8-2014

Penalized regressions for variable selection model, single index model and an analysis of mass spectrometry data.

Yubing Wan
University of Louisville

PENALIZED REGRESSIONS FOR VARIABLE SELECTION MODEL, SINGLE INDEX MODEL AND AN ANALYSIS OF MASS SPECTROMETRY DATA

By

Yubing Wan
B.S., China University of Mining and Technology, 2007
M.S., University of Texas-Pan American, 2009

Submitted to the Faculty of the
School of Public Health and Information Sciences of the University of Louisville
in Partial Fulfillment of the Requirements
for the degree of

Doctor of Philosophy

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, KY

August, 2014

PENALIZED REGRESSIONS FOR VARIABLE SELECTION MODEL, SINGLE INDEX MODEL AND AN ANALYSIS OF MASS SPECTROMETRY DATA

By

Yubing Wan
B.S., China University of Mining and Technology, 2007
M.S., University of Texas-Pan American, 2009

A Dissertation Approved on

July 30, 2014

by the following Dissertation Committee:

Maiying Kong, Ph.D.

Susmita Datta, Ph.D.

Karunarathna Kulasekera, Ph.D.

Dongfeng Wu, Ph.D.

Steven P. Jones, Ph.D., F.A.H.A.

DEDICATION

This dissertation is dedicated to my parents

Mr. Jianchun Wan

and

Mrs. Yanfeng Li Wan

who have given me invaluable educational opportunities.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my advisors, Dr. Maiying Kong and Dr. Susmita Datta, for their insightful guidance, caring, patience. My research would not have been possible without their help.

I would like to thank my committee members, Dr. K.B. Kulasekera and Dr. Dongfeng Wu from the Department of Bioinformatics and Biostatistics, and Dr. Steven P. Jones from the Department of Medicine, for their time to serve in my committee, for guiding my research and helping me to develop my background in biostatistics. I am very thankful to Dr. Roberto Bolli from the Department of Medicine for his financial support in the last two years.

I would like to thank the Department of Bioinformatics and Biostatistics at the University of Louisville for supporting me and making me grow and mature in Biostatistics. I extend my sincere thanks to all members of the Department of Bioinformatics and Biostatistics, and all those who help me directly or indirectly to complete the dissertation. In particular, I would like to thank Dr. Somnath Datta, who is one of the strictest but most enthusiastic and talented professors I have ever met. I thank him for his tremendous help in making me succeed in this program. I would like to thank the department administrative assistant Ms. Lynne Dosker for her kind help.

A special thanks to my friends and peers: Dake Yang, Joe Bible, Ruiqi Liu, Hyoyoung Choo, Jasmit Shah, Younathan Abdia, YuTing Chen, Xiaohong Li and more, for their inspiring discussions, friendship and all the fun we had in the past.

Last but not the least, I would like to thank my parents Jianchun Wan and Yanfeng Li, my younger sister Yufei Wan and my girl friend Fangbing Yu. They have been always supporting and encouraging me with their best wishes and love. To them I dedicate this dissertation.

ABSTRACT

PENALIZED REGRESSIONS FOR VARIABLE SELECTION MODEL, SINGLE INDEX MODEL AND AN ANALYSIS OF MASS SPECTROMETRY DATA

Yubing Wan

July 30, 2014

The focus of this dissertation is to develop statistical methods, under the framework of penalized regressions, to handle three different problems. The first research topic is to address missing data problem for variable selection models including elastic net (ENet) method and sparse partial least squares (SPLS). I proposed a multiple imputation (MI) based weighted ENet (MI-WENet) method based on the stacked MI data and a weighting scheme for each observation. Numerical simulations were implemented to examine the performance of the MI-WENet method, and compare it with competing alternatives. I then applied the MI-WENet method to examine the predictors for the endothelial function characterized by median effective dose and maximum effect in an ex-vivo experiment. The second topic is to develop monotonic single-index models for assessing drug interactions. In single-index models, the link function f is unnecessary monotonic. However, in combination drug studies, it is desired to have a monotonic link function f. I proposed to estimate f by using penalized splines with I-spline basis. An algorithm for estimating f and the parameter α in the index was developed. Simulation studies were conducted to examine the performance of the proposed models in term of accuracy in estimating f and α. Moreover, I applied

the proposed method to examine the drug interaction of two drugs in a real case study. The third topic was focused on the SPLS and ENet based accelerated failure time (AFT) models for predicting patient survival time with mass spectrometry (MS) data. A typical MS data set contains limited number of spectra, while each spectrum contains tens of thousands of intensity measurements representing an unknown number of peptide peaks as the key features of interest. Due to the high dimension and high correlations among features, traditional linear regression modeling is not applicable. Semi-parametric AFT model with an unspecified error distribution is a well-accepted approach in survival analysis. To reduce the bias caused in denoising step, we proposed a nonparametric imputation approach based on Kaplan-Meier estimator. Numerical simulations and a real case study were conducted under the proposed method.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

INTRODUCTION

This dissertation work is composed of three different while connected research projects.

In project I, I dealt with the issue of handling missing values of multiple predictors and studied its effect on variable selection and prediction. When missing values in some predictor variables exist, the statistical methods for variable selection and prediction could be challenging. Although multiple imputation (MI) (Rubin, 1987; Little and Rubin, 1987; 2002) is a universally accepted technique for solving missing data problem, how to combine the MI results for variable selection is not very clear because different imputations may result in different selected variables. The widely applied variable selection methods in the context of regression include the sparse partial least squares (SPLS) (Chun and Keleş, 2010) and the penalized least squares, e.g. the elastic net (ENet) method (Zou and Hastie, 2005). We proposed a MI-based weighted elastic net (MI-WENet) method, which is based on the stacked MI data sets and a weighting scheme in the regression procedure. In this method, MI accounts for sampling and imputation uncertainty for missing values, and the weight accounts for the observed information. Extensive numerical simulations were carried out to compare this MI-WENet method with other competing alternatives, such as the original ENet and SPLS methods. Moreover, we applied the MI-WENet method to examine the predictor variables for the endothelium dysfunction that is quantified by median effective dose (ED50) and maximum effect (Emax) in an ex-vivo acetylcholine-induced extension and phenylephrine-induced relaxation experiment.

The project II was inspired with the promising development of combination therapies within the pharmaceutical industry. In the combination drug studies,

the dose response relationship is often described by a response surface model (Greco et al., 1995). Denote the response at the combination dose (d_1, d_2) as y. A relative potency, say ρ is often used to describe how effective drug 2 is relative to drug 1. If we assume the dose-response curve for drug 1 is $y = f(d_1)$, then the dose-response curve for drug 2 is $f(\rho d_2)$. In the case that the two drugs have no interaction, i.e. the combination is additive, the effect of the combination dose (d_1, d_2) can be described by $f(d_1 + \rho d_2)$. If the effect at (d_1, d_2) is more (or less) than the effect of drug 1 at dose level $d_1 + \rho d_2$, we say the combination dose (d_1, d_2) is synergetic (or antagonistic) (Lee et al., 2007; Berenbaum, 1989). It is desirable that the response surface model is reduced to a dose-response when only one drug is applied. In the dose-response studies, the dose response relationship is often assumed to be monotonic (Kong and Eubank, 2006; Ramsay and Barahamowicz, 1989; Ramsay, 1998). Plummer and Short (1990) and Kong and Lee (2006) used monotonic parametric models to identify and quantify departures from additivity. However, the estimates of α can be biased when the parametric function f is misspecified. To avoid the problem caused by the misspecification of the function f, we propose the single-index model for assessing drug interactions. We do not assume any specific function form for f. Instead, we only assume that f is monotonic, and has continuous first and second derivatives. The function f is estimated by using penalized splines with I-splines as its basis functions, and the monotonicity of the function f is achieved by adding constraints to the coefficients of I-splines basis functions. (Kong and Eubank, 2006; Ramsay and Barahamowicz, 1989; Ramsay, 1998). Single-index models have been extensively studied in the statistical literatures (Stoker, 1986; Härdle and Stoker, 1989; Ichimura, 1993; Yu and Ruppert, 2002). A single-index model could be considered as an extension of a general linear model by replacing $X\alpha$ with a nonparametric function of $X\alpha$, $f(X\alpha)$, where X is a vector of covariates, α is the unknown param-

eter vector, and f is an unknown univariate link function. Although single-index models have been well studied, the function f in the model is unnecessary monotonic. We developed an algorithm for estimating the monotonic link function f and the parameter α in the single-index model. Simulation studies are carried out to examine the performance of the proposed model in term of accuracy in estimating f and α. In addition, we apply the proposed monotonic single-index method to examine the drug interaction of two drugs in a case study given by Harbron (2010).

Mass spectrometry (MS) data has been applied extensively and demonstrated great advantage in diagnosing and identifying proteomic biological markers to the discovery of key proteins and protein profiles associated with various types of diseases (Stoeckli et al., 2001; Adam et al., 2002; Aebersold and Mann, 2003; Rai and Chan, 2004; Datta et al., 2008; Datta and Pihur, 2010). Matrix-assisted laser desorption/ionization imaging mass spectrometry (MALDI-IMS) is a prosperous molecular technology that acquires information from intact proteins directly from thin sections of tissue. A typical MALDI-IMS data set contains hundreds of spectra, and each spectrum contains tens of thousands of intensity measurements representing an unknown number of protein/peptide peaks which are the key features of interest. Although some basic preprocesses like denoising and peak detection may identify some peaks for interested features, there are still hundreds or thousands of retained potentially important features which could be useful for the predictive modeling. Due to the high dimension as well as some high correlations among features, traditional linear regression modeling of survival times with proteomic features is not applicable. In order to predict patient survival using a predictive statistical model, one needs to consider dimension reduction and important feature selection on top of basic pre-processing of mass spectrometry data very carefully. Semi-parametric accelerated failure time (AFT) model

with an unspecified error distribution is a flexible and well-accepted approach in survival analysis. As far as we know, there are only a few publications on employment of the AFT model in high dimensional data setting, which mostly use the microarray platforms. Mostajabi et al. (2012) compared the performances of four relatively recent latent factor and/or penalized regression techniques (PLS, SPLS, LASSO and elastic net) in fitting AFT models based on high dimensional regressions, specifically to predict patient survival times using high dimensional mass spectrometry data. In project III, I focused on two popular techniques that performed best in the study of Mostajabi et al. (2012), namely SPLS and elastic net, to fit AFT models for predicting patient survivals. For identifying the subsets of features important for prediction analysis, some preprocessing steps like binning, standardizing, baseline correcting, and peak identifying are usually necessary. Depending on analysis goals, the preprocessing procedures can be different and complex in different literatures (Datta et al., 2007; Antoniadis et al., 2010; Morris et al., 2005; Mostajabi et al., 2012; Ndukum et al., 2011). In our methodology, we performed three basic preprocessing steps as baseline subtraction, alignment, and denoising to maintain as much information as possible before applying the AFT models in the subsequent survival analysis. To ensure the features used in analysis corresponding to real peaks, we applied a hard thresholding algorithm similar as in Datta et al. (2007); Ndukum et al. (2011); Mostajabi et al. (2012) to remove noise signals from the MS data. The denoising step ensures that the features used in analysis corresponding to real peaks. However, during the denoising, the intensities under thresholds are all considered as missing and are usually replaced with zeroes artificially. It is desirable to find a proper approach to retain sufficient true signals meanwhile reduce the bias for the subsequent predictive modeling analysis effectively. To solve this missing problem, we proposed a nonparametric imputation approach based on Kaplan-Meier estimator by con-

sidering the aligned intensities across all spectra as life times. We compared the predictive performance for the patient survival times with and without the imputation of the left censored peaks. Additionally, we compared different penalized regression schemes along with the AFT models to predict the patient survival times.

We anticipate that this dissertation research will significantly advance the area of variable selection and outcome prediction (dose response and patient survival time), with various types of predicting data, e.g. covariates with missing data and high dimensional mass spectrometry data.

