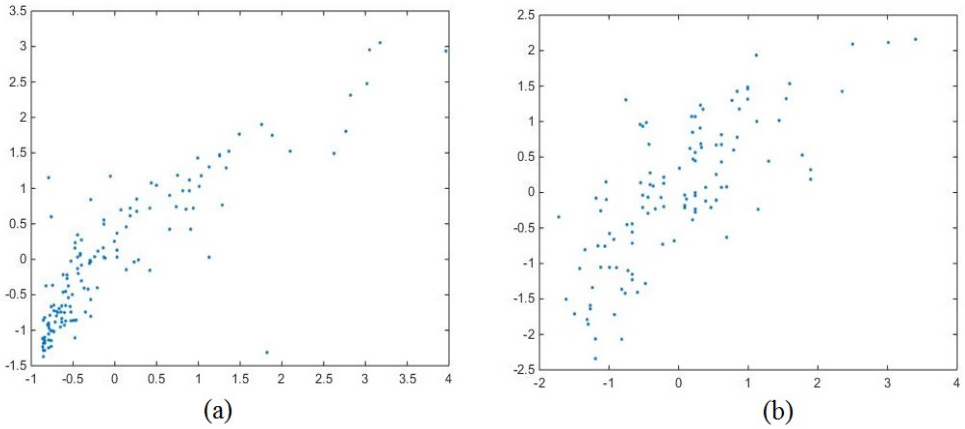

Figure 3.1: *Baseball hitter's salary data against (a) Junior (b) Veterans.*

split the data into two groups (junior/veteran) based on the years in the major leagues and the cutoff is chosen to be 7 as suggested by Chaudhuri *et al.* (1994). The response variable is the logarithm of the annual salary in 1987 as in all previous studies, and the 13 predictors used in our analysis are listed in Table 3.5. We apply the sparse adaptive MAVE to both groups, and one significant direction is identified for each group as shown in Table 3.5.

Table 3.5: *The estimated CS directions for baseball hitters data.*

	Performance		$\hat{\beta}_{junior}$	$\hat{\beta}_{veterans}$
1986	x_1	time at bat	-0.017	0.2999
	x_2	hits	0	0
	x_3	home runs	-0.1343	0
	x_4	runs	0.1727	0
	x_5	runs batted	0	0
	x_6	walks	-0.0192	0
Up to 1986	x_7	years in major leagues	0.9728	-0.3214
	x_8	time at bat	-0.0083	0
	x_9	hits	0.0152	0.9267
	x_{10}	home runs	0	0
	x_{11}	runs	0.0095	0.2264
	x_{12}	runs batted	0.0657	0
	x_{13}	walks	0.021	0

As we can see in table 3.5, for juniors' time at bat, home runs, runs, walks, and years in the major leagues are important in 1986. Furthermore, time at bat, hits, runs, runs batted,

and walks are important factor for up to 1986. For veterans, time at bat and years in major leagues are important in 1986. In addition, the number of hits and runs are important for up to 1986. The sign change of the coefficient estimates for the variable x7 (years in the major league) between the two groups supports the existence of an aging effect as discovered by Li *et al.* (2000), Xia *et al.* (2002) and Wang and Yao (2012)

3.5 Theoretical result

3.5.1 Regularity conditions

The following technical conditions are imposed in this section:

(A1) $\{(\mathbf{X}_i, y_i), i = 1, \ldots, n\}$ are i.i.d. samples from the joint density $f_{\mathbf{X},y}(\mathbf{x}, y)$.

(A2) $\{\epsilon_i\}$ are i.i.d. with $E(\epsilon_i) = 0,\ E(|\epsilon_i|^3) < \infty$. $\{\mathbf{X}_i\}$ and $\{\epsilon_i\}$ are mutually independent. Additionally, the predictor $\mathbf{X}$ has a bounded support.

(A3) $E|y|^k < \infty$ and $E||\mathbf{X}||^k < \infty$ for all $k > 0$.

(A4) $E(\mathbf{X}|y)$ and $E(\mathbf{X}\mathbf{X}^T|y)$ have bounded, continuous 3rd derivatives.

(A5) $K_h(\cdot)$ is a spherical symmetric density function with a bounded derivative and support. Specifically, we used Gaussian kernel with bandwidth $h \propto n^{-\frac{1}{d+4}}$.

(A6) The density $f_\epsilon(\cdot)$ has bounded continuous derivatives up to order 4. Let $\ell(\cdot) = \log f_\epsilon(\cdot)$. Assume $\ell'''(\cdot)$ is bounded and $E\{\ell'(\epsilon)^2 + |\ell''(\epsilon)| + |\ell'''(\epsilon)|\} < \infty$.

(A7) The smallest eigenvalue of $J^T_{\boldsymbol{\beta}_0^{(1)}} \mathbf{W}_{g0} J_{\boldsymbol{\beta}_0^{(1)}}$ is larger than ρ amd the largest eigenvalue of $\mathbf{W}_{g0}$ is less than ρ^* for some positive constant ρ and ρ^* where $\mathbf{W}_{g0} = E\left[\{(\mu_{\mathbf{B}_0}(\mathbf{X}) - \mathbf{x})(\mu_{\mathbf{B}_0}(\mathbf{X}) - \mathbf{x})^T\right.$ where $\mu_{\mathbf{B}_0}(\mathbf{X}) = E(\mathbf{X}|\mathbf{B}_0^T\mathbf{X} = \mathbf{B}_0^T\mathbf{x})$ and $J_{\boldsymbol{\beta}_0^{(1)}}$ is defined in the proof.

The above conditions are imposed to facilitate the proof and most of them are similar to Xia et al. (2002), Wang and Xia (2008), Wang et al. (2013), and Wang and Yao (2012).

3.5.2 Proof of Theorem 1

Note that the estimate $\boldsymbol{\theta} = \{\mathbf{B}, (a_j, \mathbf{b}_j), j = 1, 2, \ldots, n\}$ is the maximizer of the following objective function

$$\max_{\mathbf{B}, a_j, \mathbf{b}_j; j=1,\ldots,n} \sum_{j=1}^{n}\sum_{i=1}^{n} \log f_\epsilon \left[y_i - \left\{a_j + \mathbf{b}_j^T \mathbf{B}^T (\mathbf{X}_i - \mathbf{X}_j)\right\}\right] w_{ij} - \lambda_n \sum_{k=1}^{p} ||\beta_k||_1^\gamma. \tag{3.12}$$

where

$$\tilde{f}_\epsilon(\epsilon) = \frac{1}{2} K_{h_1}(\epsilon - \tilde{\epsilon}_i)$$

is the kernel density estimate of $f_\epsilon(\cdot)$ and $\tilde{\epsilon}_i$ is the residual based on the traditional MAVE estimate. Based on the adaptive nonparametric regression result of Linton and Xiao (2007), the convergence rate of $\hat{\boldsymbol{\theta}}$ in (3.12) is the same as the true density $f_\epsilon(\cdot)$ is used. Since the basic idea of our proof is very similar to Wang et al. (2013), we adopt the same notations for the ease of readers to follow. Therefore, we will mainly prove the theorems assuming $f_\epsilon(\cdot)$ is known. Furthermore, let $\ell(\cdot) = \log f_\epsilon(\cdot)$ and let $\boldsymbol{\beta} = vec(\mathbf{B}^T)$ where $vec(\cdot)$ is a matrix operator that stacks all columns of a matrix into a vector. Using *vec* operator, we can rewrite (3.12) as follow,

$$Q(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \lambda_n \sum_{k=1}^{p} ||\beta_k||_1^\gamma. \tag{3.13}$$

where $L(\boldsymbol{\theta}) = \sum_{j=1}^{n}\sum_{i=1}^{n} \ell\left(y_i - \left\{a_j + \left[(\mathbf{X}_i - \mathbf{X}_j)^T \otimes \mathbf{b}_j^T\right]\boldsymbol{\beta}\right\}\right) w_{ij}$.

Since $||\boldsymbol{\beta}_0||_2 = 1$, therefore $g(\boldsymbol{\beta}_0\mathbf{X})$ does not have a derivative at point $\boldsymbol{\beta}_0$. Using "delete-one-component" method, we define $\mathbf{J}_{\boldsymbol{\beta}_0^{(1)}}$ where $\boldsymbol{\beta}_0^{(1)}$ consists of all free parameters in $\boldsymbol{\beta}_0$. Let $\boldsymbol{\beta}^{*(1)} = \boldsymbol{\beta}_0^{(1)} + n^{-1/2}\boldsymbol{\eta}$, where $||\boldsymbol{\eta}||_2 = C$ for some positive constant C. Using Taylor expansion on $L(\boldsymbol{\beta}_0)$ around $\boldsymbol{\beta}_0^*$ and substituting into the $Q(\cdot)$, we have

$$\begin{aligned}
\frac{1}{n}(Q(\boldsymbol{\beta}^*) - Q(\boldsymbol{\beta}_0)) &= -n^{1/2}\boldsymbol{\eta}^T\mathbf{J}^T_{\boldsymbol{\beta}_0^{*(1)}}\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\boldsymbol{\Omega}_{ij}\ell'\left(y_i - \left\{a_j + \left[(\mathbf{X}_i - \mathbf{X}_j)^T \otimes \mathbf{b}_j^T\right]\boldsymbol{\beta}_0\right\}\right) w_{ij} \\
&\quad + \frac{1}{2}\boldsymbol{\eta}^T\mathbf{J}^T_{\boldsymbol{\beta}_0^{*(1)}}\mathbf{J}_{\boldsymbol{\beta}_0^{*(1)}}\boldsymbol{\eta}\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\boldsymbol{\Omega}_{ij}\ell''\left(y_i - \left\{a_j + \left[(\mathbf{X}_i - \mathbf{X}_j)^T \otimes \mathbf{b}_j^T\right]\boldsymbol{\beta}_0\right\}\right)\boldsymbol{\Omega}_{ij}^T w_{ij} \\
&\quad + \lambda_n\sum_{k=1}^{p}(||\beta_{0k}||_1^{\gamma} - ||\beta_k^*||_1^{\gamma}) + o_P\left(\frac{1}{\sqrt{n}}\right) \\
&= T_1 + T_2 + T_3 + o_P\left(\frac{1}{\sqrt{n}}\right),
\end{aligned} \tag{3.14}$$

where $\boldsymbol{\Omega}_{ij} = \left((\mathbf{X}_i - \mathbf{X}_j)^T \otimes \mathbf{b}_j^T\right)$ and $\boldsymbol{\beta}_0^{*(1)}$ is between $\boldsymbol{\beta}_0^{(1)}$ and $\boldsymbol{\beta}^{*(1)}$.

Following Wang and Xia (2008), we have

$$\begin{aligned}
T_1 &= -n^{1/2}\boldsymbol{\eta}^T\mathbf{J}^T_{\boldsymbol{\beta}_0^{*(1)}}\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n}\boldsymbol{\Omega}_{ij}\ell'\left(y_i - \left\{a_j + \left[(\mathbf{X}_i - \mathbf{X}_j)^T \otimes \mathbf{b}_j^T\right]\boldsymbol{\beta}_0\right\}\right) w_{ij} \\
&= -n^{1/2}\boldsymbol{\eta}^T\mathbf{J}^T_{\boldsymbol{\beta}_0^{*(1)}}\left[E\left[\left\{\nu_{\boldsymbol{\beta}_0}(\mathbf{X})\nu^T_{\boldsymbol{\beta}_0}(\mathbf{X})\right\}\ell''(\epsilon)\right](\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + n^{-1}\sum_{i=1}^{n}\{\nu_{\boldsymbol{\beta}_0}(\mathbf{x}_i)\ell'(\epsilon)\}\epsilon_i\right] \\
&\quad + o_P\left(\frac{1}{\sqrt{n}}\right)
\end{aligned} \tag{3.15}$$

where $\nu_{\boldsymbol{\beta}_0}(\mathbf{x}) = \mu(\mathbf{x}) - \mathbf{x}$ with $\mu(\mathbf{x}) = E(\mathbf{X}|\boldsymbol{\beta}_0^T\mathbf{X} = \boldsymbol{\beta}_0^T\mathbf{x})$ and $\tilde{\boldsymbol{\beta}}$ is an initial estimator of

$\boldsymbol{\beta}_0$). Furthermore, we have

$$\begin{aligned} T_2 &= \boldsymbol{\eta}^T \mathbf{J}^T_{\boldsymbol{\beta}_0^{*(1)}} \mathbf{J}_{\boldsymbol{\beta}_0^{*(1)}} \boldsymbol{\eta} \sum_{j=1}^{n}\sum_{i=1}^{n} \boldsymbol{\Omega}_{ij} \ell''\left(y_i - \left\{a_j + \left[(\mathbf{X}_i - \mathbf{X}_j)^T \otimes \mathbf{b}_j^T\right]\boldsymbol{\beta}_0\right\}\right)\boldsymbol{\Omega}_{ij}^T w_{ij} \\ &= \boldsymbol{\eta}^T \mathbf{J}^T_{\boldsymbol{\beta}_0^{*(1)}} \left[E\left[\left\{\nu_{\boldsymbol{\beta}_0}(\mathbf{X})\nu^T_{\boldsymbol{\beta}_0}(\mathbf{X})\right\}\ell''(\epsilon)\right] + o_P(1)\right]\mathbf{J}_{\boldsymbol{\beta}_0^{*(1)}}\boldsymbol{\eta} \end{aligned} \tag{3.16}$$

By assuming that the smallest eigenvalue of $\mathbf{J}^T_{\boldsymbol{\beta}_0^{*(1)}} E\left[\left\{\nu_{\boldsymbol{\beta}_0}(\mathbf{X})\nu^T_{\boldsymbol{\beta}_0}(\mathbf{X})\right\}\ell''(\epsilon)\right]\mathbf{J}_{\boldsymbol{\beta}_0^{*(1)}}$ is larger than ρ where ρ is some positive number, we have $T_2 \geq \rho||\boldsymbol{\eta}||_2^2$. Furthermore, since we assumed $||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 = O_P\left(\frac{1}{\sqrt{n}}\right)$, we have

$$n^{1/2}\boldsymbol{\eta}^T \mathbf{J}^T_{\boldsymbol{\beta}_0^{*(1)}}\left[E\left[\left\{\nu_{\boldsymbol{\beta}_0}(\mathbf{X})\nu^T_{\boldsymbol{\beta}_0}(\mathbf{X})\right\}\ell''(\epsilon)\right](\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + n^{-1}\sum_{i=1}^{n}\{\nu_{\boldsymbol{\beta}_0}(\mathbf{x}_i)\ell'(\epsilon)\}\epsilon_i\right] = O_P\left(\frac{1}{\sqrt{n}}\right) \tag{3.17}$$

Thus, by choosing a sufficiently large C, T_1 is dominated by T_2. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} T_3 = \lambda_n \sum_{k=1}^{q}\left(||\beta_{0k}||_1^{\gamma} - ||\beta_k^*||_1^{\gamma}\right) &\leq 2\lambda_n \sum_{k=1}^{q}\left(||\beta_{0k}||_1^{\gamma-1}||\beta_{0k} - \beta_k^*||_1\right) \\ &\leq 2\sqrt{d}\lambda_n \sum_{k=1}^{q}\left(||\beta_{0k}||_1^{\gamma-1}||\beta_{0k} - \beta_k^*||_2\right) \\ &\leq 2\sqrt{d}\sum_{k=1}^{q}||\beta_{0k}||_1^{\gamma-1}\lambda_n\left(\sum_{k=1}^{q}||\beta_{0k} - \beta_k^*||_2^2\right)^{\frac{1}{2}} \\ &= O(1)\lambda_n\frac{||\boldsymbol{\eta}||_2}{\sqrt{n}} = O(||\boldsymbol{\eta}||_2) \end{aligned} \tag{3.18}$$

where the last equality holds because $\lambda_n = O(n^{1/2})$. Therefore, if C is sufficiently large, then T_3 is also dominated by T_2. Thus, with a large probability, we have $Q(\boldsymbol{\beta}_0^*) \geq Q(\boldsymbol{\beta}_0)$ when $\{\boldsymbol{\eta} : ||\boldsymbol{\eta}||_2 = C\}$ which means there exists a local maximizer of $Q(\boldsymbol{\beta})$ in the ball $\{\boldsymbol{\eta} : ||\boldsymbol{\eta}||_2 \leq C\}$ such that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 = O_P\left(\frac{1}{\sqrt{n}}\right)$. This proves the first part of the theorem.