TOPIC I: VARIABLE SELECTION MODELS BASED ON MULTIPLE IMPUTATION WITH APPLICATION FOR PREDICTING MEDIAN EFFECTIVE DOSE AND MAXIMUM EFFECT

2.1 Introduction

Missing data is a common problem in various settings including clinical trials, animal studies, and survey sampling (Rubin, 1987; Little and Rubin, 1987; 2002). When analyzing data with missing values, a straightforward strategy is to conduct a complete case analysis, where the observations with any missing values are ignored. This approach is simple yet ignores the possible differences between the complete cases and incomplete cases that may result in a substantial bias when the subjects with complete observations are not a random sub-sample of all subjects (Rubin, 1987). The complete case analysis also may lose information, and thus, results in incorrect inferences (Rubin, 1987; Van Buuren, 2012). Because experiments in medical research are usually expensive, the need for adequate handling of missing data is a constantly recognized source of concern (Wood et al., 2004). Instead of the complete case analysis, a more sophisticated approach called single imputation is used to impute the missing values with plausible values, and then statistical analyses are carried out on the imputed data set. However, the single imputation method ignores the uncertainty of imputation on the missing values that may lead to the underestimation of variances and the distortion of the correlation structure of the data. Therefore, simple single imputation is usually not recommended (Rubin, 1987; Little and Rubin, 1987; 2002; Van Buuren, 2012). Multiple imputation (MI) has gradually become a more well-accepted imputation-based statistical technique for handling missing data since the pub-

lication of Rubin's pioneering work for nonresponses in survey (Rubin, 1987). MI procedure involves imputing each missing value with M (> 1) independent plausible values, and then applying the standard analysis to each imputed data set. The final estimates of the parameters and their variances are obtained from the M sets of estimates using Rubin's rules, with accounting for the uncertainty among multiple imputations (Little and Rubin, 2002; Van Buuren, 2012). The objective of MI method is not to predict missing values as close as possible to the true values but to handle missing data so that valid statistical inferences can be made (Little and Rubin, 2002; Van Buuren, 2012). Rubin's rules have become the gold standard when data are missing at random (Wood et al., 2005; Van Buuren et al., 1999; Cohen et al., 2003). By the definition of Little and Rubin in (Rubin, 1987), the three general types of missing mechanism are: 1) missing complete at random (MCAR); 2) missing at random (MAR); and, 3) not missing at random (NMAR) (Rubin, 1987; Little and Rubin, 1987; 2002). Standard implementation of MI relies on an assumption that missing data are either MCAR or MAR, while the MI procedure may also be extended to the cases where missing data are NMAR (Van Buuren et al., 1999; Wood et al., 2008; Carpenter et al., 2007).

Variable selection is increasingly important in modern data analysis. Many techniques, such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), the elastic net (ENet) (Zou and Hastie, 2005), and the sparse partial least squares (SPLS) (Chun and Keleş, 2010), have been developed to select important variables that are associated with outcome variables. LASSO minimizes the restricted least squares with the constraint on the absolute values of the parameters (*i.e.*, L_1 norm), and ENet minimizes the constrained least squares with the constraint on the combination of the absolute and the squared values of parameters (Tibshirani, 1996; Zou and Hastie, 2005; Hastie et al., 2001). SPLS maximizes the correlation between outcome variables and the linear combinations of

predictor variables (covariates) with constraints on the L_1 norm of the parameters (Chun and Keleş, 2010). The constraint for LASSO can be considered as a special case of the ENet, and several studies have shown that ENet performs better than LASSO (Zou and Hastie, 2005). These methods have assumed that the observations in the data set are complete. How to apply these variable selection methods to the situation when there are missing values is an important yet unresolved problem.

Several approaches to combine the variable selection methods with MI techniques have been proposed recently (Wood et al., 2008; Heymans et al., 2007; Chen and Wang, 2013). Wood et al. proposed a "stacked" approach in (Wood et al., 2008) by combining the multiply imputed data sets into one and using a weighting scheme to account for the fraction of missing data in each predictor variable. However, the variable selection method used by them was the classical backward stepwise selection approach. Heymans et al. developed and tested a methodology combining MI with bootstrapping techniques for studying prognostic variable selection (Heymans et al., 2007). Chen and Wang proposed a MI-LASSO variable selection method as an extension of the LASSO method to MI-based data, which is, to the best of our knowledge, the only work combining the penalized least squares method with MI-based data (Chen and Wang, 2013). In the work (Chen and Wang, 2013), the observations with missing values and those without missing values are treated with equal importance. In this chapter, I proposed a MI-based weighted ENet (MI-WENet) method as an extension of the ENet to the stacked multiple imputed data, with a weight accounting for the proportion of the observed information for each observation. The cyclical coordinate descent methods (Friedman et al., 2010) are applied to minimize the weighted penalized least squares associated with the MI-WENet variable selection method.

To describe the new approach, in Section 2.2, I first review the two most pop-

ular variable selection methods: SPLS and ENet, and then propose the MI-based SPLS (MI-SPLS) and the MI-based weighted ENet (MI-WENet) for analyzing data with missing values. In Section 2.3, I carry out extensive numerical simulations to evaluate the performance of the proposed methods, and compare the performance of the proposed methods with the other competing methods. For Section 2.4, I apply the proposed MI-WENet method to examine the predictor variables for the maximum effect and the median effective dose in an ex-vivo phenylephrine-induced extension and acetylcholine-induced relaxation experiment study. Finally, I provide a discussion of the pros and cons of our current approach in Section 2.5.

2.2 Methods

Let Y_i denote the outcome variable and X_{ij} be the j^{th} predictor variable $(j = 1, ..., p)$ for the i^{th} subject $(i = 1, ..., n)$. Without loss of generality, I assume that Y_i and X_{ij} are standardized to have zero mean and unit standard deviation. For simplicity, let us consider the following linear regression model:

$$Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i = X_i\beta + \varepsilon_i, \quad i = 1, ..., n, \tag{2.1}$$

where the regression coefficients $\beta = (\beta_1, ..., \beta_p)^T$ are unknown parameters to be estimated, and the error term ε_i are independently identically distributed as $N(0, \sigma^2)$.

2.2.1 Review of SPLS and ENet

The sparse partial least squares regression (SPLS) (Chun and Keleş, 2010) is an extension of partial least squares regression (PLS) (Wold, 1985) to achieve simultaneous dimension reduction and variable selection. The PLS begins with

calculating the first latent direction vector t_1 as $X\hat{\beta}^{(1)}$, where $\hat{\beta}^{(1)}$ is obtained by maximizing the correlation between the response variable Y and the linear combination of covariates, $X\beta$, *i.e.*,

$$\hat{\beta}^{(1)} = \arg\max_{\beta} \left\{ \beta^T X^T Y Y^T X \beta \right\}, \quad \text{subject to } \beta^T \beta = 1. \tag{2.2}$$

Suppose the k^{th} ($k \geq 1$) direction vector, $t_k = X\hat{\beta}^{(k)}$, has been obtained. Denote $T = (t_1, t_2, ..., t_k)$ and $M_T = I - T(T^T T)^{-1} T^T$, the $(k+1)^{th}$ direction vector can be obtained by solving (2.2), with Y replaced by its orthogonal projection onto the complementary of the column space of the known direction vectors T, *i.e.*, replacing Y by $M_T Y$. This process is repeated to obtain a small number of direction vectors. Regressing the original Y on those direction vectors result in a relationship between Y and X due to each direction vector is a linear combination of the covariates X. PLS has become a very popular tool in the field of chemometrics and bioinformatics (Datta, 2001; Pihur et al., 2008). The SPLS achieves the sparsity of the coefficients on X by adding the L_1 constraints on β (Chun and Keleş, 2010). For example, $\hat{\beta}^{(1)}$ is updated as

$$\arg\max_{\beta} \left\{ \beta^T X^T Y Y^T X \beta \right\}, \quad \text{subject to } \beta^T \beta = 1 \text{ and } \|\beta\|_{L_1} \leq \lambda, \tag{2.3}$$

where $\|\beta\|_{L_1} = \sum_{j=1}^{P} |\beta_j|$. The L_1 constraint is added to obtain each direction vector (Chun and Keleş, 2010). SPLS obtains good performance in prediction and variable selection by producing sparse linear combinations of the original predictors, and is especially applicable when p is much greater than n (Chun and Keleş, 2010).

The elastic net (ENet) (Zou and Hastie, 2005) is a widely applied regulation and variable selection method. The ENet estimator is obtained by undoing the

shrinkage for the naïve elastic net estimator that is obtained by minimizing the penalized least squares

$$\mathrm{L}(\lambda, \alpha, \beta_0, \beta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda P_\alpha(\beta), \tag{2.4}$$

where

$$P_\alpha(\beta) = \alpha \|\beta\|_{L_1} + \frac{1}{2}(1-\alpha)\|\beta\|_{L_2} = \sum_{j=1}^{p} \left\{ \alpha|\beta_j| + \frac{1}{2}(1-\alpha)\beta_j^2 \right\}. \tag{2.5}$$

Here P_α is the elastic net penalty that is a compromise between the ridge regression penalty ($\alpha = 0$) (Hoerl and Kennard, 1970) and the LASSO penalty ($\alpha = 1$) (Tibshirani, 1996). Ridge regression is known to shrink the coefficients of correlated predictor variables, allowing them to borrow strength from each other (Hoerl and Kennard, 1970; Hastie et al., 2001). The elastic net penalty with $\alpha = 1 - \varepsilon$, for some small $\varepsilon > 0$, performs much like the LASSO but removes any degeneracies and wild behavior caused by extreme correlations (Friedman et al., 2010). For a given λ, as α increases from 0 to 1, the sparsity of the solution to (4.4), *i.e.*, the number of coefficients being zero, increases monotonically from 0 to the sparsity of the LASSO solution. The naïve elastic net estimator obtained from (4.4) and (4.5) does not perform satisfactorily (Zou and Hastie, 2005), while the elastic net estimator that undoes the shrinkage for the naïve elastic net, performs much better even compared with LASSO and ridge regression. The ENet estimator is obtained as

$$\hat{\beta}(\mathrm{ENet}) = (1 + \lambda(1-\alpha))\, \hat{\beta}(\mathrm{naßve\ ENet}). \tag{2.6}$$

The ENet penalty is particularly useful in the cases that p is greater than n and there are many correlated predictors (Zou and Hastie, 2005), which has also been shown in our simulation studies.

2.2.2 MI-based SPLS and MI-based Weighted ENet

Both the SPLS and ENet methods assume that all covariates and outcome variables are fully observed. In the cases that there are missing values, Rubin's rules provide a general framework to handle missing problems provided missing data are missing at random (MAR) or missing completely at random (MCAR) (Rubin, 1987; Little and Rubin, 1987; 2002; Van Buuren, 2012). However, Rubin's rules can not be directly applied to SPLS or ENet, because the variables selected for one imputed data set may be quite different from those based on another imputed data set. To the best of our knowledge, there is no standard rule to combine the selected variables resulted from different imputed data sets (Cohen et al., 2003; Wood et al., 2008; Chen and Wang, 2013; Schomaker and Heumann, 2013).

To overcome the shortcoming in combining the multiple results from MI data, we propose to select variables based on the stacked MI data. To be specific, let us assume that the outcome variable is fully observed, but the predictor variables may have some missing values. The missing values in the variables are imputed M times independently to generate M imputed data sets. We denote the m^{th} imputed data set as $(y_i;\ x_{i1}^{(m)}, ..., x_{ip}^{(m)})_{i=1}^{n}$, for $m = 1, ..., M$, where $x_{ij}^{(m)}$ is the value of the j^{th} predictor variable for the i^{th} subject in the m^{th} imputed data set. If X_{ij} is observed, then we have $x_{ij}^{(1)} = ... = x_{ij}^{(M)} = x_{ij}$; and if X_{ij} is missing, then $x_{ij}^{(m)}$ may take different values in each imputation. Popular softwares for implementing MI procedure include the R-packages *mice* (Van Buuren and Groothuis-Oudshoorn, 2011) and *mi* (Su et al., 2011), the SAS software *IVEware* (Raghunathan et al., 2001), and a module named *MULTIPLE IMPUTATION* in SPSS. In the simulation studies, we applied the R-package *mice* that is based on the sequential regression MI, *i.e.* the multivariate imputation by chained equations, to impute missing data (Van Buuren and Groothuis-Oudshoorn, 2011; Raghunathan et al., 2001). In

applying the R-package *mice*, users are allowed to specify the conditional distribution of each variable on the other variables in the data. The imputation was carried out based on the specified conditional distribution for the missing variables (Van Buuren and Groothuis-Oudshoorn, 2011).

Once M imputed data sets are obtained, one may stack the M imputed data sets as a large complete data set having $M \times n$ observations. SPLS and ENet can be directly applied to this single stacked data set. These approaches are called MI-based SPLS (MI-SPLS) and MI-based ENet (MI-ENet), respectively. In general, the estimates based on the stacked MI data are unbiased if the estimates based on a single data set are unbiased, while the standard errors based on the stacked MI data will be under-estimated if they can be estimated (Cohen et al., 2003). For the MI-ENet method, a simple way to correct the underestimated errors is to apply a weight to each observation. Denote this weight by w_i for subject i. For the stacked M imputed data sets, one could assign $w_i = 1/M$ thus the overall weight for a subject is 1. This weighting scheme puts the same weight for each subject and ignores the degree of missing information. A more legitimate way is to assign weights according to the quality of the observed information. If a subject has more missing predictor variables, the weight assigned to the subject should be smaller. We propose to assign the weight $w_i = f_i/M$, where f_i is the fraction of observed values for subject i, *i.e.*, the ratio of number of observed variables for the subject i to the total number of predictor variables p. This approach is named as MI-based weighted ENet (MI-WENet) method.

The MI-WENet minimizes the following penalized weighted least squares

$$\frac{1}{2n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i \left(y_i - \beta_0 - x_i^{(m)T}\beta\right)^2 + \lambda P_\alpha(\beta), \tag{2.7}$$

where $\beta = (\beta_1, ..., \beta_p)$. The penalty here is the same as the ENet penalty in (4.4).

I propose to standardize each predictor variable first based on the available data, then carry out the multiple imputation to get M imputed data sets. In the stacked data, the values for each variable may not have mean zero and variance 1 and the intercept may not be the mean of the observed responses anymore. Thus β_0 needs to be estimated in the same manner as the other regression parameters β. By avoiding any re-standardization in the stacked data, $\sum_{m=1}^{M} w_i \left(y_i - \beta_0 - x_i^{(m)T}\beta\right)^2$ will reduce to $\left(y_i - \beta_0 - x_i^T\beta\right)^2$, if there is no missing predictor variable for subject i. Thus, the objective function is reduced exactly to the standard ENet, when there is no missing value at all in the original data.

Denote the objective function (2.7) as $R(\beta_0, \beta)$. To solve for (β_0, β), a coordinate descent method can be applied (Friedman et al., 2010). Assuming the current estimated $\tilde{\beta_0}$ and $\tilde{\beta}$ are known, we wish to update $\tilde{\beta_j}$ as $\tilde{\beta_j} + \Delta\beta_j$ by partially optimizing $R(\beta_0, \beta)$ with respect to β_j $(j = 0, 1, ..., p)$. Note that the gradient for $\Delta\beta_j$ at $\beta_j = \tilde{\beta}_j$, which only exists if $\beta_j \neq 0$, is

$$\frac{\partial R(\beta_0, \beta)}{\partial \Delta\beta_j} = \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i \left(y_i - \beta_0 - x_i^{(m)T}\beta - x_{ij}^{(m)}\Delta\beta_j\right)\left(-x_{ij}^{(m)}\right) + \lambda\left(1-\alpha\right)\left(\beta_j + \Delta\beta_j\right) + \text{sign}\left(\beta_j\right)\lambda\alpha, \quad (2.8)$$

where $(\beta_0, \beta) = (\tilde{\beta}_0, \tilde{\beta})$. Set $\frac{\partial R(\beta_0,\beta)}{\partial \Delta\beta_j} = 0$, one can get

$$\Delta\beta_j = \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i \left(y_i - \beta_0 - x_i^{(m)T}\beta\right) x_{ij}^{(m)} - \text{sign}(\beta_j)\lambda\alpha - \lambda(1-\alpha)\beta_j}{\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i x_{ij}^{(m)2} + \lambda(1-\alpha)}, \quad (2.9)$$

where $(\beta_0, \beta) = (\tilde{\beta}_0, \tilde{\beta})$. Then β_j is updated as follows:

$$\begin{aligned}
\tilde{\beta}_j^{(new)} &= \tilde{\beta}_j + \Delta\beta_j \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i \left(y_i - \beta_0 - \sum_{l=1, l\neq j}^{p} x_{il}^{(m)}\beta_l\right) x_{ij}^{(m)} - \text{sign}(\beta_j)\lambda\alpha}{\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i x_{ij}^{(m)2} + \lambda(1-\alpha)} \\
&= \frac{S\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i \left(y_i - \beta_0 - \sum_{l=1, l\neq j}^{p} x_{il}^{(m)}\beta_l\right) x_{ij}^{(m)},\ \lambda\alpha\right)}{\frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i x_{ij}^{(m)2} + \lambda(1-\alpha)},
\end{aligned} \tag{2.10}$$

where $(\beta_0, \beta) = (\tilde{\beta}_0, \tilde{\beta})$, and $S(z, \gamma)$ is the soft-thresholding operator with value

$$\text{sign}(z)\left(|z| - \gamma\right)_+ = \begin{cases} z - \gamma & \text{if } z > 0 \text{ and } \gamma < |z| \\ z + \gamma & \text{if } z < 0 \text{ and } \gamma < |z| \\ 0 & \text{if } \gamma \geq |z|. \end{cases}$$

To reduce imputation burden, for a given multiple imputed stacked data set and a given weight, one may first calculate and store the following quantities:

$$XY_j = \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i y_i x_{ij}^{(m)}, \quad \text{for } j = 0, 1, ..., p.$$

$$XX_{jj'} = \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_i x_{ij}^{(m)} x_{ij'}^{(m)}, \quad \text{for } 0 \leq j \leq j' \leq p.$$

Here $x_{ij}^{(m)}$ is set to 1 for $j = 0$. Suppose that $\tilde{\beta}_j^{(old)}$ $(j = 0, 1, ..., p)$ are the available values at the previous iteration, one may update β_j $(j = 0, 1, ..., p)$ by

$$\tilde{\beta}_j^{(new)} = \frac{S\left(XY_j - \sum_{l<j} XX_{jl}\tilde{\beta}_l^{(new)} - \sum_{l>j} XX_{jl}\tilde{\beta}_l^{(old)},\ \lambda\alpha\right)}{XX_{jj} + \lambda(1-\alpha)}. \tag{2.11}$$

The procedure is repeated until convergence to get the estimates for β_j $(j = 0, 1, ..., p)$. These estimates are similar to the naïve ENet estimates (Zou and Hastie, 2005), which can be obtained by a truncation at $\lambda\alpha$ and a shrinkage with a factor

$XX_{jj} + \lambda(1-\alpha)$ for β_j. A better estimate that undoes the shrinkage is obtained by

$$\hat{\beta}_j(\text{weighted ENet}) = \left(XX_{jj} + \lambda(1-\alpha)\right)\hat{\beta}_j(\text{weighted naive ENet}). \tag{2.12}$$

The weighted ENet estimates in (2.12) are used in the simulations in Section 2.3 and the case study in Section 2.4, and performs well in both variable selection and prediction.

In the present work, I applied 10-fold cross validation method to select the tuning parameters α and λ. Here $\alpha \in (0,1)$, and $\lambda > 0$. Because (α, λ) determines the soft-threshold boundary, I start with a sequence grid value for α. For each fixed α, I compute the solution for a decreasing sequence of values for λ starting at the largest value λ_{max} for which the entire vector $\tilde{\beta} = 0$, *i.e.*,

$$\alpha\lambda_{max} = \max_{0 \le j \le p} \mid XY_j \mid,$$

and set $\lambda_{\min} = \epsilon\lambda_{\max}$ with $\epsilon = 0.001$. I construct a sequence of λ values decreasing from $\lambda_{\max}$ to $\lambda_{\min}$ on the log-scale. The pair of (α, λ) is chosen such that the cross validation error is minimized.

2.3 Simulation

In this section, I design different simulation schemes to examine the performance of the proposed MI-WENet method and compare it with the other methods, such as MI-SPLS and MI-ENet. The different simulation scenarios are reported in Section 2.3.1; the corresponding simulation results are reported in Section 2.3.2.

2.3.1 Simulation Settings

In the simulation studies, I assume that the underlying model is known and has the form of $Y_i = X_i\beta + \varepsilon_i$, for $i = 1, ..., n$, where $X_i = (X_{i,1}, ..., X_{i,p})$, $\beta = (\beta_1, ..., \beta_p)^T$, and $\varepsilon_i \sim N(0, \sigma^2)$. The predictor variables for each subject were generated from a multivariate normal distribution with mean zero and a covariance matrix Σ. σ was set as the value such that the signal to noise ratio is 2, *i.e.*, $\sqrt{\beta^T\Sigma\beta}/\sigma = 2$.

Simulation scenarios were designed based on various assumptions of sample size n, number of predictor variables p, missing mechanism, missing pattern and correlation structure of the predictor variables. Correlation structure for the predictor variables of the i^{th} subject ($i = 1, ..., n$) was tested under three specifications: 1) compound symmetry with low correlation, *i.e.*, $\text{corr}(X_{i,j}, X_{i,j'}) = 0.1$; 2) compound symmetry with medium correlation, *i.e.*, $\text{corr}(X_{i,j}, X_{i,j'}) = 0.5$; and 3) first-order autoregressive (AR(1)), *i.e.*, $\text{corr}(X_{i,j}, X_{i,j'}) = 0.8^{|j-j'|}$, for $j, j' = 1, ..., p$ and $j \neq j'$, respectively. I set the homogenous variances as 1 for all $X_{i,j}$, so the covariance matrix Σ was same as the correlation matrix. Under each specification, I induced missing values under the MCAR and MAR mechanisms, respectively; and for each missing mechanism, missing values were generated with independent and monotone missing patterns, respectively. In total, 17 scenarios were tested in our simulations, which I believe have covered most situations in practical application. The independent missing pattern means that the missing observations for different variables are independent, and the monotone missing pattern is that a missing observation in x_{ij} (where i is the subject index, and j is the variable index) implies that all observations $x_{ij'}$ for $j \leq j' \leq p$ are missing.

For each scenario with fixed n, p, Σ, missing mechanism and missing pattern, the following steps are carried out:

1. Generate fully observed predictor variables for X_i $(i = 1, ..., n)$.

2. Generate the outcome variable for Y_i from the underlying model $Y_i = X_i\beta + \varepsilon_i$ $(i = 1, ..., n)$, where $\varepsilon_i \sim N(0, \sigma^2)$.

3. Independently generate test data set (x_t, y_t) $(t = 1, ..., n_t)$ by repeating steps 1 & 2, where the sample size n_t is larger than n ($n_t = 1000$ in our simulations).

4. Fit the full data set that has a sample size n and has been generated in steps 1 & 2 by using SPLS and ENet, respectively (see the rows named as *Full-SPLS* and *Full-ENet* in Tables 2.1-2.3).

5. Induce missing values for the predictor variables according to each pre-specified missing mechanism and missing pattern.

6. Fit the data set including complete cases only by using SPLS and ENet, respectively (see the rows named as *CC-SPLS* and *CC-ENet* in Tables 2.1-2.3).

7. Impute missing values M times (M=5), and stack the M imputed data sets into an enlarged one.

8. Perform SPLS, ENet and WENet based on the first single imputed data set (see the rows named *SI-SPLS*, *SI-ENet* and *SI-WENet* in Tables 2.1-2.3), and based on the stacked data set (see the rows named *MI-SPLS*, *MI-ENet* and *MI-WENet* in Tables 2.1-2.3).

9. Repeat Steps 1-8 100 times, and summarize the averaged key performance measures of each method.

The key performance measures for each method under each simulation scenario are predicted mean squared error (PMSE), mean squared error (MSE), sen-

sitivity and specificity. The PMSE is defined as

$$\text{PMSE}(\hat{\beta}) = \frac{1}{n_t}\sum_{t=1}^{n_t}(y_t - x_t\hat{\beta})^2,$$

where x_t and y_t are fully observed independent test data generated in Step 3, and $\hat{\beta}$ is the estimate of the underlying regression parameter β for each model. PMSE is obtained by averaging the predicted errors on a large number of observations, where I have set n_t as 1000. The MSE is defined as

$$\text{MSE} = (\hat{\beta} - \beta)^T\Sigma(\hat{\beta} - \beta),$$

where $\hat{\beta}$ and β are the same as for PMSE, Lower values of PMSE and MSE are desirable. The sensitivity is defined as the fraction of variables selected among those whose coefficients are not zero in the underlying model, and the specificity is defined as the fraction of variables not selected among those whose coefficients are zeros in the underlying model. Larger sensitivity and specificity indicate a better performance.

To examine the performance of different methods, I first fixed $p = 12$, $n = 50$ and $\beta = (3, 1.5, 0, 0, 2, 0, 3, 1.5, 0, 0, 2, 0)^T$, and I considered the combinations of different missing mechanism (MCAR and MAR), different missing pattern (independent and monotone) and different correlation structure for the predictor variables. Under the MCAR scheme, the independent missing pattern was generated by independently removing 16% of the observations from each of the first 6 predictor variables, which resulted in around 50% observations containing missing values; the monotone missing pattern was generated by first inducing missing values to the 8% of randomly sampled observations from the 1^{st} to 6^{th} predictors, and then repeatedly adding missing values to another 8% randomly sampled observations from the 2^{nd} to 6^{th}, 3^{rd} to 6^{th}, 4^{th} to 6^{th}, 5^{th} to 6^{th}, and the 6^{th} only

predictor variables, which eventually resulted in 48% subjects containing missing values. The simulation results for MCAR, with different missing patterns and different correlation structures for the predictor variables, are reported in Table 2.1. For MAR, missing values were induced by the following logistic regression model:

$$\text{logit}\{Pr(X_{ij^{(m)}} \text{ is missing} | X_{ij^{(c)}}, Y_i)\} = X_{ij^{(c)}} + Y_i, \quad \text{for } i = 1, ..., n. \tag{2.13}$$

Here $j^{(m)} = 1, ..., p_1$ are indices for the predictor variables in which missing values are to be induced, and $j^{(c)} = p_1 + j^{(m)}$ are indexes for the completely observed predictor variables. When p equals to 12, p_1 is set as 6. For independent missing pattern, the procedure to generate missing values was the same as in the MCAR cases, except that the 16% removed observations for each of the 6 missing predictor variables were selected by the highest probabilities calculated from the logistic model (2.13). For monotone missing pattern, we applied the logistic model (2.13) to the whole data set first, and removed 8% observations from the 1^{st} to 6^{th} predictor variables according to the missing probabilities for the 1^{st} predictor variable. We then applied the logistic model (2.13) to the remaining data set with complete cases only, and removed 8% additional observations from the 2^{nd} to 6^{th} predictor variables according to the missing probabilities of the 2^{nd} predictor variable. Repeating above procedure until 8% additional observations were removed for the 6^{th} predictor variable only, resulted in 48% subjects containing missing values in total. The corresponding simulation results for MAR are reported in Table 2.2.