In the second part of the theorem, we want to prove the variable selection consistency of the proposed method. Suppose, with no loss of generality, only the first $q < p$ predictor are relevant to the response variable and let $A_1 = \{1, 2, \ldots, q\}$ denotes the relevant predictors and $A_2 = \{q+1, q+2, \ldots, p\}$ denotes the irrelevant predictors. Then, we can define $\bar{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\beta}}_k I(k \in A_1)$ for $k = 1, 2, \ldots, p$ where $I(\cdot)$ shows the indicator function. Following Wang et al., (2013) and the Karush-Kuhn-Tucker condition, we have

$$\sum_{j=1}^{n}\sum_{i=1}^{n} \boldsymbol{\Omega}_{ij}^{l} \ell' \left(y_i - \left\{ a_j + \left[(\mathbf{X}_i - \mathbf{X}_j)^T \otimes \mathbf{b}_j^T \right] \hat{\boldsymbol{\beta}} \right\} \right) w_{ij} = \hat{\psi}^{1-\frac{1}{\gamma}} sgn(\hat{\beta}_k^l), \hat{\beta}_k^l \neq 0, \qquad (3.19)$$

where $\boldsymbol{\Omega}_{ij}^{l}$ is the lth component of $\boldsymbol{\Omega}_{ij}$, $\psi = \left(\frac{1-\gamma}{\tau\gamma}\right)^{\gamma} ||\beta_k||_1^{\gamma}$ where τ is defined in Wang et al. (2013), and $sgn(\cdot)$ denotes the sign function which is defined as follow

$$sgn(x) = \begin{cases} \text{-1} & \text{if } x < 0 \\ 0 & \text{if x=0} \\ 1 & \text{if } x > 0 \end{cases}$$

Since $\hat{\psi}^{1-\frac{1}{\gamma}} ||\hat{\beta}_k||_1 = \gamma\lambda_n ||\hat{\beta}_k||_1^{\gamma}$, we have

$$\sum_{j=1}^{n}\sum_{i=1}^{n} \boldsymbol{\Omega}_{ij}^{l} \ell' \left(y_i - \left\{ a_j + \left[(\mathbf{X}_i - \mathbf{X}_j)^T \otimes \mathbf{b}_j^T \right] \hat{\boldsymbol{\beta}} \right\} \right) w_{ij} = \gamma\lambda_n ||\hat{\beta}_k||_1^{\gamma} sgn(\hat{\beta}_k^l), \hat{\beta}_k^l \neq 0. \quad (3.20)$$

Therefore, we have

$$\begin{aligned}
&\sum_{j=1}^{n}\sum_{i=1}^{n}\ell'\left(y_i-\left\{a_j+\left[(\mathbf{X}_i-\mathbf{X}_j)^T\otimes\mathbf{b}_j^T\right]\hat{\boldsymbol{\beta}}\right\}\right)w_{ij}\boldsymbol{\Omega}_{ij}^T(\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}})\\
&=\sum_{j=1}^{n}\sum_{i=1}^{n}\ell'\left(y_i-\left\{a_j+\left[(\mathbf{X}_i-\mathbf{X}_j)^T\otimes\mathbf{b}_j^T\right]\hat{\boldsymbol{\beta}}\right\}\right)w_{ij}\sum_{k,l:\hat{\beta}_k^l\neq 0}\boldsymbol{\Omega}_{ij}^l(\hat{\beta}_k^l-\bar{\beta}_k^l)\\
&=\sum_{k,l:\hat{\beta}_k^l\neq 0}\gamma\lambda_n||\hat{\beta}_k||_1^{\gamma}sgn(\hat{\beta}_k^l)(\hat{\beta}_k^l-\bar{\beta}_k^l)=\sum_{k,l}\gamma\lambda_n||\hat{\beta}_k||_1^{\gamma}sgn(\hat{\beta}_k^l)(\hat{\beta}_k^l-\bar{\beta}_k^l)\\
&=\sum_{k\in A_2}\gamma\lambda_n||\hat{\beta}_k||_1^{\gamma-1}\sum_{l=1}^{d}|\hat{\beta}_k^l|=\sum_{k=1}^{p}\gamma\lambda_n||\hat{\beta}_k||_1^{\gamma-1}(||\hat{\beta}_k^l||_1-||\bar{\beta}_k^l||_1),
\end{aligned}$$

where the last two equality holds because $(\hat{\beta}_k^l-\bar{\beta}_k^l)sgn(\hat{\beta}_k^l)=|\hat{\beta}_k^l|I(k\in A_2)$. Furthermore, because for $0\leq a\leq b$ we have $\gamma b^{\gamma-1}(b-1)\leq b^{\gamma}-a^{\gamma}$, thus

$$\gamma||\hat{\beta}_k||_1^{\gamma-1}(||\hat{\beta}_k^l||_1-||\bar{\beta}_k^l||_1)\leq||\hat{\beta}_k^l||_1^{\gamma}-||\bar{\beta}_k^l||_1^{\gamma}.$$

Therefore, we have

$$\begin{aligned}
&\left|\sum_{j=1}^{n}\sum_{i=1}^{n}\ell'\left(y_i-\left\{a_j+\left[(\mathbf{X}_i-\mathbf{X}_j)^T\otimes\mathbf{b}_j^T\right]\hat{\boldsymbol{\beta}}\right\}\right)w_{ij}\boldsymbol{\Omega}_{ij}^T(\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}})\right|\\
&\qquad\leq\lambda_n\sum_{k=1}^{q}(||\hat{\beta}_k^l||_1^{\gamma}-||\bar{\beta}_k^l||_1^{\gamma})+\gamma\lambda_n\sum_{k=q+1}^{p}||\hat{\beta}_k^l||_1^{\gamma}.
\end{aligned}$$

Since by definition of $\hat{\boldsymbol{\beta}}$, we have $Q_{\lambda_n}(\hat{\boldsymbol{\beta}}) \leq Q_{\lambda_n}(\bar{\boldsymbol{\beta}})$, then

$$\begin{aligned}
&\left|\sum_{j=1}^{n}\sum_{i=1}^{n}\ell'\left(y_i - \left\{a_j + \left[(\mathbf{X}_i - \mathbf{X}_j)^T \otimes \mathbf{b}_j^T\right]\hat{\boldsymbol{\beta}}\right\}\right) w_{ij}\boldsymbol{\Omega}_{ij}^T(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})\right| + (1-\gamma)\lambda_n \sum_{k=q+1}^{p} ||\hat{\beta}_k^l||_1^\gamma \\
&\leq \lambda_n \sum_{k=1}^{q}(||\hat{\beta}_k^l||_1^\gamma - ||\bar{\beta}_k^l||_1^\gamma) \leq \lambda_n \sum_{k=1}^{q}||\hat{\beta}_k^l||_1^\gamma - \lambda_n \sum_{k=1}^{q}||\bar{\beta}_k^l||_1^\gamma \\
&\leq \sum_{j=1}^{n}\sum_{i=1}^{n}\ell'\left(y_i - \left\{a_j + \left[(\mathbf{X}_i - \mathbf{X}_j)^T \otimes \mathbf{b}_j^T\right]\hat{\boldsymbol{\beta}}\right\}\right) w_{ij}\boldsymbol{\Omega}_{ij}^T(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) \\
&- \frac{1}{2}(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})^T \sum_{j=1}^{n}\sum_{i=1}^{n}\boldsymbol{\Omega}_{ij}\ell''\left(y_i - \left\{a_j + \left[(\mathbf{X}_i - \mathbf{X}_j)^T \otimes \mathbf{b}_j^T\right]\hat{\boldsymbol{\beta}}\right\}\right) w_{ij}\boldsymbol{\Omega}_{ij}^T(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})
\end{aligned}$$

Therefore, in probability, we have

$$\begin{aligned}
(1-\gamma)\lambda_n \sum_{k=q+1}^{p} ||\hat{\beta}_k^l||_1^\gamma &\leq \frac{1}{2}(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})^T \sum_{j=1}^{n}\sum_{i=1}^{n}\boldsymbol{\Omega}_{ij}\ell''\left(y_i - \left\{a_j + \left[(\mathbf{X}_i - \mathbf{X}_j)^T \otimes \mathbf{b}_j^T\right]\hat{\boldsymbol{\beta}}\right\}\right) w_{ij}\boldsymbol{\Omega}_{ij}^T(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) \\
&\leq n\rho^* ||\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}||_2^2 = n\rho^* \sum_{k=q+1}^{p} ||\hat{\boldsymbol{\beta}}_k||_2^2 \\
&\leq n\rho^* ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2^2.
\end{aligned}$$

From the result of the first part of the theorem, we have

$$(1-\gamma)\lambda_n \sum_{k=q+1}^{p} ||\hat{\beta}_k^l||_1^\gamma \leq n\rho^* \sum_{k=q+1}^{p} ||\hat{\boldsymbol{\beta}}_k||_2^2 = O_P(1), \tag{3.21}$$

and

$$\sum_{k=q+1}^{p} ||\hat{\boldsymbol{\beta}}_k||_2^2 \geq \left(\sum_{k=q+1}^{p} ||\hat{\boldsymbol{\beta}}_k||_1\right)^\gamma \geq \left(\sum_{k=q+1}^{p} ||\hat{\boldsymbol{\beta}}_k||_1\right)^{\gamma/2}. \tag{3.22}$$

From (3.21) and (3.22), if $\sum_{k=q+1}^{p} ||\hat{\boldsymbol{\beta}}_k||_2^2 > 0$ then

$$(1-\gamma)\lambda_n \leq n\rho^* \left(\sum_{k=q+1}^{p} ||\hat{\boldsymbol{\beta}}_k||_2^2\right)^{1-\gamma/2} = n\rho^* O_P(1)(n\rho^*)^{-1+\gamma/2} = O_P(n^{\gamma/2}).$$

Since, $\lambda_n n^{-\gamma/2} \to \infty$, we have

$$P(\left(\sum_{k=q+1}^{p} ||\hat{\boldsymbol{\beta}}_k||_2^2 > 0 \right) \to 0 \quad \text{as} \quad n \to \infty. \tag{3.23}$$

This result completes the proof.

Chapter 4

Spatial Envelope

In many research areas such as health science (Lave and Seskin, 1973; Liang *et al.* 1992), epidemiology (Lekkou *et al.* 2014), business (Cooper *et al.* 2003), etc. it is common to observe a large number of simultaneous outcomes. The traditional Multivariate Linear Regression (MLR) has proved to be the standard analysis for this type of data to understand the relationship between response variables and regressors. Mathematically, the MLR model is typically given as:

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}, \tag{4.1}$$

where $\mathbf{Y} \in \mathbb{R}^p$ denotes the response vector, $\mathbf{X} \in \mathbb{R}^r$ is a vector predictor, $\boldsymbol{\alpha} \in \mathbb{R}^p$ denotes vector of intercept coefficients, $\boldsymbol{\beta} \in \mathbb{R}^{(p\times r)}$ is the matrix of regression coefficients, and $\boldsymbol{\epsilon} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ is an error vector with $\boldsymbol{\Sigma} \geq 0$ being an unknown covariance matrix (Christensen, 2001). In order to completely specify the MLR, there are p unknown parameters to specify the intercept, $r \times p$ unknown regression coefficient parameters, and $\frac{p(p+1)}{2}$ unknown variance-covariance parameters in an unstructured covariance matrix. Therefore, in practice, one must

estimate $p + pr + \frac{p(p+1)}{2}$ model parameters which will necessitate a large number of samples. The large number of parameters also leads to other problems such as the computational time required to estimate the parameters.

There are cases where the distribution of some linear combinations of the response vector $\mathbf{Y}$ do not depend on any of the predictors in $\mathbf{X}$, which are called *immaterial*. While the distribution of the other linear combinations of $\mathbf{Y}$ depend on $\mathbf{X}$ which are called *material*. For instance, one unique case that may arise in multivariate regression is when some of the regression coefficients are zero for all predictors on a few of the response variables. This means those responses do not depend on any of the predictors. Mathematically, this model is given as:

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\beta}^* \end{pmatrix} \mathbf{X} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{pmatrix},$$

which means the distribution of $\mathbf{Y}_1$ does not depend on any of the predictors in $\mathbf{X}$.

Based on this idea, Cook *et al.* (2010) proposed the *envelope* method as a new version of the classical multivariate linear model to account for the material and immaterial structure in the data. This approach separates $\mathbf{Y}$ into material ($\mathbf{Y}_1^*$) and immaterial ($\mathbf{Y}_0^*$). For the immaterial responses only the intercept is needed and hence one can reduce the number of parameters needed. This approach allows for gains in efficiency by reducing the variance of the estimate of the regression coefficients compared to the standard maximum likelihood estimates of the full model (Cook *et al.*, 2010). The envelope attempts to construct a link between the mean function and covariance matrix using a minimal reducing subspace such that the resulting number of parameters will maximally reduce. This mean the reduced subspace is the smallest subspace on which one can project the original space without losing regression information. Cook *et al.* (2010) showed that the envelope estimator for the re-

gression coefficients has asymptotically smaller variance compared to the standard maximum likelihood estimator (MLE) of MLR.

Current envelope methodology assumes observations are taken under identical conditions where independence is assured. While models based on the independence assumption are extremely useful, their use is limited in applications where the data has inherent dependency (Cressie, 2015). Spatially correlated data are one example and found in a wide range of application domains such as network screening in highway safety (Jonathan *et al.*, 2016), ecology (Rota, 2016), forensic science (Proença *et al.*, 2016), image processing (Rigaux *et al.*, 2001), etc. In many of these settings, there is an increasing need to analyze multivariate measurements obtained at spatial locations (Latimer *et al.*, 2009). An example for these types of data is environmental monitoring where each station collects data concerning several pollutants such as ozone, carbon monoxide, nitrogen dioxide, etc. The goal of this work is to extend the envelope methodology to situations where spatially correlated data are the norm.

4.1 Spatial Envelope

Suppose $\mathbf{Y}(s)$ is a multivariate response taken at site s. We assume that the data generating process is second order stationary and the covariance of the response vectors $\mathbf{Y}(s_i)$ and $\mathbf{Y}(s_j)$ at two sites s_i and s_j is given by a function of distance between the two sites. Namely the covariance can be written as:

$$Cov(\mathbf{Y}(s_i), \mathbf{Y}(s_j)) = C_{ij}(\mathbf{h}), \quad \mathbf{h} = ||s_i - s_j|| \in R^d, \tag{4.2}$$

where $||\cdot||$ denotes Euclidean distance. $C_{ij}(\cdot)$ is called the direct covariogram if $i = j$ and the cross-covariogram if $i \neq j$ and the matrix-valued function $C(\mathbf{h}) = [C_{ij}(\mathbf{h})]$ is the multivariate covariogram.

One simple model for the covariogram of a multivariate spatial response is the *proportional correlation model* (Chiles and Delfiner, 1999) given by:

$$C(\mathbf{h}) = \mathbf{V}\rho(\mathbf{h}),$$

where $\mathbf{V}$ is a $p \times p$ positive definite matrix and $\rho(\mathbf{h})$ is any valid correlation function. The proportional covariogram is also known as the *intrinsic covariogram* (Wackernagel, 2013). Using this covariance model, the covariance of response variables i.e. $\mathbf{\Sigma_Y}$ can be written as $\mathbf{V} \otimes \boldsymbol{\rho}(\mathbf{h}, \boldsymbol{\theta})$, where $\boldsymbol{\rho}(\mathbf{h}, \boldsymbol{\theta})$ is the $n \times n$ matrix with the (i, j)-th entry $\rho(||s_i - s_j||, \boldsymbol{\theta})$, and $\otimes$ denotes the Kronecker product.