I also conducted simulations with different combinations of p and n, under the specification of monotone MAR and AR(1) correlation structure, so that the performance of different methods with large p and small n (say $p = 24$, 48, and 60, with n fixed at 50) and with small p and large n (say $n = 50$, 100, and 200,

with p fixed at 12) can be examined. Here, when $p > 12$, the β in the underlying models were set as the repetitions of $(3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0)$. The procedures for generating monotone MAR missing values were similar as when $p = 12$ and $n = 50$. In the cases when $p = 24$ and $p = 48$, p_1 in (2.13) were set as $p/2$, and the the percentages of missing values in each iteration were controlled at 4% and 2%, respectively. When $p = 60$, p_1 was set as 24 and the percentage of missing value in each iteration was controlled at 2%. The total missing percentage was fixed at 48% under each scenario. The corresponding simulation results are reported in Table 2.3.

The number of simulation runs is 100 in Tables 2.1-2.3. To examine whether a large number of simulation runs impacts the simulation results, I carried out the simulations with 500 runs for each scenario showed in Table 2.3. The corresponding results are reported in Table 2.4.

2.3.2 Simulation Results

The results for MCAR with different missing patterns and different correlations for X are summarized in Table 2.1, and results for MAR are summarized in Table 2.2. The results for MCAR (Table 2.1) and MAR (Table 2.2) explain consistent improvement in the estimation and prediction errors using the MI-WENet procedure compared to others. From Tables 2.1 and 2.2, we see that: (1) Full-ENet is consistently having lower PMSE and MSE than those from Full-SPLS. When correlations of X are low, Full-ENet has both higher sensitivity and specificity compared to Full-SPLS; when correlations of X are medium to high, Full-ENet has similar or a little lower sensitivity (within 12%), while the specificity is around 30% higher than those from Full-SPLS, indicating that the ENet method has better performances than the SPLS method for the variable selection and prediction in our simulations. (2) Based on complete cases analysis, both SPLS and ENet (CC-

SPLS and CC-ENet) methods have much higher PMSE and MSE than all other imputation based methods; the sensitivity for CC-ENet dropped 30%-50% compared to the Full-ENet, and the specificity for CC-SPLS are generally low. All these measurements indicate that CC-SPLS and CC-ENet are not recommended. (3) MI-SPLS has a high sensitivity but the specificity is at least 30% lower than MI-WENet, indicating that MI-SPLS would select more variables of those should not be selected. (4) In all the tested simulation scenarios, the MI-based weighted ENet (MI-WENet) method generally obtains the lowest PMSE and MSE among all competing imputation methods considered here with an exception in Table 2.1. That is, for the independent MCAR case when the correlations are following an AR(1) process, the PMSE and MSE for MI-WEnet are slightly larger than the other imputation based Enet method. The sensitivity and specificity of the MI-WENet is always close to the full-ENet model. Opposed to that, other imputed ENet models gain sensitivity with a significant loss in specificity compared to the full-ENet model. MI-WENet also maintains a reasonable sensitivity and specificity across all the simulation scenarios. This demonstrates that the MI-WENet method outperforms all the other methods.

Table 2.3 displays the results based on different combinations of p and n, under the specification of monotone MAR and AR(1) correlation structure. The first column in Table 2.3 shows the performance of different methods for fixed $p = 12$, when n increases, say $n = 50$, 100, and 200. The results demonstrate that: as n increases, (1) the PMSE and MSE for each method decreases, which means that the prediction becomes more accurate as n goes larger; (2) the sensitivity increases, indicating that as n increases, the percentage of correctly selected variables increases; (3) the specificity stays almost the same, indicating that the sample size does not impact the percentage of correctly rejected variables effectively; (4) among all the imputation methods, the MI-WENet method has the best perfor-

Table 2.1: Simulation results for missing complete at random (MCAR) scenarios with different missing patterns (independent and monotone) and different correlation structures (compound symmetry with low correlation, compound symmetry with medium correlation and first-order autoregressive (AR(1))) for $p = 12$ and $n = 50$.

	Independent MCAR				Monotone MCAR			
	PMSE	MSE	SENS	SPEC	PMSE	MSE	SENS	SPEC
	Low correlation: $\text{corr}(X_{i,j}, X_{i,j'}) = 0.1$							
Full-SPLS	14.72	3.52	97.2	67.0	15.05	3.96	95.7	70.2
CC-SPLS	28.84	15.34	80.7	47.7	22.08	10.23	84.5	52.0
SI-SPLS	17.72	6.20	91.7	60.3	18.64	7.21	86.8	70.0
MI-SPLS	16.10	4.68	98.3	36.5	16.59	5.24	98.7	42.5
Full-ENet	13.98	2.83	97.7	74.0	14.28	3.07	97.2	74.7
CC-ENet	30.27	18.51	52.5	87.5	21.99	10.47	68.8	89.3
SI-ENet	15.85	4.61	94.5	71.5	17.18	5.96	89.0	75.3
MI-ENet	15.43	4.22	96.3	71.7	16.38	5.16	94.8	72.2
SI-WENet	15.70	4.45	94.5	74.0	16.56	5.29	91.8	75.7
MI-WENet	15.34	4.14	96.0	74.8	15.90	4.66	94.0	72.8
	Medium correlation: $\text{corr}(X_{i,j}, X_{i,j'}) = 0.5$							
Full-SPLS	32.86	7.60	90.8	41.0	33.30	7.87	86.5	44.3
CC-SPLS	44.56	17.55	81.5	34.5	37.40	11.01	86.8	31.2
SI-SPLS	34.06	8.76	88.5	36.5	34.76	9.25	86.8	40.0
MI-SPLS	33.98	8.28	93.8	23.2	34.28	8.73	91.5	32.2
Full-ENet	30.63	5.67	87.3	71.7	30.99	5.84	86.0	71.5
CC-ENet	58.10	33.02	33.8	84.0	50.03	24.86	42.7	86.0
SI-ENet	32.45	7.42	82.5	66.2	33.18	8.10	77.0	69.0
MI-ENet	31.72	6.67	84.8	68.2	31.99	6.93	81.8	69.2
SI-WENet	32.00	6.95	83.5	69.7	32.73	7.60	79.3	70.8
MI-WENet	31.35	6.35	86.7	68.2	32.01	6.87	81.8	70.3
	AR(1) correlation: $\text{corr}(X_{i,j}, X_{i,j'}) = 0.8^{\lvert j-j' \rvert}$							
Full-SPLS	27.77	5.91	87.7	28.3	27.93	6.00	88.3	30.2
CC-SPLS	38.42	15.22	85.8	23.7	31.88	9.41	89.3	22.3
SI-SPLS	28.01	5.95	91.0	23.5	29.40	6.94	85.8	35.2
MI-SPLS	28.26	6.04	94.3	17.3	29.71	7.56	89.7	25.3
Full-ENet	27.14	5.39	78.5	65.8	26.94	4.99	80.0	66.7
CC-ENet	48.79	26.79	36.8	83.0	43.19	21.26	42.7	83.3
SI-ENet	27.33	5.61	76.8	65.7	28.54	6.64	71.3	65.3
MI-ENet	27.07	5.35	78.3	63.7	27.76	5.80	74.8	66.2
SI-WENet	27.20	5.50	77.7	63.8	28.03	6.07	74.8	64.2
MI-WENet	27.40	5.65	77.3	62.8	27.53	5.56	76.8	64.2

Table 2.2: Simulation results for missing at random (MAR) scenarios with different missing patterns (independent and monotone) and different correlation structures (compound symmetry with low correlation, compound symmetry with medium correlation and first-order autoregressive (AR(1))) for $p = 12$ and $n = 50$.

	Independent MAR				Monotone MAR			
	PMSE	MSE	SENS	SPEC	PMSE	MSE	SENS	SPEC
	Low correlation: $\text{corr}(X_{i,j}, X_{i,j'}) = 0.1$							
Full-SPLS	14.45	3.33	94.8	75.7	14.66	3.54	95.3	70.5
CC-SPLS	27.27	12.69	78.7	59.2	33.38	14.95	72.5	56.3
SI-SPLS	18.03	6.61	89.3	68.0	19.66	8.19	84.0	66.0
MI-SPLS	16.52	4.93	98.0	38.7	17.18	5.72	95.0	44.3
Full-ENet	13.83	2.71	96.8	78.8	13.93	2.88	97.2	75.2
CC-ENet	26.83	12.82	64.5	92.0	31.14	15.20	58.8	91.5
SI-ENet	16.90	5.73	90.7	77.2	18.25	7.07	85.2	72.8
MI-ENet	16.38	5.07	92.8	74.7	16.88	5.70	90.0	75.5
SI-WENet	16.57	5.35	91.3	78.3	17.20	5.96	88.7	73.5
MI-WENet	15.96	4.74	93.5	76.5	16.41	5.23	89.8	73.7
	Medium correlation: $\text{corr}(X_{i,j}, X_{i,j'}) = 0.5$							
Full-SPLS	32.28	7.24	90.5	34.5	32.77	7.67	89.0	38.0
CC-SPLS	40.92	16.02	81.3	40.2	49.39	19.42	79.8	38.2
SI-SPLS	35.15	9.57	80.7	45.0	35.49	9.47	81.3	41.5
MI-SPLS	34.69	8.83	91.5	30.8	34.88	8.69	90.7	34.3
Full-ENet	31.27	6.27	83.3	72.8	30.76	6.01	84.2	73.2
CC-ENet	53.89	27.42	47.5	91.3	58.18	28.58	43.8	88.8
SI-ENet	33.21	8.31	77.3	70.5	33.31	8.17	76.2	69.3
MI-ENet	32.72	7.79	78.2	69.0	32.22	7.20	77.5	68.2
SI-WENet	33.47	8.57	76.8	71.3	33.00	8.06	75.8	71.7
MI-WENet	32.63	7.66	78.8	69.0	31.94	6.99	78.2	67.8
	AR(1) correlation: $\text{corr}(X_{i,j}, X_{i,j'}) = 0.8^{\lvert j-j' \rvert}$							
Full-SPLS	27.36	5.66	89.7	30.2	27.45	5.82	87.5	31.0
CC-SPLS	33.11	11.68	81.0	36.3	43.08	15.44	75.8	35.8
SI-SPLS	30.46	8.28	81.3	40.3	30.33	7.31	81.3	42.0
MI-SPLS	30.96	8.46	91.0	23.5	30.02	7.15	89.2	27.3
Full-ENet	26.91	5.19	78.5	66.8	26.59	5.04	78.5	70.5
CC-ENet	45.50	22.77	41.8	84.5	53.26	28.26	38.0	87.2
SI-ENet	30.21	8.30	67.8	70.5	28.28	6.68	69.3	72.2
MI-ENet	28.76	6.93	72.8	69.2	27.76	6.05	72.7	70.5
SI-WENet	28.81	7.01	70.3	67.5	27.95	6.37	69.0	72.0
MI-WENet	27.97	6.34	73.8	68.8	27.22	5.63	73.5	69.8

mance in terms of smallest PMSE and MSE, and relatively high sensitivity and specificity compared to all the ENet based imputation methods. However, we observe a higher sensitivity with a significant loss of specificity in the SPLS based imputation methods. Although we see the reduced sensitivity in MI-ENet imputations for highly correlated data compared to many SPLS based imputations. The lower sensitivity is not as severe compared to the loss of specificity in the SPLS based imputations. The second column in Table 2.3 illustrates the performance of different methods when the number of predictor variables increases from 24 to 60 with fixed sample size n at 50, from which we conclude that: (1) as p increases (say $p = 24, 48, 60$), the PMSE and MSE for each method increase apparently; (2) as p gets larger, the sensitivity decreases, and the specificity slightly decreases as well for SPLS methods, while increase slightly for ENet methods; (3) in general, the performance of MI-WENet is as good as the Full-ENet.

To examine whether a large number of simulation runs impacts the simulation results, I carried out the simulations of the same scenarios as presented in Tables 3 but with 500 simulation runs. The results are presented in Table 2.4, from which we can see the results with 500 runs are very similar to those with 100 simulation runs (See Table 2.3).

Based on all simulation results, I conclude that the MI-WENet method obtains more or less the lowest PMSE and MSE among all the imputation based methods. The sensitivity and specificity of the MI-ENet method is better than all other ENet based imputation methods. In some cases although it looses in terms of sensitivity to some of the SPLS based imputation methods its loss in sensitivity is not as severe as the loss of specificity in some of the SPLS based imputations. Moreover, in most of our simulation scenarios, the PMSE, MSE, sensitivity and specificity from MI-WENet are closest to those from ENet on fully observed data. MI-WENet is therefore recommended for variable selection and prediction when

Table 2.3: Simulation results for scenarios with different combinations of p and n under monotone MAR and AR(1) correlation structure specifications based on 100 simulation runs.

	PMSE	MSE	SENS	SPEC	PMSE	MSE	SENS	SPEC
	$p = 12,\ n = 50$				$p = 24,\ n = 50$			
Full-SPLS	27.45	5.82	87.5	31.0	24.25	7.33	87.7	54.4
CC-SPLS	43.08	15.44	75.8	35.8	43.97	18.29	78.2	47.2
SI-SPLS	30.33	7.31	81.3	42.0	26.40	9.15	80.8	59.6
MI-SPLS	30.02	7.15	89.2	27.3	29.33	11.66	91.0	26.8
Full-ENet	26.59	5.04	78.5	70.5	22.81	5.89	82.5	82.7
CC-ENet	53.26	28.26	38.0	87.2	47.90	27.29	41.7	93.4
SI-ENet	28.28	6.68	69.3	72.2	24.17	7.24	75.0	83.2
MI-ENet	27.76	6.05	72.7	70.5	23.15	6.19	78.3	82.7
SI-WENet	27.95	6.37	69.0	72.0	23.78	6.85	78.3	82.7
MI-WENet	27.22	5.63	73.5	69.8	23.03	6.08	79.3	84.3
	$p = 12,\ n = 100$				$p = 48,\ n = 50$			
Full-SPLS	24.89	2.96	90.7	35.0	64.14	26.81	84.6	38.7
CC-SPLS	40.66	13.77	80.0	40.5	125.58	63.08	68.1	39.9
SI-SPLS	27.12	4.67	82.8	47.7	71.94	35.72	75.8	44.2
MI-SPLS	26.46	4.13	91.8	33.3	122.30	81.38	87.8	24.7
Full-ENet	24.35	2.44	89.5	68.2	62.51	26.75	55.6	87.1
CC-ENet	46.01	19.01	53.0	86.8	139.63	95.08	16.2	95.6
SI-ENet	26.13	4.06	79.5	70.7	79.43	43.50	45.1	86.8
MI-ENET	25.98	3.85	81.2	70.0	72.33	36.19	52.2	83.0
SI-WENet	25.61	3.71	80.5	69.8	77.47	41.56	47.2	86.9
MI-WENet	25.34	3.41	83.0	72.2	69.98	34.01	53.5	84.4
	$p = 12,\ n = 200$				$p = 60,\ n = 50$			
Full-SPLS	23.26	1.53	97.2	35.3	86.32	39.61	85.1	32.0
CC-SPLS	33.61	9.76	87.2	34.8	172.15	89.69	65.5	42.1
SI-SPLS	24.92	3.07	90.0	54.5	95.03	48.63	79.5	38.5
MI-SPLS	24.69	2.83	95.8	39.2	276.42	218.87	91.8	11.6
Full-ENet	22.95	1.20	97.3	71.2	91.66	46.41	48.9	86.7
CC-ENet	38.12	11.94	68.0	83.8	183.00	126.53	14.5	95.4
SI-ENet	24.63	2.76	89.5	75.2	105.60	60.05	40.9	87.2
MI-ENet	24.36	2.48	92.2	71.2	97.83	52.21	46.9	86.8
SI-WENet	23.84	2.06	92.0	72.5	103.51	58.41	41.7	87.9
MI-WENet	23.72	1.97	93.0	69.7	94.69	49.27	47.9	86.6

Table 2.4: Simulation results for scenarios with different combinations of p and n under monotone MAR and AR(1) correlation structure specifications based on 500 simulation runs.

	PMSE	MSE	SENS	SPEC	PMSE	MSE	SENS	SPEC
	$p = 12,\ n = 50$				$p = 24,\ n = 50$			
Full-SPLS	27.90	6.11	88.1	30.4	24.36	7.16	88.0	52.3
CC-SPLS	44.22	16.62	77.0	38.0	44.36	18.91	77.0	47.1
SI-SPLS	30.82	8.16	79.2	45.6	26.67	8.97	81.2	60.8
MI-SPLS	30.67	8.10	87.4	29.8	29.30	11.45	92.5	27.5
Full-ENet	27.11	5.34	78.2	68.6	22.95	5.93	80.9	82.7
CC-ENet	53.80	28.20	38.1	85.6	47.73	27.26	40.1	93.0
SI-ENet	29.49	7.63	68.4	73.5	24.56	7.53	74.2	82.2
MI-ENet	28.62	6.70	70.7	73.7	23.36	6.34	77.9	82.8
SI-WENet	28.85	7.04	69.8	72.5	24.09	7.03	76.6	82.3
MI-WENet	27.97	6.16	73.0	72.4	23.25	6.20	78.1	83.1
	$p = 12,\ n = 100$				$p = 48,\ n = 50$			
Full-SPLS	24.80	3.16	92.8	32.7	64.41	27.52	84.9	40.0
CC-SPLS	37.82	12.56	82.2	38.2	117.09	56.28	74.1	36.0
SI-SPLS	26.90	4.79	83.8	48.6	75.06	38.56	75.6	45.2
MI-SPLS	26.34	4.35	91.8	33.0	116.28	76.58	85.4	25.8
Full-ENet	24.18	2.52	90.3	67.2	64.41	28.48	57.2	85.2
CC-ENet	44.37	18.25	54.4	85.2	137.51	93.63	17.0	95.4
SI-ENet	26.15	4.38	79.4	70.2	78.09	42.30	45.5	84.7
MI-ENET	25.63	3.79	82.7	70.5	72.12	36.19	52.5	83.5
SI-WENet	25.21	3.53	83.3	68.9	76.92	40.97	47.1	85.1
MI-WENet	24.88	3.20	85.6	69.1	69.80	33.94	54.7	83.6
	$p = 12,\ n = 200$				$p = 60,\ n = 50$			
Full-SPLS	23.27	1.57	96.2	36.7	84.43	37.53	83.6	34.8
CC-SPLS	33.71	9.82	87.8	33.9	164.11	82.84	69.4	39.1
SI-SPLS	24.86	2.98	88.7	55.2	96.66	49.80	76.1	40.4
MI-SPLS	24.55	2.70	96.0	36.9	285.05	225.15	91.0	15.6
Full-ENet	23.02	1.27	97.2	70.0	88.60	43.01	50.6	86.3
CC-ENet	36.95	10.69	70.1	84.8	183.30	128.18	13.8	95.7
SI-ENet	24.49	2.64	88.2	71.1	104.67	59.05	40.6	87.2
MI-ENet	24.26	2.37	92.6	70.4	99.70	54.00	45.3	86.1
SI-WENet	23.76	2.00	91.8	71.5	103.13	57.48	41.4	87.2
MI-WENet	23.60	1.85	94.2	70.2	97.24	51.57	46.3	86.4

missing data exist.

In the following section, I applied the MI-WENet method to examine which variables were associated with the median effective dose and maximum effect in an ex-vivo phenylephrine-induced extension and acetylcholine-induced relaxation experiment.

2.4 Case Study

The high-fat diet and normal chow fed mouse model has been used to examine the mechanisms by which high-fat diet impacts cardiovascular function. Early on, high fat diet feeding induces endothelium inflammation, insulin resistance and endothelium dysfunction, which precedes the onset of diabetes (Kim et al., 2008). Thus, endothelium dysfunction, characterized by decreased nitric oxide (NO) production or bioavailability, is used as a robust and early indicator of cardiovascular injury (Rizzo et al., 2010). In the mouse model, mice were randomly assigned to high-fat diet and normal chow groups. The mice were fed for 12 weeks. Their body weight (BW), organ weight, blood variables and an array of plasma compositions and the ex-vivo endothelial functional outcomes were measured. Organ weights included heart, liver, kidney and spleen weight. The blood variables included percentage of red blood counts (%RBC, *i.e.*, hematocrit) and percentage of white blood counts (%buffy). The plasma parameters included the counts of cholesterol , triglyceride, albumin, total protein (TP), high density lipoprotein (HDL), low density lipoprotein (LDL), alanine aminotransferase, aspartate aminotransferase, creatine kinase, alkaline phosphatase, creatinine, hemoglobin A1c (HbA1c), insulin, and nitrogen oxide species (NO_x, *i.e.*, the sum of nitrite (NO_2) and nitrate (NO_3)), the ratio of HDL to LDL, and the percentage of albumin to total protein (Alb/TP). Isolated aorta were contracted with phenylephrine and relaxed with acetylcholine as previously published (Conklin

et al., 2009). Percentage relaxation based on maximal contraction was calculated for each aorta. The percentage of maximal relaxation is called the Emax, and the acetylcholine concentration needed to achieve 50% relaxation is called the effective concentration producing 50% response, *i.e.*, EC50. Emax and EC50 are two important parameters used to quantify endothelial function. In this section, we examined whether the two measurements of endothelial function, Emax and EC50, were related to any of the blood variables, plasma parameters, organ and body weights of the mice.

The final data set included 22 mice and 28 measured predictor variables. Some values in the predictor variables were missing due to inadequate volume of plasma. In total, 8 mice had missing observations. In order to include the 8 mice in the analysis, we applied the MI-WENet method to examine what variables were closely associated with the measurements of endothelial function. To apply the MI-WENet method, we imputed 5 realizations for each missing value, and stacked the five imputed data sets into one large data set. Each variable was scaled to have unit variance before multiple imputation, and there was no additional standardization carried out after imputation. Thus, the subjects without missing values remained the same in the stacked data set. The log-transformation for EC50 was applied to ensure the normality of residuals. I applied the MI-WENet method to the stacked multiple imputed data set to obtain the coefficient estimates and select the important predictor variables. In addition, I applied leave-one-out cross validated samples to construct 95% confidence intervals (95% CIs) for the estimated coefficients. The predictor variables whose 95% CIs did not contain zero were selected as the important variables for predicting the measurements of endothelial function. The estimates for the selected important predictors and their 95% CIs are shown in Table 2.5.

The selected important predictors for Emax were NO_x and the ratio of kidney

Table 2.5: The estimated coefficients and their 95% CIs based on leave-one-out samples for Emax and EC50 using the MI-based weighted ENet method.

Covariate	Estimate	95% CI-low	X95% CI-up
		Emax	
NO_x	0.1894	0.0792	0.2997
kidney/BW	0.2664	0.0128	0.5200
		EC50	
NO_x	-1.0803	-1.6983	-0.4623
kidney	-1.2246	-1.9112	-0.5379
kidney/BW	-1.7004	-2.6188	-0.7821
spleen	-1.5503	-2.3522	-0.7484
spleen/BW	-1.9629	-2.5398	-1.3860
heart/BW	-0.9037	-1.5152	-0.2923
TP	-1.0097	-1.4712	-0.5481
Alb/TP	-0.9950	-1.4623	-0.5276
HDL	-1.1533	-1.6846	-0.6220
LDL	-1.0116	-1.4736	-0.5497

to body weight. The selected important predictors for the log-transformed EC50 were NO_x, kidney weight, the ratio of kidney to body weight, spleen weight, the ratio of spleen to body weight, the ratio of heart to body weight, TP, Alb/TP, HDL and LDL. Endothelium dysfunction is commonly associated with decreased nitric oxide production and/or bioavailability (Hadi et al., 2005; Davignon and Ganz, 2004; Versari et al., 2009). The current results show that the decreased NO_x is associated with decreased Emax and increased EC50, which is consistent with previous findings. The other findings, such as association between endothelium dysfunction and kidney/BW, are also interesting and may be investigated further. The selected important predictor NO_x for Emax and EC50 demonstrates the selection precision of our proposed model, and thus, re-emphasizes the importance of using these endpoints to highlight the fundamental role of the endothelium in diet-induced cardiovascular injury.

2.5 Discussion and Conclusions

Missing data is common in animal experiments and clinical studies. In this project I concentrated on the cases with missing covariate values. One of the frequently used methods in practice is the complete case analysis which ignores the covariates with missing observations. This method is easy to carry out, while it is inefficient and sometimes incorrect because the missing observations may not be a random subset of the whole sample. In this chapter, I proposed a multiple imputation based weighted elastic net method (MI-WEnet) for variable selection and prediction. The simulation studies demonstrated that the proposed MI-WENet method was able to identify important predictor variables with similar precisions as the SPLS and elastic net methods would have achieved if the data were completely observed. Sensitivity and specificity obtained by the MI-WENet method were close to the results from the ENet method based on the full data in all the tested simulation scenarios. In addition, the MI-WENet method had the lowest MSE and PMSE among almost all methods we have evaluated. The simulations also showed that the use of SPLS and ENet on complete cases only resulted in models with poor sensitivity and much larger PMSE and MSE than MI-WENet, especially when proportion of missing data is high and the missing patterns are MAR. This again indicates that the use of MI-WENet is especially recommended when proportion of missing values of the covariates is moderate to high.

MI-WENet maintained a balanced sensitivity and specificity in all the simulation scenarios and all the imputation schemes. The MI-WENet is also easy to implement. By applying the cyclical coordinate descent algorithm (Friedman et al., 2010), the coefficients of MI-WENet can be easily estimated by iteratively minimizing the weighted penalized least squares. The computational cost is mainly affected by the number of predictor variables not the sample size. R code

for implementing the MI-WENet method can be obtained upon request. At last, it should be pointed out that the weights we proposed account for the available information in an observation; how to account for the available information more accurately is challenging and is beyond the scope of the current work.