Suppose the response vector can be decomposed into the material and immaterial part, $\mathbf{Y}_1$ and $\mathbf{Y}_0$, respectively. Using the envelope idea, $\mathbf{V}$ can be written as $\mathbf{V}_0 + \mathbf{V}_1$ where $\mathbf{V}_0\mathbf{V}_1 = 0$ where $\mathbf{V}_0$ denotes the covariance matrix associated with the immaterial part of response and $\mathbf{V}_1$ denotes the covariance matrix associated with the material part. Hence, the covariance matrix of $\mathbf{Y}$ can be written as follows:

$$\begin{aligned}\mathbf{\Sigma_Y} &= \mathbf{V} \otimes \rho(\mathbf{h}, \boldsymbol{\theta}) \\ &= \mathbf{V}_0 \otimes \rho(\mathbf{h}, \boldsymbol{\theta}) + \mathbf{V}_1 \otimes \rho(\mathbf{h}, \boldsymbol{\theta}), \quad \mathbf{h} \in R^d.\end{aligned}$$

This fact will be used later in derivation for different formulas in the appendix. For simplicity of notation, $\rho(\mathbf{h}, \boldsymbol{\theta})$ is denoted by $\rho(\boldsymbol{\theta})$. The multivariate spatial linear model can be written

as:

$$\mathbf{Y}(s) = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}(s) + \boldsymbol{\epsilon}(s). \tag{4.3}$$

where

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{X}(s) = \begin{pmatrix} \mathbf{X} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{X} \end{pmatrix} = \mathbf{I}_p \otimes \mathbf{X}(s), \quad \mathbf{Y}(s) = \begin{pmatrix} \mathbf{Y}_1^T \\ \vdots \\ \mathbf{Y}_p^T \end{pmatrix}$$

The likelihood function of this model will be as follows:

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{V}, \boldsymbol{\theta}) &= [\det(\mathbf{V} \otimes \boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{1}{2}} \\ &\times exp\left\{ -\frac{1}{2}(\mathbf{Y} - \boldsymbol{\alpha} \otimes \mathbf{1}_n - (\mathbf{I}_p \otimes \mathbf{X})\boldsymbol{\beta})^T (\mathbf{V} \otimes \boldsymbol{\rho}(\boldsymbol{\theta}))^{-1} (\mathbf{Y} - \boldsymbol{\alpha} \otimes \mathbf{1}_n - (\mathbf{I}_p \otimes \mathbf{X})\boldsymbol{\beta}) \right\}. \end{aligned} \tag{4.4}$$

Following the envelope idea by Cook *et al.*, (2010), this likelihood for fixed dimension u, where $0 < u < r$ denotes the dimension of the envelope, can be rewritten as follows

$$L^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{V}_0, \mathbf{V}_1, \boldsymbol{\theta}) = L_1^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{V}_1, \boldsymbol{\theta}) \times L_2^{(u)}(\boldsymbol{\alpha}, \mathbf{V}_0, \boldsymbol{\theta}), \tag{4.5}$$

where

$$\begin{aligned} L_1^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{V}_1, \boldsymbol{\theta}) &= [det_0(\mathbf{V}_1)]^{-\frac{n}{2}} [\det(\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{p}{2}} \\ &\times exp\left\{ -\frac{1}{2}(\mathbf{Y} - \boldsymbol{\alpha} \otimes \mathbf{1}_n - (\mathbf{I}_p \otimes \mathbf{X})\boldsymbol{\beta}))^T \left(\mathbf{V}_1^{\dagger} \otimes \boldsymbol{\rho}(\boldsymbol{\theta})^{-1} \right) (\mathbf{Y} - \boldsymbol{\alpha} \otimes \mathbf{1}_n - (\mathbf{I}_p \otimes \mathbf{X})\boldsymbol{\beta})) \right\}, \\ L_2^{(u)}(\boldsymbol{\alpha}, \mathbf{V}_0, \boldsymbol{\theta}) &= [det_0(\mathbf{V}_0)]^{-\frac{n}{2}} [\det(\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{p}{2}} \\ &\times exp\left\{ -\frac{1}{2}(\mathbf{Y} - \boldsymbol{\alpha} \otimes \mathbf{1}_n)^T \left(\mathbf{V}_0^{\dagger} \otimes \boldsymbol{\rho}(\boldsymbol{\theta})^{-1} \right) (\mathbf{Y} - \boldsymbol{\alpha} \otimes \mathbf{1}_n) \right\}, \end{aligned} \tag{4.6}$$

where † denotes the Moore-Penrose inverse and $det_0(\mathbf{A})$ denotes the product of non-zero

eigenvalues of $\mathbf{A}$ where $\mathbf{A}$ is a non-zero symmetric matrix. The likelihood can be factored because the $span(\boldsymbol{\beta}) \subseteq span(\mathbf{V}_1)$, therefore $\mathbf{V}_0\boldsymbol{\beta} = 0$ and $\mathbf{V} = \mathbf{V}_0 + \mathbf{V}_1$. This factorization is detailed in the appendix.

The objective is to maximize the likelihood in (4.5) over $\boldsymbol{\beta}, \mathbf{V}_0, \mathbf{V}_1$, and $\boldsymbol{\theta}$ subject to the constraints:

$$\begin{aligned} &span(\boldsymbol{\beta}) \subseteq span(\mathbf{V}_1), \quad (a) \\ &\mathbf{V}_0\mathbf{V}_1 = 0, \qquad\qquad\quad (b). \end{aligned} \tag{4.7}$$

The coordinate free version of this maximization is detailed in the theoretical results section. Here, we are presenting the coordinate version of the algorithm.

The optimization depends on being able to maximize the logarithm of $\mathbf{D}$ over the Grassmann manifold $\mathbb{G}^{p\times u}$, where

$$\mathbf{D} = \det(\mathbf{P}_{\mathbf{V}_1}\hat{\boldsymbol{\Sigma}}_{\mathbf{res}}\mathbf{P}_{\mathbf{V}_1} + \mathbf{Q}_{\mathbf{V}_1}\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}}\mathbf{Q}_{\mathbf{V}_1})$$

As mentioned by Cook *et al.* (2010), the gradient-based algorithms for Grassmann optimization (Edelman *et al.*, 1998; Liu *et al.*, 2004) require a coordinate version of the objective function which must have continuous directional derivatives. Let $\hat{\boldsymbol{\Gamma}}_1$ and be the semi-orthogonal bases for $span(\mathbf{V}_1)$ and $\hat{\boldsymbol{\Gamma}}_0$ be the semi-orthogonal bases for $span(\mathbf{V}_0)$. Then $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Gamma}}_1^T\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\Omega}}_1 = \hat{\boldsymbol{\Gamma}}_1^T\hat{\boldsymbol{\Sigma}}_{res}\hat{\boldsymbol{\Gamma}}_1$ and $\hat{\boldsymbol{\Omega}}_0 = \hat{\boldsymbol{\Gamma}}_0^T\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}}\hat{\boldsymbol{\Gamma}}_0$. Let $\log\det(\cdot)$ denote the composite function $\log \circ \det(\cdot)$. Then, the coordinate form of the $\log\mathbf{D}$

$$\log\mathbf{D} = \log\det\left(\boldsymbol{\Gamma}_1^T\left(\mathbf{H}^T\hat{\boldsymbol{\rho}}(\boldsymbol{\theta})\mathbf{H} - \mathbf{H}^T\hat{\boldsymbol{\rho}}(\boldsymbol{\theta})\mathbf{G}\left(\mathbf{G}^T\hat{\boldsymbol{\rho}}^{-1}(\boldsymbol{\theta})\mathbf{G}\right)^{-1}\mathbf{G}^T\hat{\boldsymbol{\rho}}(\boldsymbol{\theta})\mathbf{H}\right)\boldsymbol{\Gamma}_1 + \boldsymbol{\Gamma}_0^T(\mathbf{H}^T\hat{\boldsymbol{\rho}}(\boldsymbol{\theta})\mathbf{H})\boldsymbol{\Gamma}_0\right) \tag{4.8}$$

where $\mathbf{H} = \mathbf{Y} - \bar{\mathbf{Y}} \otimes \mathbf{1}_n$, $\mathbf{U} = vec(\mathbf{H})$, $\mathbf{G} = \mathbf{X} - \bar{\mathbf{X}} \otimes \mathbf{1}_n$, and $\mathbf{F} = \mathbf{I} \otimes \mathbf{G}$. The objective

function (4.8) can be maximized by the coordinate version of spatial envelope using following algorithm.

Algorithm

1. Obtain an initial value for $\hat{\mathbf{\Sigma}}_{\mathbf{Y}}$, $\hat{\mathbf{\Sigma}}_{\mathbf{res}}$, and $\hat{\boldsymbol{\beta}}_{MLE}$, the marginal covariance matrix of $\mathbf{Y}$, the residual covariance matrix, and the maximum likelihood estimate for $\boldsymbol{\beta}$ from the fit of the full model (4.3).

2. Estimate $\mathbf{P}_{\mathbf{V}_1}$ by minimizing the objective function (4.8) over the Grassmann manifold $\mathbb{G}^{(r\times u)}$, and estimate $\mathbf{P}_{\mathbf{V}_0}$ by $\hat{\mathbf{P}}_{\mathbf{V}_0} = \mathbf{I} - \hat{\mathbf{P}}_{\mathbf{V}_1}$.

3. Fix $\boldsymbol{\theta}$ and estimate $\mathbf{V}_0$ and $\mathbf{V}_1$ by $\hat{\mathbf{V}}_0 = \hat{\mathbf{P}}_{\mathbf{V}_0}\hat{\mathbf{\Sigma}}_{\mathbf{Y}}\hat{\mathbf{P}}_{\mathbf{V}_0}$ and $\hat{\mathbf{V}}_1 = \hat{\mathbf{P}}_{\mathbf{V}_1}\hat{\mathbf{\Sigma}}_{\mathbf{res}}\hat{\mathbf{P}}_{\mathbf{V}_1}$.

4. Fix $\mathbf{V}_0$ and $\mathbf{V}_1$ and maximize $L^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{V}_0, \mathbf{V}_1, \boldsymbol{\theta})$ over $\boldsymbol{\theta}$ by solving the following maximization problem:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \Big\{ -\frac{p}{2}\det(\boldsymbol{\rho}(\boldsymbol{\theta})) - \frac{1}{2} tr\left(\left(\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}} \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H} \right) \mathbf{V}_1^{\dagger} \left(\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}} \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H} \right)^T + \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H}\mathbf{V}_0^{\dagger}\mathbf{H}^T\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \right). \tag{4.9}$$

5. Update $\hat{\mathbf{\Sigma}}_{\mathbf{Y}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{res}}$.

6. Iterate between step (2) through step (5) until the matrix norm between mth and $(m+1)$th iteration can be used to compare with some pre-specified tolerance value i.e. $||\mathbf{\Theta}^{m+1} - \mathbf{\Theta}^m|| < \delta$ where $\mathbf{\Theta} = \{\boldsymbol{\theta}, \mathbf{V}_0, \mathbf{V}_1\}$.

7. Estimate $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}} = \hat{\mathbf{P}}_{\mathbf{V}_1}\hat{\boldsymbol{\beta}}_{MLE}$.

As mentioned in Cook *et al.*, (2010) it is possible for an objective function that is defined on Grassmann manifolds to have multiple local optimal points. One way to check this is to run the simulation with different starting values and compare their results. However, after using this approach we have not found the local optima to be a problem for our method.

4.2 Asymptotic Variance

The parameters of spatial envelope model in equation (4.3), without loss of generality $\boldsymbol{\alpha}$ is not included, can be combined into the vector as follows:

$$\phi = \begin{bmatrix} vec(\boldsymbol{\eta}) \\ vec(\boldsymbol{\Gamma}_1) \\ vech(\boldsymbol{\Omega}_1) \\ vech(\boldsymbol{\Omega}_0) \end{bmatrix} \equiv \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{bmatrix} \tag{4.10}$$

where the $vec(\cdot)$ denotes the vector operator and $vech(\cdot)$ denotes vector half operator. For background on these operators, see Seber (2008) and Harville (2008). Here we focus on the following estimable functions under the spatial envelope model:

$$\psi(\phi) = \begin{bmatrix} vec(\boldsymbol{\beta}) \\ vech(\boldsymbol{\Sigma}) \end{bmatrix} = \begin{bmatrix} vec(\boldsymbol{\Gamma}_1\boldsymbol{\eta}) \\ vech\left((\boldsymbol{\Gamma}_1\boldsymbol{\Omega}\boldsymbol{\Gamma}_1^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) \otimes \boldsymbol{\rho}(\boldsymbol{\theta})\right) \end{bmatrix} \equiv \begin{bmatrix} \psi_1(\phi) \\ \psi_2(\phi) \end{bmatrix} \tag{4.11}$$

Let

$$\Psi = \begin{bmatrix} \frac{\partial \psi_1}{\phi_1^T} & \cdots & \frac{\partial \psi_1}{\phi_4^T} \\ \frac{\partial \psi_2}{\phi_1^T} & \cdots & \frac{\partial \psi_2}{\phi_4^T} \end{bmatrix} \tag{4.12}$$

denote the gradient matrix. Following Cook *et al.*, (2010) and using the result of Shapiro (1986), we have following theorem.

Theorem 1: Suppose $\bar{\mathbf{X}} = 0$ and $\mathbf{J}$ is the Fisher information for $\psi(\phi)$ in the model (4.3):

$$\mathbf{J} = \begin{bmatrix} \mathbf{\Sigma_X} \otimes \mathbf{V}^{-1} \otimes \rho(\boldsymbol{\theta})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{E}_r^T \left(\mathbf{V}^{-1} \otimes \boldsymbol{\rho}(\boldsymbol{\theta})^{-1} \otimes \mathbf{V}^{-1} \otimes \rho(\boldsymbol{\theta})^{-1}\right) \mathbf{E}_r \end{bmatrix} \tag{4.13}$$

where $\mathbf{\Sigma_X} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T$, and $\mathbf{E}_r \in R^{r^2 \times r(r+1)/2}$ is expansion matrix which is defined such that for a given matrix such as $\mathbf{A}$, $vec(\mathbf{A}) = \mathbf{E}_r vech(\mathbf{A})$. Let $\mathbf{\Lambda} = \mathbf{J}^{-1}$ be the asymptotic variance of the MLE under the full model. Then

$$\sqrt{n}(\hat{\phi} - \phi) \to N(\mathbf{0}, \mathbf{\Lambda}_0) \tag{4.14}$$

where $\mathbf{\Lambda}_0 = \Psi(\Psi^T \mathbf{\Lambda} \Psi)^{\dagger} \Psi$. Furthermore, $\mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{\Lambda} - \mathbf{\Lambda}_0)\mathbf{\Lambda}^{-\frac{1}{2}} \geq 0$, so the spatial envelope model decreases the asymptotic variance. Proof of this theorem may be found in the appendix. □

This theorem shows that using the spatial envelope will lead to an estimate of the parameters with smaller variance compared to the ML estimator.

4.3 Prediction

Prediction of the response variables at a new unsampled location is often a major objective of a study. Let $\mathbf{Y}_{new}$ be the $vec(\mathbf{Y}_{new})$ of the new multivariate response at unsampled location.

The model in this case can be written as:

$$\begin{pmatrix} \mathbf{Y}_{new} \\ \mathbf{Y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha} \otimes \mathbf{1}_{n_{new}} + \boldsymbol{\Gamma}_1 \boldsymbol{\eta}(\mathbf{I}_p \otimes \mathbf{X}_{new}) \\ \boldsymbol{\alpha} \otimes \mathbf{1}_n + \boldsymbol{\Gamma}_1 \boldsymbol{\eta}(\mathbf{I}_p \otimes \mathbf{X}) \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_{new} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N \left(\boldsymbol{\alpha} + \boldsymbol{\Gamma}_1 \boldsymbol{\eta} \begin{pmatrix} \mathbf{X}_{new} \\ \mathbf{X} \end{pmatrix}, \boldsymbol{\Sigma} \right). \tag{4.15}$$

where $\boldsymbol{\Sigma}$ is as follows

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} = \begin{pmatrix} (\mathbf{V}_0 + \mathbf{V}_1) \otimes \boldsymbol{\rho}_{new,new}(\boldsymbol{\theta}) & (\mathbf{V}_0 + \mathbf{V}_1) \otimes \boldsymbol{\rho}_{new,\mathbf{Y}}(\boldsymbol{\theta}) \\ (\mathbf{V}_0 + \mathbf{V}_1) \otimes \boldsymbol{\rho}_{\mathbf{Y},new}(\boldsymbol{\theta}) & (\mathbf{V}_0 + \mathbf{V}_1) \otimes \boldsymbol{\rho}_{\mathbf{Y},\mathbf{Y}}(\boldsymbol{\theta}) \end{pmatrix}. \tag{4.16}$$

The conditional distribution of the normal distribution is used to find $\mathbf{Y}_{new}|\mathbf{Y}$ which is

$$\mathbf{Y}_{new}|\mathbf{Y}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \mathbf{V}_0, \mathbf{V}_1, \boldsymbol{\theta} \sim N\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right), \tag{4.17}$$

where $\boldsymbol{\mu}_1 = \boldsymbol{\alpha} \otimes \mathbf{1}_{n_{new}} + \boldsymbol{\Gamma}_1 \boldsymbol{\eta}(\mathbf{I}_p \otimes \mathbf{X}_{new})$ and $\boldsymbol{\mu}_2 = \boldsymbol{\alpha} \otimes \mathbf{1}_n + \boldsymbol{\Gamma}_1 \boldsymbol{\eta}(\mathbf{I}_p \otimes \mathbf{X})$. Using (4.17), we can make prediction for an unsampled location.

4.4 Simulation

To evaluate the finite sample performance of the proposed spatial envelope and compare it with the traditional ordinary least squares multivariate regression (MLR), linear coregionalization model (LCM; Zhang, 2007), and envelope (Cook *et al.*, 2010).

The data $\{(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)\}$ were generated from the model

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.18}$$

where $\mathbf{Y}_i \in \mathbb{R}^5$, $\mathbf{X}_i \in \mathbb{R}^6$, and the structural dimension of two i.e. $u = 2$. The matrix $(\boldsymbol{\Gamma}_1; \boldsymbol{\Gamma}_0)$ is obtained by orthogonalizing an 5×5 matrix of random uniform $(0, 1)$ variables, and the elements in $\boldsymbol{\eta}$ were sampled from a standard normal population. We generated $\boldsymbol{\Sigma}_Y = \left(\boldsymbol{\Gamma}_1\boldsymbol{\Omega}_1\boldsymbol{\Gamma}_1^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\right) \otimes \boldsymbol{\rho}(\boldsymbol{\theta})$ where $\boldsymbol{\Omega}_1 = [\omega_{i,j}^1]$ where $\omega_{i,j}^1 = (-0.9)^{|i-j|}$ and $\boldsymbol{\Omega}_0 = [\omega_{i,j}^0]$ where $\omega_{ij}^0 = (-0.5)^{|i-j|}$. Three error distributions of ϵ were investigated:

1. $N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \left(\boldsymbol{\Gamma}_1\boldsymbol{\Omega}_1\boldsymbol{\Gamma}_1^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\right)$. This density serves as a benchmark where the error are independent from each other;

2. $\boldsymbol{\epsilon}$ follows a Matern covariance function with $\theta_1 = 0.5$ and $\theta_2 = 1$; This case represents a spatial correlation in the data with a small range of dependency. We call this case as an example of weak spatial correlation.