TOPIC II: MONOTONIC SINGLE-INDEX MODELS WITH APPLICATION TO ASSESSING DRUG INTERACTION

3.1 Introduction

Single-index models have been extensively studied in the statistical literatures (Stoker, 1986; Härdle and Stoker, 1989; Ichimura, 1993; Yu and Ruppert, 2002). A single-index model could be considered as an extension of a general linear model. Recall that a general linear model is defined as $Y = X\alpha + \varepsilon$, where X is a vector of covariates $X = (X_0, X_1, ..., X_p)$, and $\alpha = (\alpha_0, \alpha_1, ..., \alpha_p)^T$ is an unknown parameter vector. A single-index model generalizes the general linear regression by replacing $X\alpha$ with a nonparametric function of $X\alpha$, say $f(X\alpha)$, where f is an unknown univariate link function, and α remains the same. In the literature, f is usually estimated by using kernel spline (Ichimura, 1993; Härdle and Stoker, 1989; Xia and Härdle, 2006), or using penalized spline (Yu and Ruppert, 2002). Both α and f are unknown and need to be estimated. For identification, one may either restrict $\alpha_0 = 1$ or add constraints on α such that $\| \alpha \| = \sqrt{\alpha_0^2 + \alpha_1^2 + ... + \alpha_p^2} = 1$ and $\alpha_0 > 0$. The asymptotic properties for α have been established (Stoker, 1986; Härdle and Stoker, 1989; Ichimura, 1993; Yu and Ruppert, 2002). A single-index model reduces the dimensionality from multivariate predictors to a univariate index z (say $z = X\alpha$), while it still captures important features in high-dimensional data (Yu and Ruppert, 2002). Any interactions between the covariates can also be included in the single index z. Single-index model has wide application in econometrics (Stoker, 1986; Härdle and Stoker, 1989; Ichimura, 1993) as well as in biometrics (Yu and Ruppert, 2002).

Although single-index models have been well studied, the function f in the

single-index models is not necessary monotonic. In certain applications, it is desirable to have the function f monotonic. For example, in the dose-response studies, the dose response relationship is often assumed to be monotonic (Kong and Eubank, 2006; Ramsay and Barahamowicz, 1989; Ramsay, 1998). In the combination drug studies, the dose response relationship is often described by a response surface model (Greco et al., 1995). It is desirable that the response surface model is reduced to a dose-response curve when only one drug is applied. Without loss of generality, let denote the response at the combination dose (d_1, d_2) is y. A relative potency, say ρ, is often used to describe how effective drug 2 is relative to drug 1, that is the effect of drug 2 at dose level d_2 when applied alone, is the same as that of drug 1 at dose level ρd_2. If we assume the dose-response curve for drug 1 is $y = f(d_1)$, then the dose-response curve for drug 2 is $f(\rho d_2)$. In case that the two drugs do not have any interaction, i.e. the combination is additive, the effect of the combination dose (d_1, d_2) can be described by $f(d_1 + \rho d_2)$. If the effect at (d_1, d_2) is more (or less) than the effect of drug 1 at dose level $d_1 + \rho d_2$, we say the combination dose (d_1, d_2) is synergetic (or antagonistic) (Lee et al., 2007; Berenbaum, 1989). Plummer and Short (1990) used a model of the form $f(d_1 + \rho d_2 + k\sqrt{d_1 d_2})$ identify and quantify departures from additivity, where $k > 0$ indicates synergy of the combination dose at (d_1, d_2), and $k < 0$ indicates antagonism of the combination dose at (d_1, d_2). Kong and Lee (2006) extended the model of Plummer and Short (1990) by replacing k with a quadratic function of (d_1, d_2) so that the model has the flexibility to capture different patterns of drug interaction, i.e. some combination dose may be synergistic and some may be antagonistic (Savelev et al., 2003). These models can be rewritten as the form of $f(x^T\alpha)$. For example, one may set $x = (d_1, d_2, \sqrt{d_1 d_2})$ and $\alpha = (1, \rho, \kappa)$ for Plummer and Short (1990)'s model.

Note that in the approaches by Plummer and Short (1990) and the extension by

Kong and Lee (2006), the function is assumed to be monotonic parametric model. The estimates of α can be biased when the parametric function f is misspecified. To avoid the problem caused by the misspecification of the function f, we do not assume any specific function form for f. Instead, we only assume that f is monotonic and has continuous first and second derivatives. The function f is estimated by using penalized splines with I-splines as its basis functions (Ramsay, 1998). The monotonicity of the function f is achieved by adding constraints to the coefficients of I-splines basis functions (Kong and Eubank, 2006; Ramsay and Barahamowicz, 1989; Ramsay, 1998). The presentation of this topic is organized as follows. In Section 3.2, I propose the single-index model for assessing drug interactions, and develop algorithm for estimating the monotonic function f and the parameter α in the single-index model. Simulation studies are carried out in Section 3.3 to examine the performance of the proposed model in term of accuracy in estimating f and α. In Section 3.4, I apply the proposed monotonic single-index method to examine the drug interaction of two drugs in a case study given by Harbron (2010). The last section is devoted to a discussion.

3.2 Monotonic Single-index Model

3.2.1 Monotonic Single-index Model

Let y_i denote the response observed at the combination dose (d_{1i}, d_{2i}) $(i = 1, ..., n)$. In case that drug 1 is applied alone, d_{2i} is set to zero. Similarly, d_{1i} is set to zero if drug 2 is applied alone. In the literature, the dose-response curve is usually described by a parametric function, which may not be specified correctly. In the project, we develop dose-response surface model under the minimal assumption that the dose-response curve is monotonic. Similar to Kong and Lee (2008), a

quadratic function of (d_1, d_2) of the form

$$g(d_1, d_2; \kappa) = \kappa_0 + \kappa_1\sqrt{d_1} + \kappa_2\sqrt{d_2} + \kappa_3 d_1 + \kappa_4 d_2 + \kappa_5\sqrt{d_1 d_2}, \tag{3.1}$$

is used to capture different patterns of drug interaction. I propose the following response surface model to assess different patterns of drug interactions:

$$y = f\left(d_1 + \rho d_2 + g(d_1, d_2; \kappa)\sqrt{d_1 d_2}\right) + \varepsilon, \tag{3.2}$$

where f is monotonic function and is estimated by cubic splines, and ε is a random error with mean zero and variance σ^2.

Denote

$$\begin{aligned} \alpha^T &= (1, \rho, \kappa_0, \kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5) \\ &= (\alpha_0, \alpha_1, ..., \alpha_7), \end{aligned} \tag{3.3}$$

$$x^T = \left(d_1, d_2, \sqrt{d_1 d_2}, \sqrt{d_1} d_2, d_1\sqrt{d_2}, \sqrt{d_1^3 d_2}, \sqrt{d_1 d_2^3}, d_1 d_2\right).$$

Model (3.1) and (3.2) can be expressed as single-index models of the form

$$y_i = f\left(x_i^T \alpha\right) + \varepsilon_i, \quad i = 1, ..., n \tag{3.4}$$

subject to $\alpha_0 = 1$ and f is monotonic.

Let us denote $z_i = x_i^T \alpha$ $(i = 1, ..., n)$, z_i is often called single index for $x_i^T = (x_{i0}, ..., x_{ip})$. Let us also denote a knot sequence $\tau = \{\tau_j\}_{j=1}^K$, where $\min\{z_i, i = 1, ..., n\} = L = \tau_1 < \tau_2 < ... < \tau_K = U = \max\{z_i, i = 1, ..., n\}$. The function f is defined on the domain $[L, U]$ over which f is approximated by a piecewise polynomial function in each interval $[\tau_l, \tau_{l+1}]$, and the two polynomials in the two adjacent intervals, say $[\tau_{l-1}, \tau_l]$ and $[\tau_l, \tau_{l+1}]$, are required to join smoothly. The

most commonly used spline is the cubic spline in which each polynomial is a cubic, and the piece-wise polynomials are joined at each knot with continuous first and second derivatives so that the curve changes smoothly. There are different basis functions for cubic splines, such as B-splines, M-spines, and I-splines (Ramsay, 1998). I found I-splines are convenient for constructing monotonic curves (Ramsay, 1998).

In general, the I-spline basis function of degree k is determined by $k+1$ knots, say $\{\tau_j, \tau_{j+1}, ...\tau_{j+k}\}$, and can be expressed as

$$I_j^k(z) = \int_L^z M_j^k(u)du, \tag{3.5}$$

with $z \in [L, U]$ and $j = 1, 2, ..., K-k$. Here the M-splines can be iteratively obtained by the following formula:

$$M_j^k(z) = \frac{k\left[(z-\tau_j)M_j^{k-1}(z) + (\tau_{j+k}-z)M_{j+1}^{k-1}(z)\right]}{(k-1)(\tau_{j+k}-\tau_j)}, \tag{3.6}$$

with $k > 1$ and

$$M_j^1(z) = \begin{cases} \frac{1}{\tau_{j+1}-\tau_j}, & \tau_j \le z < \tau_{j+1}; \\ 0, & \text{otherwise.} \end{cases} \tag{3.7}$$

Because the M-spline basis $M_j^k(z)$ is a piecewise nonnegative polynomial of degree $k-1$, the corresponding I-spline basis function $I_j^k(z)$ is therefore a piecewise monotone polynomial of degree k (Ramsay, 1988). The I-spine function $I_j^k(z)$ can also be put in a more convenient form

$$I_j^k(z) = \begin{cases} 0, & j > l \\ \sum_{m=j}^{l} \frac{\tau_{m+k+1}-\tau_m}{k+1} M_m^{k+1}(z), & l-k+1 \le j \le l \\ 1, & j < l-k+1, \end{cases} .$$

for any $z \in [\tau_l, \tau_{l+1}]$. In the rest of this presentation, I set k=3. The monotonic function f is approximated by a linear combination of the I-splines in the following form

$$f(z) = \sum_{j=1}^{K-3} \beta_j I_j^3(z), \tag{3.8}$$

subject to $\beta_j \geq 0$ for $j = 1, ..., K-3$. The parameters $\beta^T = (\beta_1, ..., \beta_{K-3})$ in equation (3.8) is obtained by minimizing the following penalized residuals sum of squares (PRSS),

$$\text{PRSS} = \sum_{i=1}^{n} \{y_i - f(z_i)\}^2 + \lambda \int_L^U f''(u)^2 du, \tag{3.9}$$

subject to $\beta_j \geq 0$ $(j = 1, ..., K-3)$. Note that

$$f'(z) = \sum_{j=1}^{K-3} \beta_j M_j^3(z). \tag{3.10}$$

and

$$f''(z) = \sum_{j=1}^{K-3} \beta_j (M_j^3(z))'. \tag{3.11}$$

The second term in (3.9), $\int_L^U f''(u)^2 du$, can be written in the form of $\beta^T D \beta$, where D is a $(K-3) \times (K-3)$ tri-diagonal matrix, and the (j, j') entry is

$$\int_L^U \left[M_j^3(u)\right]' \left[M_{j'}^3(u)\right]' du,$$

which is specified in Appendix A. Thus, given the indices $z_i = x_i^T \alpha$ $(i = 1, ..., n)$ and knots sequence, one can obtain the estimate of f from minimizing the PRSS in equation (3.9).

To obtain the estimate for α, for fixed function f, one may minimize the following residual sum of squares for errors

$$\text{RSSE}(\alpha) = \sum_{i=1}^{n} \left\{y_i - f(x_i^T \alpha)\right\}^2. \tag{3.12}$$

Consider the first order Taylor expansion for $f(x_i^T\alpha)$ at $\alpha^{(0)}$:

$$\begin{aligned} f(x_i^T\alpha) &\approx f(x_i^T\alpha^{(0)}) + f'(x_i^T\alpha^{(0)})x_i^T(\alpha - \alpha^{(0)}) \\ &= f(x_i^T\alpha^{(0)}) - f'(x_i^T\alpha^{(0)})x_i^T\alpha^{(0)} + f'(x_i^T\alpha^{(0)})x_i^T\alpha, \end{aligned}$$

one can update α by minimizing

$$\text{RSSE}^*(\alpha) = \sum_{i=1}^{n} \left\{ y_i^* - x_i^{*T}\alpha \right\}^2, \tag{3.13}$$

where $x_i^* = f'(x_i^T\alpha^{(0)})x_i$ and $y_i^* = y_i - f(x_i^T\alpha^{(0)}) + f'(x_i^T\alpha^{(0)})x_i^T\alpha^{(0)}$. By replacing $\alpha^{(0)}$ by the current available α and using the first order Taylor expansion, one can minimize the RSSE for α. It should be noticed that the estimates for α and f are obtained iteratively. That is, given α, one estimates f; and given f, one estimates α. The iteration continues until the estimates of α from two adjacent estimations of f are close enough.

3.2.2 Algorithm for Estimating f and α

The algorithm for estimating α and f can be described as the follows:

Step 1. Set a grid of 100 values for λ ($\lambda > 0$) by equally spacing λ in log-scale. For each fixed λ (given α) obtain the monotonic link function by minimizing the PRSS in (3.9). Set the optional link function as the one which has the minimum generalized cross validation (GCV) score, which has the following form

$$GCV(\lambda) = \frac{n^{-1}\sum_{i=1}^{n}\{y_i - \sum_{j=1}^{K-3}\hat{\beta}_j I_j(x_i^T\hat{\alpha})\}^2}{\{1 - n^{-1}tr(S_\lambda)\}^2},$$

where $tr(S_\lambda) = tr\{I(I^TI + n\lambda D)^{-1}I^T\}$, and $I = (I_1, ..., I_{K-3})$ is the matrix of I-spline basis on the estimated single-index $\hat{z}_i = x_i^T\hat{\alpha}$, for $i = 1, ..., n$.

Step 2: Given the estimate of f, α is obtained by minimizing the RSSE in (3.13).

Step 3: Repeat step 1 and step 2 until the estimates of α from two repeats are converged within a small tolerance error.

Remark: In nonparametric regression, the trace of smoothing matrix S_λ is often called the degrees of freedom of the fit (Hastie and Tibshirani, 1990), i.e., $df_{fit} = tr(S_\lambda)$. It has the rough interpretation as the equivalent number of parameters as defined in linear regression. The residual degrees of freedom is defined as $df_{res} = n - 2tr(S_\lambda) + tr(S_\lambda S_\lambda^T)$. Consequently, the error variance σ^2 in model (3.4) can be estimated by $\hat{\sigma}^2 = RSSE/df_{res}$, where $RSSE = \sum_{i=1}^{n}\{y_i - \hat{f}(x_i^T\hat{\alpha})\}^2$.

3.2.3 Variance Estimate for $\hat{\alpha}$ and Assessing Drug Interaction

The asymptotic properties for $\hat{\alpha}$ in single-index model have been studied by Hardle et al (2000) and Xia et al. (2006) using kernal estimation for f and by Yu and Ruppert (2002) using penalized splines. Under the assumption that $\hat{f}$ is an unbiased estimate for f, $\hat{\alpha}$ has the following asymptotic property:

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{D} N(0, V^{-1}\Sigma V^{-T}), \tag{3.14}$$

where

$$V = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial f(x_i^T\alpha)}{\partial \alpha}\frac{\partial f(x_i^T\alpha)}{\partial \alpha^T} = \frac{1}{n}\sum_{i=1}^{n}(f'(z_i))^2 x_i x_i^T, \tag{3.15}$$

and

$$\begin{aligned}\Sigma &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial f(x_i^T\alpha)}{\partial \alpha}(y_i - f(z_i))^2\frac{\partial f(x_i^T\alpha)}{\partial \alpha^T} \\ &= \frac{1}{n}\sum_{i=1}^{n}(f'(z_i))^2(y_i - f(z_i))^2 x_i x_i^T.\end{aligned} \tag{3.16}$$

Therefore, the variance of $\hat{\alpha}$ can be estimated by

$$\widehat{Var}(\hat{\alpha}) = \frac{1}{n}\hat{V}^{-1}\hat{\Sigma}\hat{V}^{-T}, \tag{3.17}$$

where $\hat{V}$ and $\hat{\Sigma}$ are obtained by replacing every quantity in (3.15) and (3.16) by their final estimates, respectively. Note that the model is developed to assess drug interaction. It is important to estimate the quantity of the function $g(d_1, d_2; \alpha)$ in (3.1) and estimate its variation so that the inference for drug interaction can be made with statistic rigor. Indeed, the function $g(d_1, d_2; \alpha)$ can be written as a linear function of α, say $u^T\alpha$, where $u^T = \left(1, \sqrt{d_1}, \sqrt{d_2}, d_1, d_2, \sqrt{d_1 d_2}\right)$. Once the estimate $\hat{\alpha}$ and its variance become available, the variance for $g(d_1, d_2; \alpha)$ can be obtained as $u^T Var(\hat{\alpha})u$. Thus the 95% confidence interval (CI) for $g(d_1, d_2; \hat{\alpha})$ can be constructed as $g(d_1, d_2; \hat{\alpha}) \pm z_{\alpha/2}\sqrt{u^T Var(\hat{\alpha})u}$. An alternative approach for the variances of $g(d_1, d_2; \hat{\alpha})$ can be obtained by using bootstrap method that is shown in Appendix B.

3.3 Simulation

Extended simulation studies were performed to examine the finite-sample properties of the estimates of the proposed model, with a set of parameters $\alpha = (1, 1, 0, 0, 0, 0.5, -0.5, 0)$ was considered. The corresponding response surface model was defined as

$$\begin{aligned} Y &= \log\left(\frac{E}{1-E}\right) \\ &= \log\left(d_1 + d_2 + (0.5d_1 - 0.5d_2)(d_1 d_2)^{\frac{1}{2}}\right) + \varepsilon, \end{aligned} \tag{3.18}$$

where $\varepsilon \sim N(0, \sigma^2)$. I generated 100 random samples with $\sigma = 0.01$, and d_1 and d_2 taking values among $(0, 0.1, 0.5, 1, 2, 4)$. The replicates for each sample was set as 3. Then the total sample size in each simulation run was $3 \times 6 \times 6 = 108$. I fitted each random sample to the full model as specified in Equations (3.1) and (3.2), and obtained the estimated parameters and their corresponding standard errors (SE). The averages of the estimated parameters (Est.), averaged SE (SE.Ave), and

the standard error of the estimates from the 100 estimates of α (SE.Emp) were reported in Table 3.1.

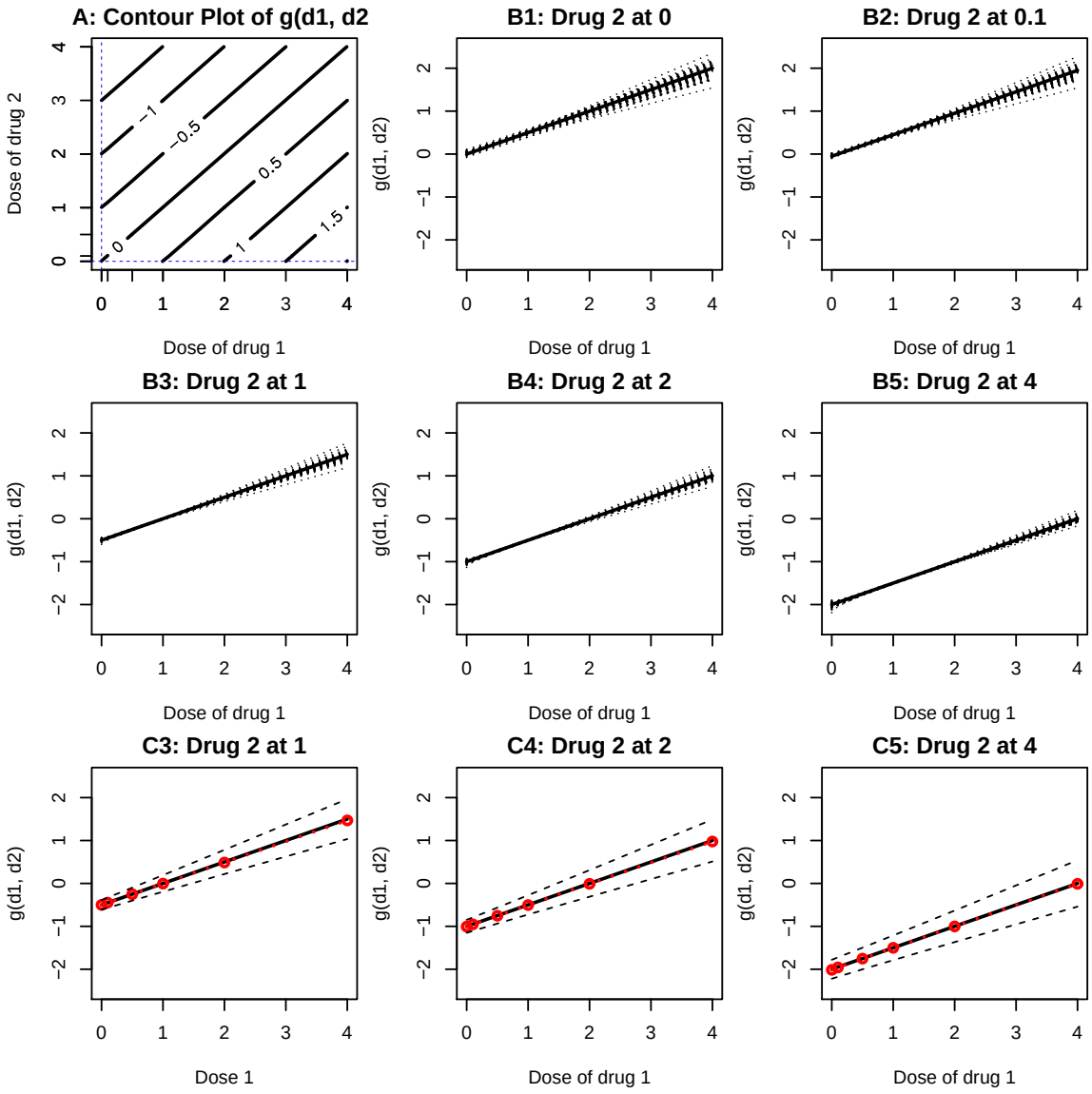
Table 3.1: Simulation results from fitting the full model based on the 100 simulated data.

$g(d_1, d_2; \alpha) = 0.5d_1 - 0.5d_2$				
Parameters	True value	Est.	SE.Ave	SE.Emp
α_0	1.0	1.000	0.001	0.000
α_1	1.0	0.999	0.004	0.004
α_2	0.0	0.022	0.031	0.034
α_3	0.0	-0.012	0.038	0.057
α_4	0.0	-0.026	0.028	0.044
α_5	0.5	0.489	0.013	0.048
α_6	-0.5	-0.496	0.010	0.016
α_7	0.0	0.018	0.012	0.029

Figure 3.1 showed the contour plot of the underlying polynomial function $g(d_1, d_2)$ in panel A. We can see that some combination doses are synergistic (i.e., $g(d_1, d_2) > 0$), and some combinations doses are antagonistic (i.e., $g(d_1, d_2) < 0$). The underlying polynomial function $g(d_1, d_2)$ and the fitted polynomial function for the 100 simulations are plotted in panels B1-B5 of Figure 3.1, where the x-axis is the dose level for drug 1 under each fixed drug 2 dose level. The underlying function $g(d_1, d_2)$ and the 95% limit bounds based on the 100 simulation data are shown in panels C3-C5 of Figure 3.1.

Based on the simulation results shown in Table 3.1 and Figure 3.1, I conclude that (1) the estimates of the parameters in the single-index model were unbiased (Table 3.1); (2) the functions $g(d_1, d_2)$ ware estimated correctly (panels B1-B5 and C3-C5 of Figure 3.1); (3) the empirical standard errors (Table 3.1 under column "SE.Emp") were close to the formula based SE (Table 3.1 under column "SE.Ave"), indicating the variance estimates were reasonable.

Figure 3.1: The results from simulation studies. Panel A shows the contour plot of the underlying polynomial function $g(d_1, d_2)$. Each of the panels B1-B5 shows the underlying curve (solid lines) and the fitted curves (dotted lines) based on 100 simulation runs. Panels of C3-C5 present the underlying curve (solid lines), the mean of fitted curve (dotted line), and the 95% point-wise confidence bounds (dashed lines).



3.4 Case Study

In this section, I studied a real case data resulting from an in vitro combination experiment (Harbron, 2010). In the data set, there were two compounds A and B, both dosed in monotherapy along a 9 dose levels with 3 replicates in each dose level. The two compounds were combined in a factorial design manner for all of the 6 lower doses (Table 3.2). All design points were tested in triplicate. The

Table 3.2: Experimental results (per cent inhibition) from a combination study.

		Dose of compound B								
Dose of compound A	0	0.037	0.11	0.33	1	3	9	27	81	243
		3.1	1.0	1.0	8.5	13.3	23.7	53.1	78.9	93.5
0		1.5	1.0	8.8	1.0	14.7	30.2	59.0	82.9	98.8
		1.0	1.0	5.9	4.5	18.1	42.5	62.0	81.5	96.2
	1.0	1.0	1.0	1.0	8.4	21.8	38.5			
0.037	5.8	2.0	1.0	2.9	10.0	4.7	34.9			
	1.0	1.0	1.0	4.2	7.6	9.5	35.2			
	1.0	1.0	2.6	1.0	5.4	22.2	32.8			
0.11	1.0	1.0	1.0	2.5	9.8	22.5	34.8			
	1.0	1.0	9.2	2.0	8.9	15.6	30.4			
	1.0	1.0	1.0	4.7	8.5	22.5	37.9			
0.33	4.2	6.2	4.9	6.3	12.3	19.8	41.7			
	13.3	6.1	9.5	5.6	7.2	15.9	34.3			
	1.9	16.0	3.4	21.2	22.9	34.0	52.9			
1	4.2	6.0	6.6	19.6	23.4	37.7	46.4			
	5.7	15.8	15.5	14.7	26.4	42.1	53.9			
	20.6	41.1	49.4	43.0	50.5	55.8	66.8			
3	31.7	42.1	50.4	48.3	40.0	56.6	59.2			
	23.9	43.1	51.3	46.1	52.5	61.8	64.2			
	56.2	69.2	66.8	76.8	84.7	75.6	77.5			
9	58.5	82.1	83.5	83.4	79.3	68.6	77.6			
	66.6	71.1	72.8	83.1	84.0	85.5	79.8			
	89.4									
27	84.9									
	85.8									
	92.9									
81	97.6									
	90.9									
	99.0									
243	93.7									
	99.0									

percentage of growth inhibition of a cell culture was measured as the endpoint. It is calculated from an optical density, corrected for background and normalized

Table 3.3: The estimates of α and their standard error estimates based on model formula and bootstrap method.