3. $\boldsymbol{\epsilon}$ follows a Matern covariance function with $\theta_1 = 0.5$ and $\theta_2 = 5$; This case represents a spatial correlation in the data with a large range of dependency. We call this case as an example of strong spatial correlation.

Sample size was chosen as 100, 225, and 400. There are two different ways that we took these samples. One is based on 10×10, 15×15 and 20×20 evenly spaced grids on $[0, 1]^2$, respectively. The second method that that we took the sample was as follows. First we made a 101×101 grid on $[0, 1]^2$, and then we chose $n = 100,\ 225$ and 400 locations. All results reported here were based on 200 replications from the simulation model in each case. In order to compare the different estimators, we used *Leave One Out Cross Validation* (LOCV) method which provides a convenient approximation for the prediction error under squared-error loss given by

$$MSPE = \frac{\sum_{i=1}^{n}(\hat{Y}^{(-i)}(s_i) - Y(s_{i,obs}))^2}{n}, \tag{4.19}$$

where $Y(s_{i,obs})$ is the observed value for response in location s and $\hat{Y}^{(-i)}(s_i)$ is the predicted values of $Y(s_i)$ computed with the ith row of the data removed. Tables 4.1 and 4.2 summarize the results of these simulations. These tables provide the LOCV for different methods and different errors and smaller LOCV shows better performance.

Table 4.1: *Prediction accuracy comparison based on the mean (standard deviation) of leave one out cross validation (LOCV) for all 200 data sets for equally spaced samples. Smaller LOCV shows better estimation.*

ϵ	n	MLR	LCM	Envelope	Spatial Envelope
1	100	19.02 (1.537)	20.01 (1.754)	13.71 (1.547)	14.28 (1.644)
	225	18.49 (1.153)	19.75 (1.659)	11.49 (1.124)	12.51 (1.234)
	400	18.27 (0.828)	19.02 (1.002)	10.37 (0.812)	10.87 (0.989)
2	100	102.79 (35.570)	22.54 (3.246)	91.98 (36.379)	20.21 (1.988)
	225	101.57 (32.495)	20.46 (2.897)	89.24 (33.083)	18.34 (1.450)
	400	99.98 (32.185)	18. 89 (2.051)	88.95 (31.855)	17.68 (1.056)
3	100	117.79 (48.834)	24.19 (4.125)	119.08 (47.852)	21.36 (2.353)
	225	103.22 (39.065)	21.78 (3.278)	104.73 (39.023)	20.76 (2.012)
	400	99.08 (37.718)	19.45 (3.001)	100.39 (36.896)	18.10 (1.651)

Table 4.2: *Prediction accuracy comparison based on the mean (standard deviation) of leave one out cross validation (LOCV) for all 200 data sets for random location samples. Smaller LOCV shows better estimation.*

ϵ	n	MLR	LCM	Envelope	Spatial Envelope
1	100	20.12 (1.613)	21.01 (1.863)	14.32 (1.699)	14.98 (1.722)
	225	19.34 (1.231)	19.68 (1.542)	13.12 (1.234)	13.19 (1.201)
	400	17.83 (0.804)	18.22 (1.101)	11.73 (0.718)	12.37 (0.819)
2	100	104.02 (36.702)	23.32 (4.111)	93.02 (30.433)	19.21 (2.004)
	225	102.41 (34.521)	21.41 (3.758)	91.34 (27.211)	17.34 (1.352)
	400	100.39 (30.822)	19.20 (3.201)	89.21 (25.581)	16.68 (1.110)
3	100	116.34 (45.089)	25.21 (4.821)	97.01 (43.021)	20.79 (2.115)
	225	108.15 (34.211)	22.35 (3.555)	95.52 (31.774)	18.92 (1.944)
	400	101.54 (32.102)	20.44 (2.998)	90.94 (30.234)	17.03 (1.234)

From the summary of all three different error distributions, in both scenarios, it can be seen that for the standard normal errors, where the data are actually independent from

each other, the spatial envelope provides a comparable result to the envelope method and both of these methods provide better results compared to MLR and LCM. Furthermore, in error distributions 2 and 3, where there exists spatial dependency in the data, the spatial envelope method performed almost equally as well as they did in the cases without spatial dependency. The spatial envelope method performs drastically better than the original envelope. In addition, spatial envelope outperformed LCM in both of the cases that there exists spatial dependency in the data. Therefore, we can conclude that the proposed spatial envelope model provided consistent estimates with good prediction accuracy in all error distributions considered.

4.5 Real data

Air pollution is the existence of one or several pollutant elements such as dust, gases, smoke, etc. in the air that has a serious impact on the health of plants and animals (including humans). There is evidence that shows exposure to air pollutions such as particulate matter (PM) and nitrogen dioxide has significant effect on human health. For instance PM is associated with increases in cardiopulmonary disease (Pope *et al.*, 2002; Pope *et al.*, 2004). Furthermore, nitrogen dioxide increases allergic responses to inhaled pollens, risk of respiratory symptoms such as acute bronchitis and cough and phlegm, particularly in children, and decreases lung function (World Health Organization, 2003). Most of the air pollutant are study concentrate on one of the pollutant in the air but since a relation among these pollutant seems to exist, it would be interesting to study the behavior of all of these pollutants together. Here, we are going to apply the proposed methodology to the air pollution data in the northeastern United States. This dataset has drawn much attention from statis-

ticians and other scientists. These researchers looked at this data from different points of view including but not restricted to climate change (Phelan *et al.*, 2016), health science (Kioumourtzoglou *et al.*, 2016; Zeng *et al.*, 2016), and air quality (Battye *et al.*, 2016).

The pollutant and weather data that we used in this study are the average levels of the following variables in January 2015. In this study, the response variables are:

1. Criteria Gases: ozone, sulfur dioxide (SO_2), carbon monoxide (CO), and nitrogen dioxide (NO_2).

2. Particulates: particulate matter which are PM2.5 FRM/FEM Mass, PM2.5 non FRM/FEM Mass, PM10 Mass, and PM2.5 speciation. PM10 includes particles less than or equal to 10 micrometers in diameter. Similarly, PM2.5 includes particles less than or equal to 2.5 micrometers and is also called fine particle pollution.

3. Toxics: core Hazardous Air Pollutants (HAPS) and Volatile Organic Compounds (VOCs). Hazardous air pollutants (HAPs) (also called toxic air pollutants or air toxics) are pollutants that are known or suspected to cause serious health problems such as cancer.

This data is combined with the following meteorological variables: wind, temperature, and relative humidity as our regressors. Along with this information, latitude and longitude of the monitoring locations were used to model the spatial structure in the data. Our study area consists of 9 states in the northeast of the United States of America which are: Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont. This dataset can be found at http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata. Figure 4.1 shows the study area in red.

Figure 4.1: *Study area in the United States of America. States of interest are shaded in red.*

The study area contains 270 sites which measure the air quality data. Figure 4.2 shows the location of these sites on the map. The results of the cross-validation showed that best choice for the dimension is 3. The Matern's covariance parameter for this data set are estimated to be 0.51 and 0.92 for θ_1 and θ_2 respectively. The corresponding direction estimates using the spatial envelope are shown in Table 4.3. In addition, Table 4.4 shows the regression coefficients and their asymptotic standard deviation (in parenthesis) using spatial envelope.

Table 4.3: *The corresponding direction estimates using spatial envelope for the air pollution data in northeastern United States of America.*

Variable	Direction 1	Direction 2	Direction 3
Ozone	-0.0001	0.0009	0.0003
Carbon monoxide	-0.0006	-0.0041	-0.0011
Nitrogen dioxide	0.0643	-0.9615	-0.1195
Sulphur dioxide	0.0110	0.0076	-0.0116
PM10 Mass	-0.9754	-0.0319	-0.1164
PM2.5 FRM/FEM Mass	-0.1304	-0.1732	-0.1811
PM2.5 FRM/FEM non Mass	0.0992	-0.1350	-0.0358
PM2.5 Speciation	0.1302	0.1601	-0.9684
Hazardous Air Pollutants	-0.0014	-0.0018	0.0006
Volatile Organic Compounds	0.0238	0.0247	0.0125

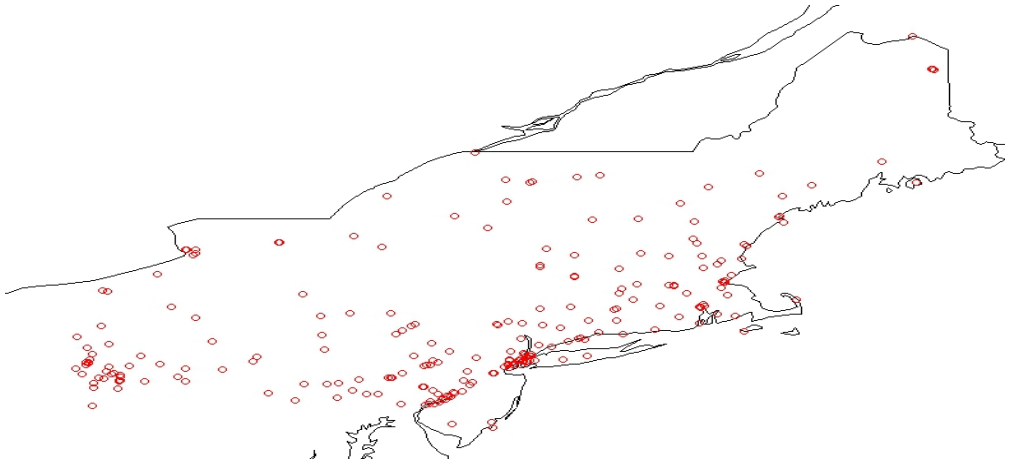

Figure 4.2: *Location of different sites in the study area. It can be seen that there is a higher number of sites in places with larger population compare to other palaces in the study area.*

Table 4.4: *Regression coefficients (asymptotic standard deviation) using spatial envelope the air pollution data in northeastern United States of America.*

Variable	Wind	Temperature	Relative humidity
Ozone	-0.0008 (0.010)	0.0015 (0.012)	0.0007 (0.010)
Carbon monoxide	0.0066 (0.061)	-0.0128 (0.040)	-0.0079 (0.030)
Nitrogen dioxide	0.8911 (0.435)	-1.8826 (0.382)	-0.9927 (0.317)
Sulphur dioxide	-0.1372 (0.147)	0.2748 (0.101)	0.2278 (0.078)
PM10 Mass	3.298 (0.904)	-5.9855 (0.975)	-5.1187 (0.769)
PM2.5 FRM/FEM Mass	0.2905 (0.473)	-0.4577 (0.442)	-0.2823 (0.349)
PM2.5 FRM/FEM non Mass	-0.1804 (0.243)	0.3048 (0.166)	0.3617 (0.077)
PM2.5 Speciation	-0.0063 (0.016)	0.011 (0.006)	0.0105 (0.004)
Hazardous Air Pollutants	0.0094 (0.014)	-0.0187 (0.007)	-0.0141 (0.005)
Volatile Organic Compounds	0.0157 (0.102)	-0.0439 (0.069)	-0.041 (0.053)

By checking the estimated coefficients (directions), we can see PM10 mass, PM2.5 mass, and PM 2.5 speciation are important in the first direction. As it can be seen, this direction mainly involves with particulates. Among these particulates, PM10 mass has the largest impact in this direction. In the second direction, nitrogen dioxide, PM2.5 FRM/FEM mass, PM2.5 FRM/FEM non mass, and PM2.5 speciation are important. In this direction nitrogen dioxide has the largest effect. In the third direction, nitrogen dioxide, PM2.5 FRM/FEM mass, and PM2.5 speciation are important. In the third direction, PM 2.5 speciation has the largest effect.

Using fossil fuels creates nitrogen monoxide and nitrogen dioxide. The nitrogen monoxide will also become nitrogen dioxide in the atmosphere. Almost 80 percent of nitrogen dioxide in the urban areas is because of motor vehicles. The rest comes from the oil industry, metals industry, and plants that use fossil fuels such as coal power plants. The amount of the released nitrogen dioxide from one ton of fossil fuel is 36 kilograms (Laegreid, 1999). Therefore, because there exists a lot of industry and crowded centers in the northeastern of the United States the amount of this pollutant is high. On the other hand, in January, due to the extremely low temperature in the the United States there is a high use of the fossil fuels for warming the houses and buildings which will increase the amount of the nitrogen dioxide. This explains the importance of the nitrogen dioxide that we found in this study.

Temperature decreases in the troposphere with increasing altitude. This Phenomena is called lapse rate. In cases where the atmospheric layer near the earth surface loses heat energy due to the night radiation resulting from the contact of atmosphere with warm ground, the track of temperature decreases with height adjustment and the temperature rarely rises. This phenomena is called inversion and it happens in cold season which leads to the increasing of the pollutant close to earth surface. There the inversion is a common phenomena

in the winter in different continents (Byers, 1959). Therefore in the northeastern United States which has cold winter, there exist more favorable conditions for the occurrence of the inversion. The inversion layer is very close to earth surface such that it will vanish in 1 kilometer altitude on the oceans and 2 kilometers altitude on the continents. This layer does not let the particulates go higher into the atmosphere and it keeps them close to the earth surface which will support the founding of our statistical analysis.

In general, we find out that the most important pollutants in January are particulates and nitrogen and other pollutants have small effect. These statistical conclusions support the chemical claim that the reaction among gases is affected by chemical reactions between weather conditions and pollutants in the presence of sunlight and heat which are not available in January in the northeastern of the USA.

4.6 Theoretical results and prediction maps

4.6.1 Derivation of the factorization of the likelihood function in section 4.1

The likelihood function of the model (4.3) will be as follows:

$$
\begin{aligned}
&L^{(u)}(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{V}_0,\mathbf{V}_1,\boldsymbol{\theta}) = [\det((\mathbf{V}_0+\mathbf{V}_1)\otimes\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{1}{2}} \\
&\times exp\left\{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\alpha}\otimes\mathbf{1}_n-(\mathbf{I}_p\otimes\mathbf{X})\boldsymbol{\beta})^T\left((\mathbf{V}_0+\mathbf{V}_1)\otimes\boldsymbol{\rho}(\boldsymbol{\theta})\right)^{-1}(\mathbf{Y}-\boldsymbol{\alpha}\otimes\mathbf{1}_n-(\mathbf{I}_p\otimes\mathbf{X})\boldsymbol{\beta})\right\} \\
&= [\det(\mathbf{V}_0\otimes\boldsymbol{\rho}(\boldsymbol{\theta})+\mathbf{V}_1\otimes\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{1}{2}} \\
&\times exp\left\{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\alpha}\otimes\mathbf{1}_n-(\mathbf{I}_p\otimes\mathbf{X})\boldsymbol{\beta})^T((\mathbf{V}_0+\mathbf{V}_1)^{-1}\otimes\boldsymbol{\rho}(\boldsymbol{\theta})^{-1})(\mathbf{Y}-\boldsymbol{\alpha}\otimes\mathbf{1}_n-(\mathbf{I}_p\otimes\mathbf{X})\boldsymbol{\beta})\right\} \\
&= [\det(\mathbf{V}_0\otimes\boldsymbol{\rho}(\boldsymbol{\theta})+\mathbf{V}_1\otimes\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{1}{2}} \\
&\times exp\left\{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\alpha}\otimes\mathbf{1}_n-(\mathbf{I}_p\otimes\mathbf{X})\boldsymbol{\beta})^T\left((\mathbf{V}_0^{\dagger}\otimes\boldsymbol{\rho}(\boldsymbol{\theta})^{-1})+(\mathbf{V}_1^{\dagger}\otimes\boldsymbol{\rho}(\boldsymbol{\theta})^{-1})\right)(\mathbf{Y}-\boldsymbol{\alpha}\otimes\mathbf{1}_n-(\mathbf{I}_p\otimes\mathbf{X})\boldsymbol{\beta})\right\},
\end{aligned}
\tag{4.20}
$$

where † denotes Moore-Penrose inverse and $\mathbf{V}_0 = \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0$ and $\mathbf{V}_1 = \boldsymbol{\Gamma}_1\boldsymbol{\Omega}_1\boldsymbol{\Gamma}_1$. Since $span(\boldsymbol{\beta}) \subseteq span(\mathbf{V}_1)$, therefore $\mathbf{V}_0\boldsymbol{\beta} = 0$ and because $\mathbf{V} = \mathbf{V}_0 + \mathbf{V}_1$, this likelihood can be factored as:

$$
\begin{aligned}
&L^{(u)}(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{V}_0,\mathbf{V}_1,\boldsymbol{\theta}) = [\det(\mathbf{V}_0+\mathbf{V}_1)]^{-\frac{n}{2}}[\det(\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{p}{2}} \\
&\times exp\left\{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\alpha}\otimes\mathbf{1}_n-(\mathbf{I}_p\otimes\mathbf{X})\boldsymbol{\beta}))^T\left(\mathbf{V}_1^{\dagger}\otimes\boldsymbol{\rho}(\boldsymbol{\theta})^{-1}\right)(\mathbf{Y}-\boldsymbol{\alpha}\otimes\mathbf{1}_n-(\mathbf{I}_p\otimes\mathbf{X})\boldsymbol{\beta}))\right\} \\
&\times exp\left\{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\alpha}\otimes\mathbf{1}_n)^T\left(\mathbf{V}_0^{\dagger}\otimes\boldsymbol{\rho}(\boldsymbol{\theta})^{-1}\right)(\mathbf{Y}-\boldsymbol{\alpha}\otimes\mathbf{1}_n)\right\} \\
&= L_1^{(u)}(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{V}_1,\boldsymbol{\theta})\times L_2^{(u)}(\boldsymbol{\alpha},\mathbf{V}_0,\boldsymbol{\theta}),
\end{aligned}
\tag{4.21}
$$

where

$$
\begin{aligned}
L_1^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{V}_1, \boldsymbol{\theta}) &= [det_0(\mathbf{V}_1)]^{-\frac{n}{2}} [\det(\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{p}{2}} \\
&\quad \times exp\left\{ -\frac{1}{2} (\mathbf{Y} - \boldsymbol{\alpha} \otimes \mathbf{1}_n - (\mathbf{I}_p \otimes \mathbf{X})\boldsymbol{\beta}))^T \left(\mathbf{V}_1^{\dagger} \otimes \boldsymbol{\rho}(\boldsymbol{\theta})^{-1} \right) (\mathbf{Y} - \boldsymbol{\alpha} \otimes \mathbf{1}_n - (\mathbf{I}_p \otimes \mathbf{X})\boldsymbol{\beta})) \right\}, \\
L_2^{(u)}(\boldsymbol{\alpha}, \mathbf{V}_0, \boldsymbol{\theta}) &= [det_0(\mathbf{V}_0)]^{-\frac{n}{2}} [\det(\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{p}{2}} \\
&\quad \times exp\left\{ -\frac{1}{2} (\mathbf{Y} - \boldsymbol{\alpha} \otimes \mathbf{1}_n)^T \left(\mathbf{V}_0^{\dagger} \otimes \boldsymbol{\rho}(\boldsymbol{\theta})^{-1} \right) (\mathbf{Y} - \boldsymbol{\alpha} \otimes \mathbf{1}_n) \right\},
\end{aligned}
\tag{4.22}
$$

where $det_0(\mathbf{A})$ denotes the product of non-zero eigenvalues of $\mathbf{A}$ where $\mathbf{A}$ is a non-zero symmetric matrix.

4.7 Coordinate free version of the algorithm of the spatial envelope

The objective is to maximize the likelihood in (4.21) over $\boldsymbol{\beta}, \mathbf{V}_0, \mathbf{V}_1$, and $\boldsymbol{\theta}$ subject to the constraints:

$$
\begin{aligned}
&span(\boldsymbol{\beta}) \subseteq span(\mathbf{V}_1), \quad (a) \\
&\mathbf{V}_0 \mathbf{V}_1 = 0, \qquad\qquad\quad (b).
\end{aligned}
\tag{4.23}
$$

Based on this factorization given in equation (4.21), we can decompose the likelihood maximization into the following steps:

1. Fix $\boldsymbol{\beta}, \mathbf{V}_0$, $\mathbf{V}_1$, and $\boldsymbol{\theta}$, and maximize $L^{(u)}$ in (4.4) over $\boldsymbol{\alpha}$ which will be:

$$\hat{\boldsymbol{\alpha}} = \bar{\mathbf{Y}} - \bar{\mathbf{X}}\boldsymbol{\beta}.$$

Let $\mathbf{H} = \mathbf{Y} - \bar{\mathbf{Y}} \otimes \mathbf{1}_n$, $\mathbf{U} = vec(\mathbf{H})$, $\mathbf{G} = \mathbf{X} - \bar{\mathbf{X}} \otimes \mathbf{1}_n$, and $\mathbf{F} = \mathbf{I} \otimes \mathbf{G}$. Therefore, the profile likelihood can be written as the following:

$$\begin{aligned} L_1^{(u,p)}(\boldsymbol{\beta}, \mathbf{V}_1, \boldsymbol{\theta}) = [det_0(\mathbf{V}_1)]^{-\frac{n}{2}} [\det(\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{p}{2}} \\ \times exp\left\{ -\frac{1}{2} (\mathbf{U} - \mathbf{F}\boldsymbol{\beta})^T \left(\mathbf{V}_1^\dagger \otimes \boldsymbol{\rho}(\boldsymbol{\theta})^{-1} \right) (\mathbf{U} - \mathbf{F}\boldsymbol{\beta}) \right\}, \end{aligned} \tag{4.24}$$

and

$$L_2^{(u,p)}(\boldsymbol{\alpha}, \mathbf{V}_0, \boldsymbol{\theta}) = [det_0(\mathbf{V}_0)]^{-\frac{n}{2}} exp\left\{ -\frac{1}{2} \mathbf{U}^T \left(\mathbf{V}_0^\dagger \otimes \boldsymbol{\rho}(\boldsymbol{\theta})^{-1} \right) \mathbf{U} \right\}. \tag{4.25}$$

2. Fix $\mathbf{V}_1$, and $\boldsymbol{\theta}$ and maximize the function $L_1^{(u,p)}$ over $\boldsymbol{\beta}$, subject to (4.23a), to obtain $L_{21}^{(u,p)}(\mathbf{V}_1, \boldsymbol{\theta})$. Since $vec(\mathbf{X}\boldsymbol{\beta}) = (\mathbf{I} \otimes \mathbf{X}) vec(\boldsymbol{\beta})$ and

$$tr(\mathbf{D}^T(\mathbf{C}^T\mathbf{B}^T\mathbf{A}^T)) = (vec(\mathbf{D}))^T(\mathbf{A} \otimes \mathbf{C}^T)(vec(\mathbf{B}))^T,$$

we have

$$\begin{aligned} (\mathbf{U} - \mathbf{F}\boldsymbol{\beta})^T \left(\mathbf{V}_1^\dagger \otimes \boldsymbol{\rho}(\boldsymbol{\theta})^{-1} \right) (\mathbf{U} - \mathbf{F}\boldsymbol{\beta}) &= tr\left((\mathbf{H} - \mathbf{G}\boldsymbol{\beta})^T \boldsymbol{\rho}(\boldsymbol{\theta})^{-1} (\mathbf{H} - \mathbf{G}\boldsymbol{\beta}) \mathbf{V}_1^\dagger \right) \\ &= tr\left((\mathbf{H} - \mathbf{G}\boldsymbol{\beta})^T \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} (\mathbf{H} - \mathbf{G}\boldsymbol{\beta}) \mathbf{V}_1^\dagger \right) \\ &= tr\left(\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} (\mathbf{H} - \mathbf{G}\boldsymbol{\beta}) \mathbf{V}_1^\dagger (\mathbf{H} - \mathbf{G}\boldsymbol{\beta})^T \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \right) \\ &= tr\left(\left(\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{H} - \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{G}\boldsymbol{\beta} \right) \mathbf{V}_1^\dagger \left(\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{H} - \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{G}\boldsymbol{\beta} \right)^T \right) \\ &= tr\left(\left(\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{H} - \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{G}\boldsymbol{\beta}\mathbf{I}_p \right) \mathbf{V}_1^\dagger \left(\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{H} - (\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{G}\boldsymbol{\beta}\mathbf{I}_p \right)^T \right) \end{aligned} \tag{4.26}$$

where $tr(\cdot)$ denotes the trace of the matrix. The last equality in equation (4.26) is

from Lemma 4.1 in Cook *et al.*, (2010). Thus, the optimal $\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}\boldsymbol{\beta}\mathbf{I}_p$ is

$$\mathbf{P}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}}\left(\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H}\right)\mathbf{P}^T_{\mathbf{I}_p(\mathbf{V}_1^\dagger)} = \mathbf{P}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}}\left(\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H}\right)\mathbf{P}_{\mathbf{V}_1},$$

where $\mathbf{P}_{(.)}$ is the projection onto the subspace indicated by its argument. This implies following

$$\boldsymbol{\beta}^T = \left(\mathbf{G}^T\boldsymbol{\rho}(\boldsymbol{\theta})\mathbf{G}\right)^{-1}\mathbf{G}\boldsymbol{\rho}(\boldsymbol{\theta})\mathbf{H}\mathbf{P}_{\mathbf{V}_1} \Rightarrow \boldsymbol{\beta} = \mathbf{P}_{\mathbf{V}_1}\hat{\boldsymbol{\beta}},$$

where $\boldsymbol{\beta}$ is the MLE estimate of $\boldsymbol{\beta}$ from the full model (4.1). Substituting this into (4.25), and using the relation $\mathbf{P}_{\mathbf{V}_1}\mathbf{V}_1^\dagger = \mathbf{V}_1^\dagger$, we see that the maximum of $L_2^{(u,p)}$ for fixed $\mathbf{V}_1$ over $\boldsymbol{\beta}$ is

$$\begin{aligned}
&L_{11}^{(u,p)}(\mathbf{V}_1,\boldsymbol{\theta}) = [det_0(\mathbf{V}_1)]^{-\frac{n}{2}}[\det(\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{p}{2}}\\
&\times exp\left\{-\frac{1}{2}tr\left(\left(\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H} - \mathbf{P}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}}\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H}\mathbf{P}_{\mathbf{V}_1}\right)\mathbf{V}_1^\dagger\left(\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H} - \mathbf{P}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}}\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H}\mathbf{P}_{\mathbf{V}_1}\right)^T\right)\right\}\\
&= [det_0(\mathbf{V}_1)]^{-\frac{n}{2}}[\det(\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{p}{2}}\\
&\times exp\left\{-\frac{1}{2}tr\left(\left(\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H} - \mathbf{P}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}}\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H}\right)\mathbf{V}_1^\dagger\left(\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H} - \mathbf{P}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}}\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H}\right)^T\right)\right\}\\
&= [det_0(\mathbf{V}_1)]^{-\frac{n}{2}}[\det(\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{p}{2}}exp\left\{-\frac{1}{2}tr\left(\left(\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}}\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H}\right)\mathbf{V}_1^\dagger\left(\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}}\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H}\right)^T\right)\right\}
\end{aligned} \tag{4.27}$$

where $\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}} = \mathbf{I}_n - \mathbf{P}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}}$.

3. Maximize $L^{(u)}(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{V}_0,\mathbf{V}_1,\boldsymbol{\theta})$ over all $\mathbf{V}_0$, $\mathbf{V}_1$, and $\boldsymbol{\theta}$. Since $L^{(u)}(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{V}_0,\mathbf{V}_1,\boldsymbol{\theta}) =$

$L_1^{(u,p)}(\boldsymbol{\beta}, \mathbf{V}_1, \boldsymbol{\theta}) \times L_2^{(u,p)}(\boldsymbol{\alpha}, \mathbf{V}_0, \boldsymbol{\theta})$, we have

$$\begin{aligned}
L^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{V}_0, \mathbf{V}_1, \boldsymbol{\theta}) &= [det_0(\mathbf{V}_0)]^{-\frac{n}{2}} [det_0(\mathbf{V}_1)]^{-\frac{n}{2}} [\det(\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{p}{2}} \\
&\times exp\left\{-\frac{1}{2} tr\left(\left(\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}} \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{H}\right) \mathbf{V}_1^{\dagger} \left(\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}} \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{H}\right)^T\right)\right\} \\
&\times exp\left\{-\frac{1}{2} \mathbf{U}^T \left(\mathbf{V}_0^{\dagger} \otimes \boldsymbol{\rho}(\boldsymbol{\theta})^{-1}\right) \mathbf{U}\right\} \\
&= [det_0(\mathbf{V}_0)]^{-\frac{n}{2}} [det_0(\mathbf{V}_1)]^{-\frac{n}{2}} [\det(\boldsymbol{\rho}(\boldsymbol{\theta}))]^{-\frac{p}{2}} \\
&\times exp\left\{-\frac{1}{2} tr\left(\left(\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}} \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{H}\right) \mathbf{V}_1^{\dagger} \left(\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}} \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{H}\right)^T\right)\right\} \\
&\times exp\left\{-\frac{1}{2} tr\left(\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{H} \mathbf{V}_0^{\dagger} \mathbf{H}^T \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\right)\right\}.
\end{aligned} \tag{4.28}$$

This maximization can be as follows:

(a) Fix $\mathbf{V}_0$ and $\mathbf{V}_1$ and maximize $L^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{V}_0, \mathbf{V}_1, \boldsymbol{\theta})$ over $\boldsymbol{\theta}$ by solving the following maximization problem:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{-\frac{p}{2} \det(\boldsymbol{\rho}(\boldsymbol{\theta})) - \frac{1}{2} tr\left(\left(\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}} \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{H}\right) \mathbf{V}_1^{\dagger} \left(\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}} \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{H}\right)^T + \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{H} \mathbf{V}_0^{\dagger} \mathbf{H}^T \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\right)\right\} \tag{4.29}$$

(b) Fix the $\boldsymbol{\theta}$ and maximize $L^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{V}_0, \mathbf{V}_1, \boldsymbol{\theta})$ over $\mathbf{V}_0$ and $\mathbf{V}_1$. This means maximize $L_{11}^{(u,p)}(\boldsymbol{\beta}, \mathbf{V}_1, \boldsymbol{\theta})$ over $\mathbf{V}_1$ and $L_{12}^{(u,p)}(\boldsymbol{\alpha}, \mathbf{V}_0, \boldsymbol{\theta})$ over $\mathbf{V}_0$. Maximization $L_{11}^{(u,p)}(\mathbf{P}_{\mathbf{V}_1})$ over $\mathbf{V}_1$ is

$$L_{11}^{(u,p)}(\mathbf{P}_{\mathbf{V}_1}) \propto \left[det_0\left(\mathbf{P}_{\mathbf{V}_1}\left(\mathbf{H}^T \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}} \boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}} \mathbf{H}\right) \mathbf{P}_{\mathbf{V}_1}\right)\right]^{-\frac{n}{2}} \tag{4.30}$$

and maximization $L_{12}^{(u,p)}(\mathbf{P}_{\mathbf{V}_0})$ over $\mathbf{V}_0$ is

$$L_{12}^{(u,p)}(\mathbf{P}_{\mathbf{V}_0}) \propto \left[det_0\left(\mathbf{P}_{\mathbf{V}_0} \mathbf{H}^T \boldsymbol{\rho}(\boldsymbol{\theta})^{-1} \mathbf{H} \mathbf{P}_{\mathbf{V}_0}\right)\right]^{-\frac{n}{2}}. \tag{4.31}$$

Therefore, maximization $L^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{V}_0, \mathbf{V}_1, \boldsymbol{\theta})$ over $\mathbf{V}_0$ and $\mathbf{V}_1$ is equivalent to maximization of $L_{11}^{(u,p)}(\mathbf{P}_{\mathbf{V}_1}) \times L_{12}^{(u,p)}(\mathbf{P}_{\mathbf{V}_0})$ which is proportion to

$$\begin{aligned}
\mathbf{D} &= \left[det_0\left(\mathbf{P}_{\mathbf{V}_1}\left(\mathbf{H}^T\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}}\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H}\right)\mathbf{P}_{\mathbf{V}_1}\right)\right]^{-\frac{n}{2}} \times \left[det_0\left(\mathbf{P}_{\mathbf{V}_0}\mathbf{H}^T\boldsymbol{\rho}(\boldsymbol{\theta})^{-1}\mathbf{H}\mathbf{P}_{\mathbf{V}_0}\right)\right]^{-\frac{n}{2}} \\
&= \left[det_0\left(\mathbf{P}_{\mathbf{V}_1}\left(\mathbf{H}^T\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}}\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H}\right)\mathbf{P}_{\mathbf{V}_1} + \mathbf{P}_{\mathbf{V}_0}\mathbf{H}^T\boldsymbol{\rho}(\boldsymbol{\theta})^{-1}\mathbf{H}\mathbf{P}_{\mathbf{V}_0}\right)\right]^{-\frac{n}{2}} \\
&= \left[det_0\left(\mathbf{P}_{\mathbf{V}_1}\left(\mathbf{H}^T\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}}\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H}\right)\mathbf{P}_{\mathbf{V}_1} + \mathbf{Q}_{\mathbf{V}_0}\mathbf{H}^T\boldsymbol{\rho}(\boldsymbol{\theta})^{-1}\mathbf{H}\mathbf{Q}_{\mathbf{V}_0}\right)\right]^{-\frac{n}{2}}
\end{aligned} \tag{4.32}$$

where $\mathbf{Q}_{\mathbf{V}_0} = \mathbf{I}_p - \mathbf{P}_{\mathbf{V}_1}$. Since $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}} = \mathbf{H}^T\boldsymbol{\rho}(\boldsymbol{\theta})\mathbf{H}$ and

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{res} &= \mathbf{H}^T\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{Q}_{\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{G}}\boldsymbol{\rho}(\boldsymbol{\theta})^{-\frac{1}{2}}\mathbf{H} \\
&= \mathbf{H}^T\boldsymbol{\rho}(\boldsymbol{\theta})\mathbf{H} - \mathbf{H}^T\boldsymbol{\rho}(\boldsymbol{\theta})\mathbf{G}\left(\mathbf{G}^T\boldsymbol{\rho}^{-1}(\boldsymbol{\theta})\mathbf{G}\right)^{-1}\mathbf{G}^T\boldsymbol{\rho}(\boldsymbol{\theta})\mathbf{H}.
\end{aligned} \tag{4.33}$$

Therefore we have $\mathbf{D} = det(\mathbf{P}_{\mathbf{V}_1}\hat{\boldsymbol{\Sigma}}_{\mathbf{res}}\mathbf{P}_{\mathbf{V}_1} + \mathbf{Q}_{\mathbf{V}_1}\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}}\mathbf{Q}_{\mathbf{V}_1})$ and $\hat{\mathbf{V}}_1 = \underset{\mathbf{V}_1}{\text{argmax}}(\mathbf{D})$ and $\mathbf{P}_{\hat{\mathbf{V}}_0} = \mathbf{I} - \mathbf{P}_{\hat{\mathbf{V}}_1}$

Repeat (a) and (b) until it converges.

4.7.1 Proof of Theorem 1

In this section, we derive the an explicit expression for Ψ as given by (4.37). In order to find these expression, we need to find expressions for the eight partial derivatives $\frac{\partial \Psi_i}{\partial \phi_j^T}$ for $i = 1, 2$ and $j = 1, 2, 3, 4$. Before starting the derivation, the following properties hold:

1. Suppose $\mathbf{A}$ and $\mathbf{X}$ are both $n \times n$, and $\mathbf{X}$ is symmetric, then

$$\begin{aligned} vech\left(\mathbf{A}\mathbf{X}\mathbf{A}^T\right) &= \mathbf{M}_n vec(\mathbf{A}\mathbf{X}\mathbf{A}) \\ &= \mathbf{M}_n(\mathbf{A}\otimes\mathbf{A})vec(\mathbf{X}) \\ &= \mathbf{M}_n(\mathbf{A}\otimes\mathbf{A})\mathbf{E}_n vech(\mathbf{X}) \\ &= \mathbf{C}vech(\mathbf{X}), \end{aligned}$$

   where for a given matrix such as $\mathbf{A}$, $vech(\mathbf{A}) = \mathbf{M}_n vec(\mathbf{A})$, $vec(\mathbf{A}) = \mathbf{E}_n vech(\mathbf{A})$, and $\mathbf{C} = \mathbf{M}_n(\mathbf{A}\otimes\mathbf{A})\mathbf{E}_n$, a $k \times k$ matrix where $k = \frac{n(n+1)}{2}$.

2. If $\mathbf{Y} = \mathbf{A}\mathbf{X}\mathbf{B}$, then $vec(\mathbf{Y}) = \left(\mathbf{B}^T \otimes \mathbf{A}\right) vec(\mathbf{X})$ and

$$\frac{\partial \mathbf{Y}}{\partial \left(vec(\mathbf{X})\right)^T} = \mathbf{B}^T \otimes \mathbf{A}$$

3. If $\mathbf{Y} = \mathbf{A}\mathbf{X}^T\mathbf{B}$, and $\mathbf{X}$ is $m \times n$ then $vec(\mathbf{Y}) = (\mathbf{B}^T \otimes \mathbf{A})vec\left(\mathbf{X}^T\right)$ where $vec(\mathbf{X}^T) = \mathbf{I}_{(m,n)}vec(\mathbf{X})$, where $\mathbf{I}_{(m,n)}$ is the $mn \times mn$ permutation matrix and

$$\frac{\partial \mathbf{Y}}{\partial \left(vec(\mathbf{X})\right)^T} = \left(\mathbf{B}^T \otimes \mathbf{A}\right)\mathbf{I}_{(m,n)}.$$

4. If $\mathbf{X}$ is $m \times n$ and $\mathbf{Y} = \mathbf{X}\mathbf{B}\mathbf{X}^T$ where $\mathbf{B}$ is symmetric then

$$\frac{\partial \mathbf{Y}}{\partial \left(vec(\mathbf{X})\right)^T} = \left(\mathbf{I}_{m^2} + \mathbf{I}_{(m,m)}\right)(\mathbf{X}\mathbf{B} \otimes \mathbf{I}_m).$$

5. Let $\mathbf{F}$ and $\mathbf{G}$ be $m \times n$ and $p \times q$ matrices, respectively, which are functions of $\mathbf{x}$, then

we have

$$\frac{\partial\left(vec(\mathbf{F})\otimes vec(\mathbf{G})\right)}{\partial vec(\mathbf{x})^T}=\left(vec(\mathbf{F})\otimes\frac{\partial vec(\mathbf{G})}{\partial vec(\mathbf{x})^T}\right)+\left(\frac{\partial vec(\mathbf{F})}{\partial vec(\mathbf{x})^T}\otimes vec(\mathbf{G})\right)$$

and

$$\frac{\partial vec\left(\mathbf{F}\otimes\mathbf{G}\right)}{\partial vec(\mathbf{x})^T}=\left(\mathbf{I}_n\otimes\mathbf{I}_{(m,q)}\otimes\mathbf{I}_p\right)\frac{\partial\left(vec(\mathbf{F})\otimes vec(\mathbf{G})\right)}{\partial vec(\mathbf{x})^T}$$

Proof of above properties can be found in Seber (2008).

Theorem 1: Suppose $\bar{\mathbf{X}}=0$ and $\mathbf{J}$ is the Fisher information for $\psi(\phi)$ in the model (4.3):

$$\begin{aligned}\mathbf{J}&=\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\boldsymbol{\Sigma}^{-1} & \mathbf{0}\\ \mathbf{0} & \frac{1}{2}\mathbf{E}_r^T\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbf{E}_r\end{bmatrix}\\&=\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{X}}\otimes\mathbf{V}^{-1}\otimes\rho(\boldsymbol{\theta})^{-1} & \mathbf{0}\\ \mathbf{0} & \frac{1}{2}\mathbf{E}_r^T\left(\mathbf{V}^{-1}\otimes\boldsymbol{\rho}(\boldsymbol{\theta})^{-1}\otimes\mathbf{V}^{-1}\otimes\rho(\boldsymbol{\theta})^{-1}\right)\mathbf{E}_r\end{bmatrix}\end{aligned}\tag{4.34}$$

where $\boldsymbol{\Sigma}_{\mathbf{X}}=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\mathbf{X}_i\mathbf{X}_i^T$, and $\mathbf{E}_r\in R^{r^2\times r(r+1)/2}$ is expansion matrix which is defined such that for a given matrix such as $\mathbf{A}$, $vec(\mathbf{A})=\mathbf{E}_r vech(\mathbf{A})$. Let $\boldsymbol{\Lambda}=\mathbf{J}^{-1}$ be the asymptotic variance of the MLE under the full model. Then

$$\sqrt{n}(\hat{\phi}-\phi)\to N(\mathbf{0},\boldsymbol{\Lambda}_0)\tag{4.35}$$

where $\boldsymbol{\Lambda}_0=\Psi(\Psi^T\boldsymbol{\Lambda}\Psi)^{\dagger}\Psi$ and Ψ is as follows:

$$\begin{bmatrix}\mathbf{I}_p\otimes\boldsymbol{\Gamma}_1 & \boldsymbol{\eta}^T\otimes\mathbf{I}_r & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \Psi_{22} & \Psi_{23} & \Psi_{24}\end{bmatrix}\tag{4.36}$$

where

$$\begin{aligned}
\Psi_{22} &= \mathbf{M}_{rn}(\mathbf{I}_r \otimes \mathbf{I}_{rn} \otimes \mathbf{I}_n) \\
&\quad \times \left\{ \left[\left(\boldsymbol{\Gamma}_1(\boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_0)^T \otimes \mathbf{I}_r \right) + (\mathbf{I}_u \otimes \boldsymbol{\Gamma}_1(\boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_0)) - \boldsymbol{\Omega}_0 \otimes \mathbf{I}_r - (\mathbf{I}_u \otimes \boldsymbol{\Omega}_0)\mathbf{I}_{ru} \right] \otimes vec(\boldsymbol{\rho}(\boldsymbol{\theta})) \right\}, \\
\Psi_{23} &= \mathbf{M}_{rn}(\mathbf{I}_r \otimes \mathbf{I}_{rn} \otimes \mathbf{I}_n) \left[(\boldsymbol{\Gamma}_1 \otimes \boldsymbol{\Gamma}_1)\mathbf{G}_u \otimes vec(\boldsymbol{\rho}(\boldsymbol{\theta})) \right], \\
\Psi_{24} &= \mathbf{M}_{rn}(\mathbf{I}_r \otimes \mathbf{I}_{rn} \otimes \mathbf{I}_n) \left[(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{G}_{r-u} \otimes vec(\boldsymbol{\rho}(\boldsymbol{\theta})) \right],
\end{aligned} \tag{4.37}$$

where $\mathbf{M}_{rn} \in R^{rn(rn+1)/2 \times (rn)^2}$ is the contraction matrix which is defined such that for a given matrix such as $\mathbf{A}$, $vech(\mathbf{A}) = \mathbf{M}_{rn} vec(\mathbf{A})$. Furthermore, $\boldsymbol{\Lambda}^{-\frac{1}{2}}(\boldsymbol{\Lambda} - \boldsymbol{\Lambda}_0)\boldsymbol{\Lambda}^{-\frac{1}{2}} \geq 0$, so the spatial envelope model decreases the asymptotic variance.

The partial derivatives $\frac{\partial \Psi_i}{\partial \phi_j^T}$ for $i = 1, 2$ and $j = 1, 2, 3, 4$ derivations are presented as follow:

In order to calculate $\frac{\partial \Psi_1}{\partial \phi^T}$, first we have:

$$\frac{\partial vec(\Psi_1(\phi))}{\partial \phi_1} = \frac{\partial vec(\boldsymbol{\Gamma}_1 \boldsymbol{\eta})}{\partial vec(\boldsymbol{\eta})^T} = \frac{\partial (\mathbf{I}_p \otimes \boldsymbol{\Gamma}_1) vec(\boldsymbol{\eta})}{\partial vec(\boldsymbol{\eta})^T} = \mathbf{I}_p \otimes \boldsymbol{\Gamma}_1$$

Similarly, we have:

$$\frac{\partial vec(\Psi_1(\phi))}{\partial \phi_2} = \frac{\partial vec(\boldsymbol{\Gamma}_1 \boldsymbol{\eta})}{\partial vec(\boldsymbol{\Gamma}_1)^T} = \frac{\partial (\boldsymbol{\eta}^T \otimes \mathbf{I}_r) vec(\boldsymbol{\Gamma}_1)}{\partial vec(\boldsymbol{\Gamma}_1)^T} = \boldsymbol{\eta}^T \otimes \mathbf{I}_r$$

Clearly, $\frac{\partial \Psi_1}{\partial \phi_3^T} = \mathbf{0}$, $\frac{\partial \Psi_1}{\partial \phi_4^T} = \mathbf{0}$.

In order to compute $\frac{\partial \Psi_2}{\partial \phi^T}$, since Ψ_2 does not depend on ϕ_1, therefore $\frac{\partial \Psi_2}{\partial \phi_1^T} = \mathbf{0}$. For $\frac{\partial \Psi_2}{\partial \phi_2^T}$, we have

$$\begin{aligned}
\frac{\partial vec(\Psi_2(\phi))}{\partial \phi_2} &= \frac{\partial vec\left[\left(\boldsymbol{\Gamma}_1\boldsymbol{\Omega}_1\boldsymbol{\Gamma}_1^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T\right) \otimes \boldsymbol{\rho}(\boldsymbol{\theta})\right]}{\partial vec(\boldsymbol{\Gamma}_1)} \\
&= \frac{\partial vech\left[\left(\boldsymbol{\Gamma}_1\boldsymbol{\Omega}_1\boldsymbol{\Gamma}_1^T + (\mathbf{I}-\boldsymbol{\Gamma}_1)\boldsymbol{\Omega}_0(\mathbf{I}-\boldsymbol{\Gamma}_1)^T\right) \otimes \boldsymbol{\rho}(\boldsymbol{\theta})\right]}{\partial vec(\boldsymbol{\Gamma}_1)} \\
&= \frac{\mathbf{M}_{rn}\partial vec\left[\left(\boldsymbol{\Gamma}_1\boldsymbol{\Omega}_1\boldsymbol{\Gamma}_1^T + (\mathbf{I}-\boldsymbol{\Gamma}_1)\boldsymbol{\Omega}_0(\mathbf{I}-\boldsymbol{\Gamma}_1)^T\right) \otimes \boldsymbol{\rho}(\boldsymbol{\theta})\right]}{\partial vec(\boldsymbol{\Gamma}_1)} \\
&= \mathbf{M}_{rn}(\mathbf{I}_r \otimes \mathbf{I}_{(r,n)} \otimes \mathbf{I}_n)\left[\frac{\partial vec\left[\left(\boldsymbol{\Gamma}_1\boldsymbol{\Omega}_1\boldsymbol{\Gamma}_1^T + (\mathbf{I}-\boldsymbol{\Gamma}_1)\boldsymbol{\Omega}_0(\mathbf{I}-\boldsymbol{\Gamma}_1)^T\right)\right]}{\partial vec(\boldsymbol{\Gamma}_1)} \otimes vec(\boldsymbol{\rho}(\boldsymbol{\theta}))\right] \\
&= \mathbf{M}_{rn}(\mathbf{I}_r \otimes \mathbf{I}_{(r,n)} \otimes \mathbf{I}_n) \\
&\times \left\{\left[\left(\boldsymbol{\Gamma}_1(\boldsymbol{\Omega}_1+\boldsymbol{\Omega}_0)^T \otimes \mathbf{I}_r\right) + (\mathbf{I}_u \otimes \boldsymbol{\Gamma}_1(\boldsymbol{\Omega}_1+\boldsymbol{\Omega}_0)) - \boldsymbol{\Omega}_0 \otimes \mathbf{I}_r - (\mathbf{I}_u \otimes \boldsymbol{\Omega}_0)\mathbf{I}_{(r,u)}\right] \otimes vec(\boldsymbol{\rho}(\boldsymbol{\theta}))\right\}
\end{aligned}$$

For $\frac{\partial \Psi_3}{\partial \phi_2^T}$, we have

$$\begin{aligned}
\frac{\partial vec(\Psi_2(\phi))}{\partial \phi_3} &= \frac{\partial vech(\boldsymbol{\Gamma}_1\boldsymbol{\Omega}_1\boldsymbol{\Gamma}_1^T \otimes \boldsymbol{\rho}(\boldsymbol{\theta}))}{\partial vech(\boldsymbol{\Omega}_1)^T} \\
&= \frac{\mathbf{M}_{rn}\partial vec(\boldsymbol{\Gamma}_1\boldsymbol{\Omega}_1\boldsymbol{\Gamma}_1^T \otimes \boldsymbol{\rho}(\boldsymbol{\theta}))}{\partial vech(\boldsymbol{\Omega}_1)^T} \\
&= \mathbf{H}_{rn}(\mathbf{I}_r \otimes \mathbf{I}_{(r,n)} \otimes \mathbf{I}_n)\left[\frac{\partial vec(\boldsymbol{\Gamma}_1\boldsymbol{\Omega}_1\boldsymbol{\Gamma}_1^T)}{\partial vech(\boldsymbol{\Omega}_1)^T} \otimes vec(\boldsymbol{\rho}(\boldsymbol{\theta}))\right] \\
&= \mathbf{M}_{rn}(\mathbf{I}_r \otimes \mathbf{I}_{(r,n)} \otimes \mathbf{I}_n)\left[(\boldsymbol{\Gamma}_1 \otimes \boldsymbol{\Gamma}_1)\mathbf{E}_u \otimes vec(\boldsymbol{\rho}(\boldsymbol{\theta}))\right]
\end{aligned}$$

Similarly, for $\frac{\partial \Psi_4}{\partial \phi_2^T}$, we have

$$\begin{aligned}
\frac{\partial vec(\Psi_2(\phi))}{\partial \phi_4} &= \frac{\partial vech(\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T \otimes \boldsymbol{\rho}(\boldsymbol{\theta}))}{\partial vech(\boldsymbol{\Omega}_0)^T} \\
&= \frac{\mathbf{M}_{rn}\partial vec(\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T \otimes \boldsymbol{\rho}(\boldsymbol{\theta}))}{\partial vech(\boldsymbol{\Omega}_0)^T} \\
&= \mathbf{M}_{rn}(\mathbf{I}_r \otimes \mathbf{I}_{(r,n)} \otimes \mathbf{I}_n)\left[\frac{\partial vec(\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)}{\partial vech(\boldsymbol{\Omega}_0)^T} \otimes vec(\boldsymbol{\rho}(\boldsymbol{\theta}))\right] \\
&= \mathbf{M}_{rn}(\mathbf{I}_r \otimes \mathbf{I}_{(r,n)} \otimes \mathbf{I}_n)\left[(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{(r-u)} \otimes vec(\boldsymbol{\rho}(\boldsymbol{\theta}))\right]
\end{aligned}$$

Having these derivatives together lead to obtain (4.37).

The asymptotic distribution (4.35) follows from Shapiro (1986, Proposition 4.1). In order to prove that $\boldsymbol{\Lambda}_0 \leq \boldsymbol{\Lambda}$, we have

$$\boldsymbol{\Lambda}_0 - \boldsymbol{\Lambda} = \mathbf{J}^{-1} - \Psi(\Psi^T\boldsymbol{\Lambda}\Psi)^\dagger\Psi = \mathbf{J}^{-\frac{1}{2}}\left[\mathbf{I}_{pr+r(r+1)/2} - \mathbf{J}^{\frac{1}{2}}\Psi(\Psi^T\boldsymbol{\Lambda}\Psi)^\dagger\Psi\mathbf{J}^{\frac{1}{2}}\right]\mathbf{J}^{-\frac{1}{2}}$$

Since the matrix $\mathbf{I}_{pr+r(r+1)/2} - \mathbf{J}^{\frac{1}{2}}\Psi(\Psi^T\boldsymbol{\Lambda}\Psi)^\dagger\Psi\mathbf{J}^{\frac{1}{2}}$ is the projection on to orthogonal complement of $span(\mathbf{J}^{\frac{1}{2}}\Psi)$, it is positive semidefinite, which implies that $\boldsymbol{\Lambda}_0 - \boldsymbol{\Lambda}$ is also positive semidefinite. In addition, we have

$$\boldsymbol{\Lambda}^{-\frac{1}{2}}(\boldsymbol{\Lambda} - \boldsymbol{\Lambda}_0)\boldsymbol{\Lambda}^{-\frac{1}{2}} = \mathbf{I}_{pr+r(r+1)/2} - \mathbf{J}^{\frac{1}{2}}\Psi(\Psi^T\boldsymbol{\Lambda}\Psi)^\dagger\Psi\mathbf{J}^{\frac{1}{2}}$$

which proves the last statement of the theorem.

4.7.2 Prediction Plot for Response Variables

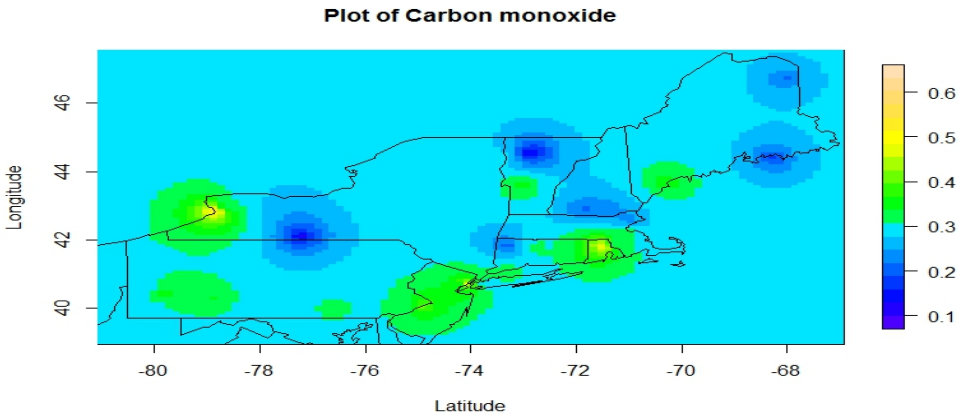

Figure 4.3: *Prediction plot of carbon monoxide for the study area. As it can be seen, the carbon monoxide is high in Rhodes Island, New York, New Jersey, and Buffalo which are highly populated and therefore there will be a lots of car and usage of fossil fuels which leads to high concentration of carbon monoxide in the air.*

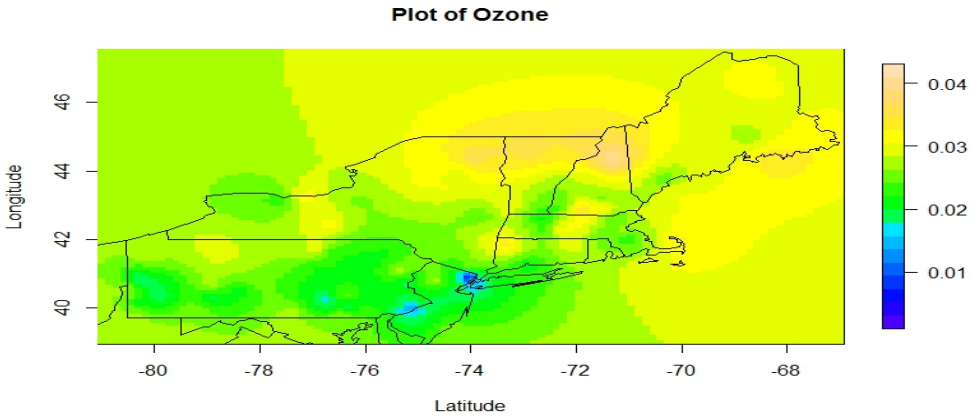

Figure 4.4: *Prediction plot of the log of the ground level Ozone for the study area. as it can be seen, the Ozone level is not high in the study area.*

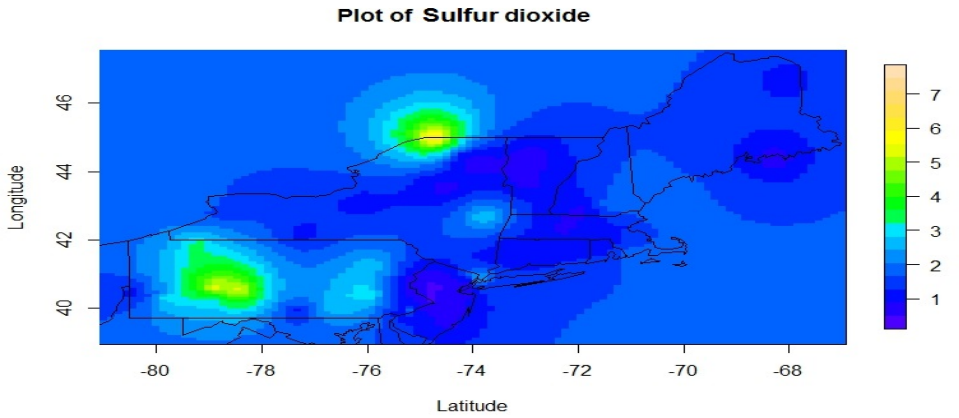

Figure 4.5: *Prediction plot of the log of the Sulfur dioxide for the study area. as it can be seen, the Sulfur dioxide is low for the most part of the study area. However, it is high in Johnstown where there exists a lot of defense manufacturing.*

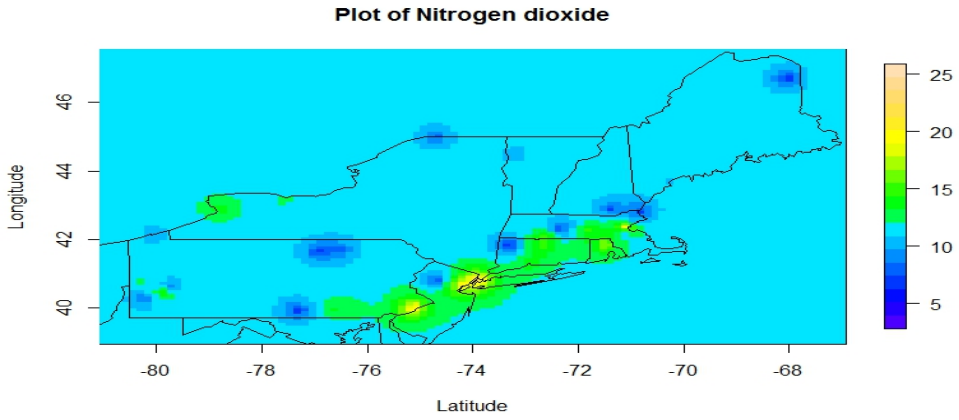

Figure 4.6: *Prediction plot of the log of the Nitrogen dioxide for the study area. as it can be seen, the Nitrogen dioxide is high in Newark, New York, Philadelphia, and Rhodes Island which are all highly populated areas.*

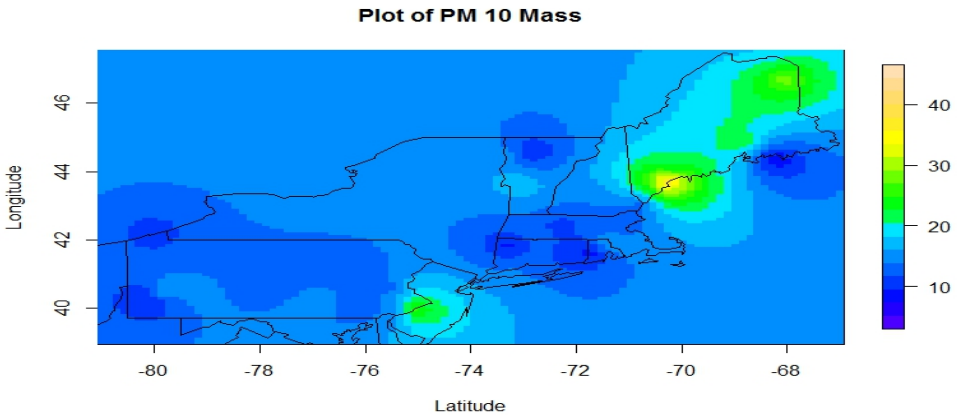

Figure 4.7: *Prediction plot of the log of the PM 10 Mass for the study area. as it can be seen, the PM 10 Mass is low for most part of the study area. However, it is high in New Jersey and Concord.*

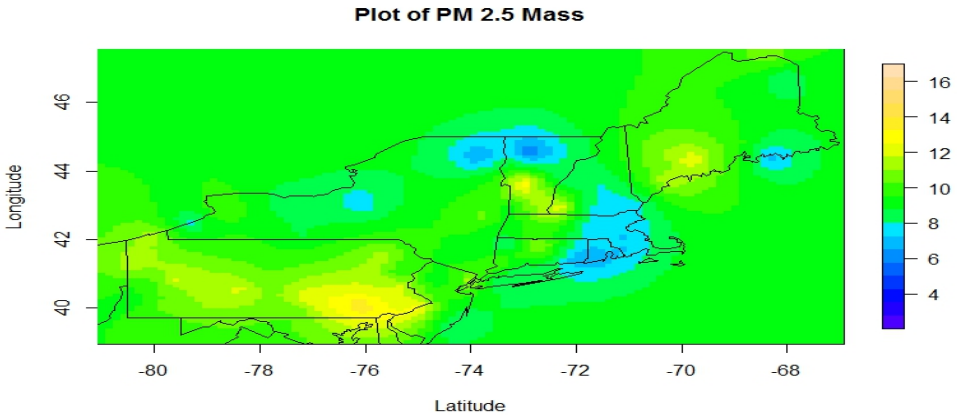

Figure 4.8: *Prediction plot of the log of the PM 2.5 Mass for the study area. as it can be seen, the PM 2.5 Mass is moderate in almost every place in the study area except for Philadelphia where it is high.*

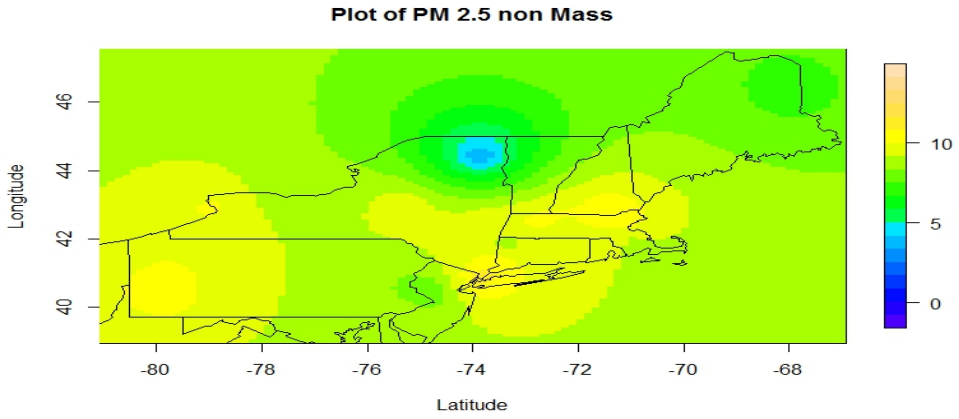

Figure 4.9: *Prediction plot of the PM 2.5 non Mass for the study area. as it can be seen, the PM 2.5 non Mass is moderately high in almost every place in the study area especially in Rhodes Island, Massachusetts, and New York.*

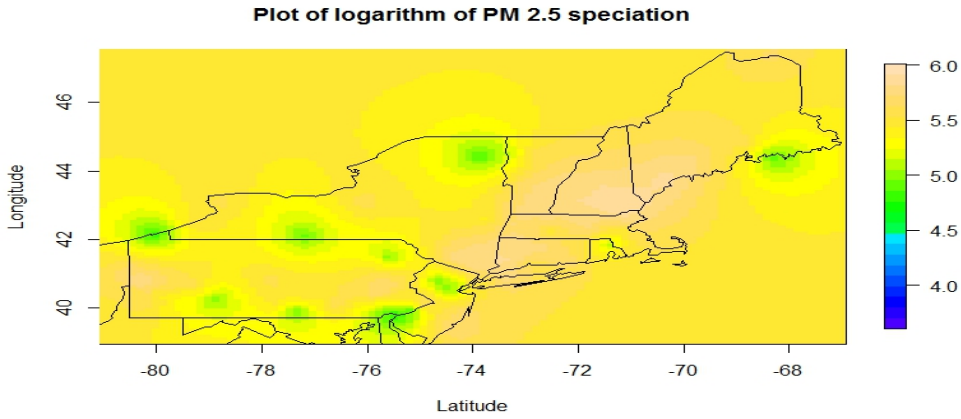

Figure 4.10: *Prediction plot of the log of the PM 2.5 speciation for the study area. as it can be seen, the PM 2.5 speciation is high in almost every place in the study area.*

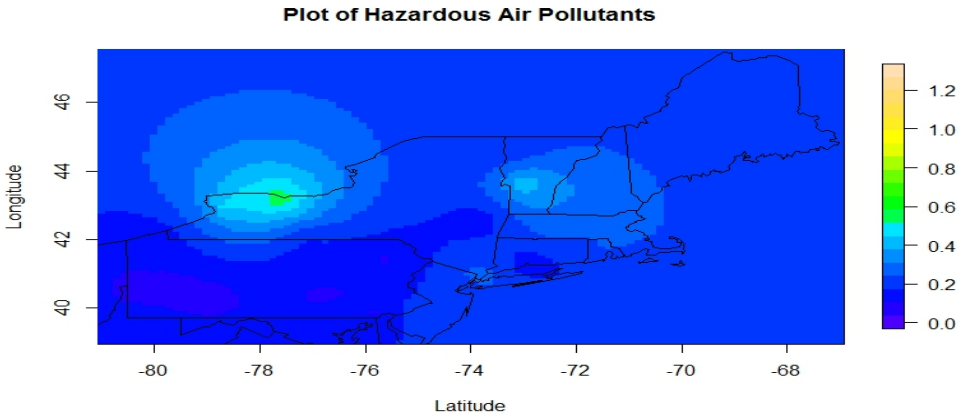

Figure 4.11: *Prediction plot of Hazardous air pollutants (HAPs) for the study area. As it can be seen, the HAPs is high in Rochester.*

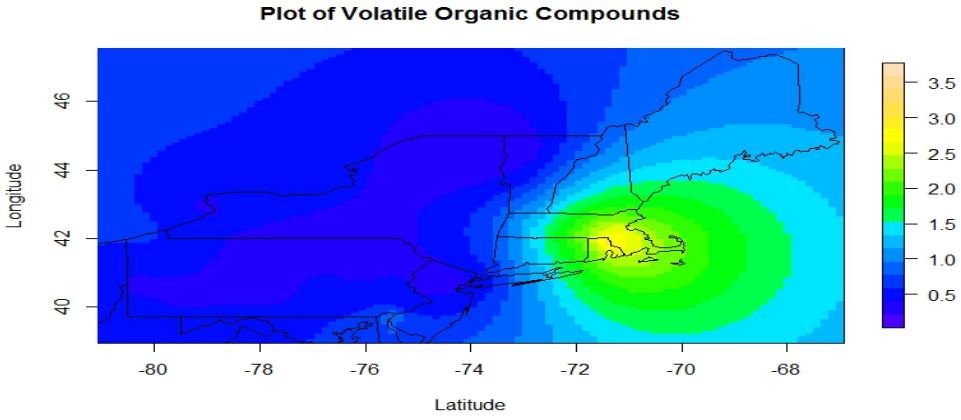

Figure 4.12: *Prediction plot of Volatile organic compounds (VOCs) for the study area. As it can be seen, the VOCs is high in Rhodes Island and Massachusetts.*

References

1. Abramowitz, M., and Stegun, I. A. (1964). Handbook of mathematical functions: with formulas, graphs, and mathematical tables. *Courier Corporation.*

2. Adler, R. J., and Taylor, J. E. (2009). Random fields and geometry. *Springer Science and Business Media.*

3. Adragni, K. P., and Cook, R. Dennis. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **367** 4385–4405.

4. Adcock, R. J. (1878). A problem in least squares. *The Analyst.* 53–54.

5. Akike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium of Information Theory.* Akademiai Kiado, Budapest.

6. Amato, U., Antoniadis, A., and De Feis, I. (2006). Dimension reduction in functional regression with applications. *Computational Statistics and Data Analysis.* **50** 2422–2446.

7. Antoniadis, A., Lambert-Lacroix, S., and Leblanc, F. (2003). Effective di-

mension reduction methods for tumor classification using gene expression data. *Bioinformatics.* **19** 563–570.

8. Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). Hierarchical modeling and analysis for spatial data. Crc Press.

9. Battye, William H and Bray, Casey D and Aneja, Viney P and Tong, Daniel and Lee, Pius and Tang, Youhua. (2016). Evaluating ammonia (NH 3) predictions in the NOAA National Air Quality Forecast Capability (NAQFC) using in situ aircraft, ground-level, and satellite measurements from the DISCOVER-AQ Colorado campaign. *Atmospheric Environment.* **140**, 342-351.

10. Bellman, R. E. (1961). *Adaptive Control Processes.* Princeton University Press, Princeton, NJ.

11. Bickel, P. J., Ritov, Y., and Tsybakov, A. B.. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics.* 1705–1732.

12. Bondell, H. D., and Li, L. (2009). Shrinkage inverse regression estimation for model-free variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71:1** 287–299.

13. Bradic, J., Fan, J., and Wang, W. (2011) Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B.* **73**, 325–349.

14. Bura E., and Pfeiffer R. M. (2003). Graphical Methods for Class Prediction Using Dimension Reduction Techniques on DNA Microarray Data. *Bioinformatics.* **19**, 1252–1258.

15. Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics.* **37** 373–384.

16. Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics.* **24** 2350–2383.

17. Byers, H. R. (1959). *General meteorology.* McGraw-Hill

18. Candes, E. J. (2006). Modern statistical estimation via oracle inequalities. *Acta Numerica*, **15**, 257–325.

19. Candes, E., and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics.* 2313–2351.

20. Chaudhuri, P., Hang, M. C., Loh, W. Y., and Yao, Y. (1991). Piecewise-polynomial regression trees. *Statistica Sinica.* **4** 143–167.

21. Chiles, J., and Delfiner, P. (1999). Geostatistics: Modeling spatial uncertainty. New York: John Wiley and Sons.

22. Christensen, R. (2001). *Advanced Linear Modeling.* Springer, New York.

23. Čížek, P., and Härdle, W. (2006). Robust estimation of dimension reduction space. *Computational Statistics and Data Analysis.* **51** 545–555.

24. Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association.* **89** 177–189.

25. Cook, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. *Proceedings of the section on Physical and Engineering Sciences.* 18–25.

26. Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91** 983–992.

27. Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics.* Wiley, New York.

28. Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science* 1–26.

29. Cook, R. D., and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association.* **104**, 197–208.

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

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

32. Cook, R. D., Li, B. (2002). Dimension reduction for the conditional mean in regression. *The Annals of Statistics.* **30** 455–474.

33. Cook, R. D., Li, B., and Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika.* **94** 569–584.

34. Cook, R. D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression (with discussion). *Statistica Sinica.* **20** 927–1010.

35. Cook, R. D., and Ni, L. (2005). Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association.* **100**, 410–428.

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

37. Cook, R. D., and Su, Z. (2015). Scaled predictor envelopes and partial least squares regression. *Technometrics.* To appear.

38. Cook, R. D., and Su, Z. (2016). Scaled predictor envelopes and partial least squares regression. *Technometrics*, **58**, 155–165.

39. Cook, R. D., Su, Z. and Yang, Y. (2015b). envlp: A MATLAB Toolbox for Computing Envelope Estimators in Multivariate Analysis. *Journal of Statistical Software.* **62** 1–20.

40. Cook, R. D., and Weisberg, S. (1991). Comment. *Journal of the American Statistical Association.* **86** 328–332.

41. Cook, R. D., and Weisberg, S. (2009). *An introduction to regression graphics.* John Wiley and Sons.

42. Cook, R. D., and Yin, X. (2001) Theory and Methods: Special Invited Paper: Dimension Reduction and Visualization in Discriminant Analysis (with discussion) *Australian and New Zealand Journal of Statistics.* **43**, 147–199.

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

44. Cook, R. D. and Zhang, X. (2015c). Simultaneous envelopes for multivariate linear regression. *Technometrics.* **57** 11-25.

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

46. Cooper, D. R., Schindler, P. S., and Sun, J. (2003). *Business research methods.* McGraw-Hill/Irwin New York, NY.

47. Cressie, N. (2015). *Statistics for spatial data.* John Wiley and Sons.

48. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm, with Applications. *Journal of the Royal Statistical Society, Series B.* **39** 1–38.

49. Duan, N., and Li, K. C. (1985). The Ordinary Least Squares Estimation for the General-Link Linear Models, with Applications. *Technical Report.*

50. Edelman, A., Tomas, A. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications.* **20**, 303–353.

51. Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics.* **32** 407–499.

52. Einbeck. J. and Tutz. G. (2006). Modelling beyond regression functions: an application of multimodal regression to speedflow data. *Journal of the Royal Statistical Society: Series C.* **55(4)** 461–475.

53. Fan, J., and Jiang, J. (2000). Variable bandwidth and one-step local M-estimator. *Science in China Series A: Mathematics.* **43** 65–81.

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

55. Fan, J., and Peng, H. (2004) Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics.* **32**, 928–961.

56. Fan, J., and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133.*

57. Fan, J., Fan, Y., and Barut, E. (2014). Adaptive robust variable selection. *The Annals of Statistics.* **42**, 324–351.

58. Friedman, J. H., and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association.* **76**, 817–823.

59. Friedman, J. H. (1994). An overview of computational learning and function approximation. *From Statistics to Neural Networks. Theory and Pattern Recognition Applications.* **1**.

60. Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software.* **33** 1–22.

61. Fruchter, B. (1954). Introduction to factor analysis. *Van Nostrand*

62. Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics.* **7** 397–416.

63. Fukumizu, K., Bach, F. R., and Jordan, M. I. (2009). Kernel dimension reduction in regression. *The Annals of Statistics.* 1871–1905.

64. Ghosh, D., and Chinnaiyan, A. M. (2005). Classification and selection of biomarkers in genomic data using LASSO. *BioMed Research International.* **2005:2** 147–154.

65. Guisan, A., Edwards, T. C. and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling.* **157:2** 89–100.

66. Härdle, W., and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association.* **84** 986–995.

67. Harville, D. A. (2008). *Matrix Algebra From a Statisticians Perspective.* Springer.

68. Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *The Annals of Statistics.* **29** 1537–1566.

69. Huber, P. J. Projection pursuit. *The Annals of Statistics.* 435–475.

70. Jonathan, A. V., Wu, K. F. K. and Donnell, E. T. (2016). A multivariate spatial crash frequency model for identifying sites with promise based on crash types. *Accident Analysis & Prevention.* **87**, 8–16.

71. Khoshnevisan, D. (2002). *Multiparameter Processes: an introduction to random fields.* Springer Science and Business Media.

72. Kioumourtzoglou, M. A., Schwartz, J. D., Weisskopf, M. G., Melly, S. J., Wang, Y., Dominici, F. and Zanobetti, A. (2016). Long-term PM2. 5 exposure

and neurological hospital admissions in the Northeastern United States. *Environmental health perspectives.* **124**, 23–29.

73. Kitanidis, P. K. (1983). Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water resources research.* **19:4** 909–921.

74. Koltchinskii, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli.* **15**, 799–828.

75. Knight, K., and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics.* 1356–1378.

76. Lægreid, M., Bockman, O. C., Kaarstad, O. (1999). *Agriculture, fertilizers and the environment.* CABI publishing.

77. Lave, L. B., and Seskin, E. P. (1973). An analysis of the association between US mortality and air pollution. *Journal of the American Statistical Association.* **68:342**, 284–290.

78. Latimer, A. M., Banerjee, S., Sang Jr, H., Mosher, E. S., and Silander Jr, J.A. (2009). Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. *Ecology Letters.* **12: 2**, 144–154.

79. Lekkou, A., Mouzaki, A., Siagris, D., Ravani, I. and Gogos, C. A. (2014). Serum lipid profile, cytokine production, and clinical outcome in patients with severe sepsis. *Journal of critical care.* **29:5**, 723–727.

80. Leng, C., Lin, Y., and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica.* **16**, 1273–1284.

81. Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics.* 1580–1616.

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

83. Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika.* **94** 603–613.

84. Li, L., Cook, R.D., Nachtsheim, C.J. (2005). Model-free variable selection. *Journal of the Royal Statistical Society Series B.* **67** 285–299.

85. Li, L., and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. *Technometrics*, **48** 503510.

86. Li, L., and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics*, **64**, 124–131.

87. Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association.* **86** 316–327.

88. Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association.* **87** 1025–1039.

89. Li, K. C. and Duan, N. H. (1989). Regression analysis under link violation. *The Annals of statistics.* **17** 1009–1052.

90. Li, K. C., Lue, H. H., and Chen, C. H. (2000). Interactive tree-truncated regression via principal Hessian directions. *Journal of the American Statistical Association.* **95** 547–560.

91. Liang, K. Y., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*. 3–40.

92. Linton, O., and Xiao, Z. (2007). A nonparametric regression estimator that adapts to error distribution of unknown form. *Econometric Theory*. **23** 371–413.

93. Liu, X., Srivastava, A. and Gallivan, K. (2004). Optimal linear representations of images for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **26**, 662–666.

94. Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press.

95. Mardia, K. V. and Marshall, R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*. **71:1** 135–146.

96. Meinshausen, N., and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*. 1436–1462.

97. Miller, A. (2002). *Subset selection in regression*. CRC Press.

98. Moradi Rekabdarkolaee, H., Boone, E., and Wang, Q. (2016). Robust estimation and variable selection in sufficient dimension reduction, *Computational Statistics & Data Analysis*.

99. Ni, L., and Cook, R. D., and Tsai, C. (2005). A note on shrinkage sliced inverse regression. *Biometrika*. **92:1** 242–247.

100. Naik, P., and Tsai, C. L. (2000). Partial least squares estimator for single-index models. *Journal of the Royal Statistical Society: Series B.* **62**, 763–771.

101. Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the lasso and its dual. *Journal of Computational and Graphical statistics.* **9** 319–337.

102. Phelan, J., Belyazid, S., Jones, P., Cajka, J., Buckley, J., and Clark, C. (2016). Assessing the Effects of Climate Change and Air Pollution on Soil Properties and Plant Diversity in Sugar Maple–Beech–Yellow Birch Hardwood Forests in the Northeastern United States: Model Simulations from 1900 to 2100. *Water, Air, & Soil Pollution.* **227**, 1–30.

103. Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., and Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Journal of American Medical Association.* **287:9**, 1132–1141.

104. Pope, C. A., Burnett, R. T., Thurston, G. D., Thun, M. J., Calle, Eugenia E. E., Krewski, D., and Godleski, J. J. (2004). Cardiovascular mortality and long-term exposure to particulate air pollution epidemiological evidence of general pathophysiological pathways of disease. *Circulation.* **109:1**, 71–77.

105. Proença, M. C., Rebelo, M. T., Alves, M. J., and Cunha, S. (2016). Ports and Airports: Gateways to Vector-Borne Diseases in Portugal Mainland. *World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering.* **10**, 232–237.

106. Rigaux, P., Scholl, M., and Voisard, A. (2001). *Spatial databases: with appli-*

cation to GIS. Morgan Kaufmann.

107. Rota, C. T., Wikle, C. K., Kays, R. W., Forrester, T. D., McShea, W. J., Parsons, A. W., and Millspaugh, J. J. (2016). A two-species occupancy model accommodating simultaneous spatial and interspecific dependence. *Ecology.* **97**, 48–53.

108. Samarov, A. M. (1993). Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association.* **88** 836–847.

109. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics.* **6** 461–464.

110. Seber, G. A. F. (2008). *A matrix handbook for statisticians.* John Wiley and Sons.

111. Seya, H., Murakami, D., Tsutsumi, M. and Yamagata, Y. (2015). Application of LASSO to the Eigenvector Selection Problem in Eigenvector-based Spatial Filtering. *Geographical analysis.* **47:3** 284–299.

112. Shi, P., and Tsai, C. L. (2002). Regression model selection: a residual likelihood approach. *Journal of the Royal Statistical Society: Series B.* **64** 237–252.

113. Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of American Statistical Association* **81**, 142–149.

114. Stein, M. L. (2001). *Spatial databases: with application to GIS*. Morgan Kaufmann.

115. Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging.* Springer Science and Business Media.

116. Stone, M. (1974). Cross-validation and multinomial prediction. *Biometrika.* **61** 509–515.

117. Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika.* **98** 133–146.

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

119. Su, Z. and Cook, R. D. (2012). Estimation of multivariate means with heteroscedastic errors using envelope models. *Statistica Sinica.* **23** 213–230.

120. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.* **162** 267–288.

121. Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B.* **73** 273–282.

122. Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics.* 614–645.

123. author=Wackernagel, H. (2013) *Multivariate geostatistics.* Springer Science and Business Media.

124. Wang, Y., Lin, X., Zhu, M., and Bai, Z. (2007). Robust estimation using Huber fuction with a data-dependent tuning constant. *Journal of Computational and Graphical Statistics.* **14** 468–481.

125. Wang, H., and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association.* **103** 811–821.

126. Wang, Q., and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Computational Statistics and Data*

Analysis. **52** 4512–4520.

127. Wang, Q., and Yao, W. (2012). An adaptive estimation of MAVE. *Journal of Multivariate Analysis.* **104** 88–100.

128. Wang, L., Wu, Y., and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association.* **107**, 214–222.

129. Wang, X., Jiang, Y., Huang, M., and Zhang, H. (2013). Robust variable selection with exponential squared loss. *Journal of the American Statistical Association.* **108** 632–643.

130. Wang, T., Xu, P., and Zhu, L. (2013). Penalized minimum average variance estimation. *Statist. Sinica.* **23** 543–569.

131. Welsh, A. H. and Ronchetti, E. (2002). A journey in single steps: robust one-step M-estimation in linear regression. *Journal of Statistical Planning and Inference.* **103** 287–310.

132. Wold, H. (1985). Partial least squares. *Encyclopedia of statistical sciences.* Wiley Online Library.

133. World Health Organization. (2003). *Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide: report on a WHO working group, Bonn, Germany 13-15 January 2003.* Copenhagen: WHO Regional Office for Europe.

134. Xia, Y., Tong, H., Li, W., and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B.* **64** 363–410.

135. Yao, W., Lindsay, B. G., and Li, R. (2012). Local modal regression. *Journal of Nonparametric Statistics.* **24** 647–663.

136. Yao, W., and Wang, Q. (2013). Robust variable selection through MAVE. *Computational Statistics and Data Analysis.* **63** 42–49.

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

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

139. Zeng, X. W., Vivian, E., Mohammed, K., Jakhar, S., Vaughn, M., Huang, J., Zelicoff, A., Xaverius, P., Bai, Z., and Lin, S. (2016). Long-term ambient air pollution and lung function impairment in Chinese children from a high air pollution range area: The Seven Northeastern Cities (SNEC) study. *Atmospheric Environment.* **138**, 144–151.

140. Zhang, H. (2007). Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics.* **18**, 125–139.

141. Zhang, H. H., Liu, Y., Wu, Y., and Zhu, J. (2008). Variable selection for the multicategory SVM via adaptive sup-norm regularization. *Electronic Journal of Statistics.* **2**, 149–167.

142. Zhang, C. H., and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics.* 1567–1594.

143. Zhou, J., and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *The Annals of Statistics.* 1649–1668.

144. Zhu, L. X., and Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics.* **24**,1053–1068.

145. Zhu, L., and Xue, L. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *Journal of the Royal Statistical Society: Series B.* **68** 549–570.

146. Zhu, L. X. Ohtaki, M., and Li, Y. (2007). On hybrid methods of inverse regression-based algorithms. *Computational Statistics and Data Analysis.* **51**, 2621–2635.

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

148. Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association.* **101** 1418–1429.

149. Zou, H., and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics.* **36**, 1509–1533.