	Direct estimates				Bootstrap estimates		
Par.	Est.	SE	CI.low	CI.up	SE	CI.low	CI.up
α_0	1.00	0.00	1.00	1.00	0.00	1.00	1.00
α_1	0.34	0.14	0.07	0.62	0.03	0.29	0.40
α_2	-0.20	0.66	-1.49	1.08	0.61	-1.25	1.16
α_3	1.62	0.54	0.57	2.68	0.51	0.61	2.61
α_4	0.26	0.86	-1.43	1.95	0.48	-0.76	1.13
α_5	-0.03	0.14	-0.30	0.24	0.15	-0.36	0.24
α_6	-0.06	0.30	-0.64	0.52	0.10	-0.25	0.16
α_7	-0.41	0.17	-0.74	-0.07	0.13	-0.65	-0.14

against an average no treatment response (Harbron, 2010). I set the grid of penalty parameter as $\lambda \in [10^{-10}, 10^{10}]$ by equally spacing $\log(\lambda)$ on $[-10, 10]$.

The plot of GCV versus λ at final step is shown in Figure 3.2. The optimal λ was $\lambda = 0.012$ after 19 iterations. The estimates of α and their 95% confidence intervals based on model-based variances and bootstrap variances were presented in Table 3.3, and both gave similar results.

Figure 3.2: The plots of penalty parameter λ vs GCV. The top plot is for $\log(\lambda) \in [-10, 10]$, and the bottom plot is for $\log(\lambda) \in [-10, 0]$. The minimum GCV corresponds to $\lambda = 0.012$.

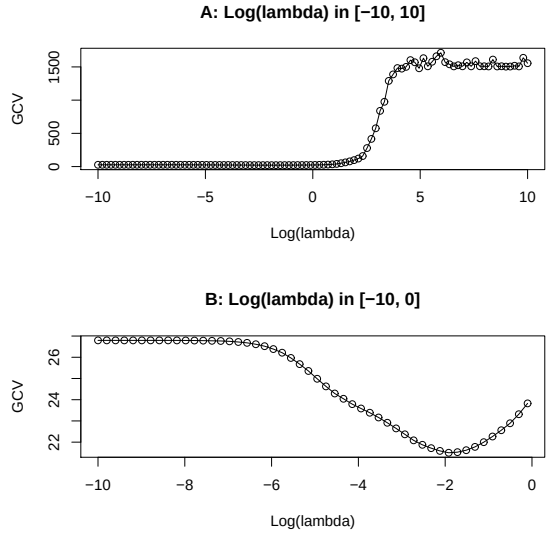
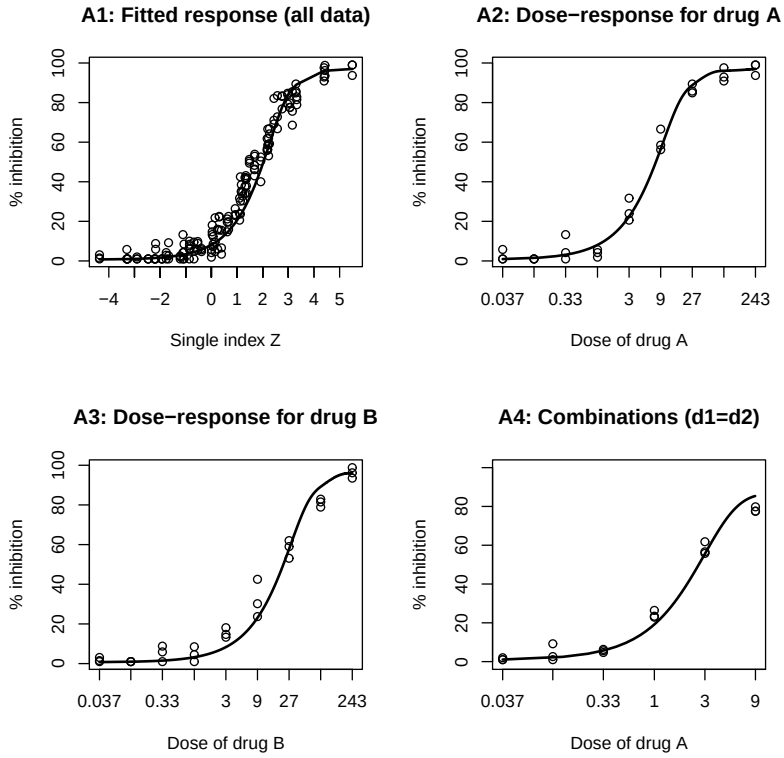


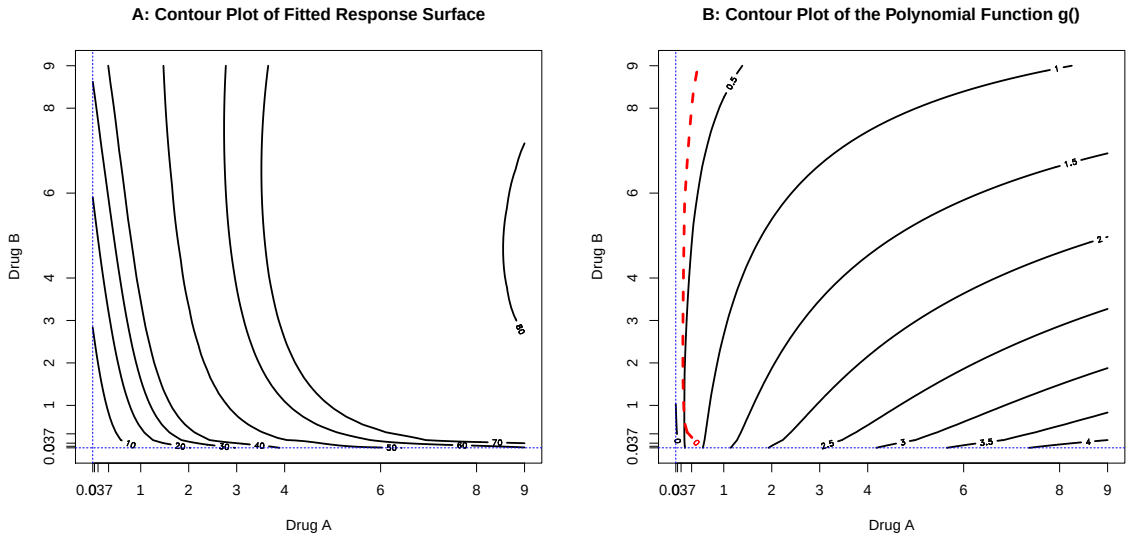
Figure 3.3 showed the fitted $\hat{f}(x^T\hat{\alpha})$ versus the indices $x^T\hat{\alpha}$ as the solid lines and the observed responses as circles (panel A1), the fitted function $\hat{f}$ when only drug A was applied (panel A2), when only drug B was applied (panel A3), and when drug A and drug B were combined with equal amount (panel A4). To facilitate viewing, Figure 3.3 is shown on a logged scale, though the analysis was performed on the unlogged dose scale. From Figure 3.3, it is clear that the model fitted the data very well for all data (panel A1), marginal dose-response (panels A2 and A3), and a typical combinations (panel A4).

Figure 3.3: Fitted response versus the estimated single-index (panel A1), fitted responses versus dose of drug A when drug A was applied alone (panel A2), dose of drug B when drug B was applied alone (panel A3), and dose of drug A when drug A was combined with drug B in equal amount (panel A4).



In Figure 3.4, panel A showed the contour plot of the fitted response surface, and panel B showed the contour plot of the polynomial function $g(d_1, d_2; \kappa)$, where the dotted curve is the intercept of lower 95% confidence surface with

Figure 3.4: Contour plot of response surface of the combination of compounds A and B (panel A), and contour plot of polynomial function $g(d_1, d_2; \hat{\kappa})$ (panel B).



$\hat{g}(d_1, d_2; \kappa) = 0$. From panel B of Figure 3.4, I concluded that the combination doses for all but left side of the dashed lines were synergistic, the combination dose with low level of drug A (left side of the dashed lines) were additive. The results were consistent with the findings of Harbron (2010).

3.5 Discussion and Conclusions

Based on simulation and case study, I conclude that the proposed monotonic single-index modeling approach worked effectively in assessing the interactions of two drugs. By using I-spline basis, we can easily construct a monotonic link function f in the single-index model to describe the dose-response relationship. The estimates of both f and α are unbiased. The polynomial function $g(d_1, d_2; \kappa)$ is estimated accurately to capture different patterns of drug interaction. The approach can be extended to assess drug interaction of multiple drugs. The algorithm was implemented in R and will be made public available to facilitate the use of this method.

According to (Ramsay, 1998), the primary distinction between smoothing splines and regression splines is how to place the knots. The knots selection technique for classical smoothing splines is to select all unique values of z. The knots for regression penalized splines are usually selected from a smaller set of candidates. The more knots on the interested region of z, the more flexible the spline is. This principle applies locally, such that if we need a lot of flexibility in a particular region of z, we use more knots in this region, and we may use less knots in the region if we don't need much curvature. However, to decide how many knots needed and where to position the knots is often challenging. Practically, user first simply make knots equally spaced, while paying attention to the requirement of having at least one observation in every subinterval. The optimal number of knots K should be sufficiently large to fit the data, meanwhile K can not be so large that computation time is excessive or the estimated curve f is over-fitted. Unfortunately, there is no standard rules to decide the number of K. Penalized splines tend to select more knots but add penalty to make fitted curve more smooth. Ruppert (2002) has made a detailed study of the choice for K for penalized splines. However, the algorithm developed by Ruppert (2002) may either stop prematurely or there are multiple local minimums of GCV(λ) at the sequence of K. Therefore, the selection of K seems challenging.

In this chapter, I applied a penalty approach that is similar to smoothing splines but with a little fewer knots. In order to fit data to capture as enough features as possible while to save computation burden, I proposed a new knots selection approach. Use all the unique values of z as knot candidates, and include $min(z)$ and $max(z)$ into the knot sequence first. By setting up a jump width w, say $w = 0.1$, suppose z_j is an element selected in the knot sequence, if $z_{j+1} - z_j > w$, we maintain z_{j+1} in the knot sequence; otherwise we drop z_{j+1} and examine whether $z_{j+2} - z_j > w$. A subset of indexes z is then formed as knots with the

distance of any two knots being larger than w. By using this approach, we avoid the dense knot values and include important z values in the knot sequence. The simulation and case study showed this approach provided efficient way for knots determination.

TOPIC III: A STUDY FOR PREDICTING PATIENT SURVIVAL TIME WITH HIGH THROUGHPUT MASS SPECTROMETRY DATA

4.1 Introduction

Genomic and proteomic technologies have become more and more important in biomedical studies in a recent time. The use of mass spectrometry as a diagnostic tool and identification of proteomic biological markers has risen extensively and demonstrated great advantage that led to the discovery of numerous proteins and protein profiles associated with various types of diseases (Stoeckli et al., 2001; Adam et al., 2002; Aebersold and Mann, 2003; Rai and Chan, 2004; Datta et al., 2008; Datta and Pihur, 2010). In this chapter, I aim to develop an effective model for predicting the survival time of cancer patients via penalized linear regression modeling on log-transformed failure times by the proteomic features as obtained from the matrix-assisted laser desorption/ionization Time-of-Flight mass spectrometry (MALDI-TOF-MS) data.

A typical MALDI-TOF-MS data set contains hundreds of spectra, and each spectrum contains tens of thousands of intensity measurements representing an unknown number of protein/peptide peaks which are the key features of interest (Hardesty et al., 2011). The data of a single spectrum is usually given in two columns, with the first column containing the mass-to-charge ratio (m/z) and the second column containing the corresponding intensity. Before applying this data directly to the final modeling analysis, it is also important to conduct basic preprocessing analyses such as baseline correction, denoising, alignment and peak detection to identify key interested features (Satten et al., 2004; Jeffries, 2005; Renard et al., 2008; Atlas and Datta, 2009; Ndukum et al., 2011). Although these

basic preprocessing steps may generally extract some peaks for further interest, there are still hundreds or thousands of retaining potentially important features which could be used for the subsequent predictive modeling analysis. On the other hand, the preprocessing procedures should be performed well and carefully for assuring the precision of feature extraction and quantification, as subsequent analysis depends on these determinations. It has been shown that the use of inadequate or ineffective methods in the preprocessing steps make it difficult to extract meaningful biological information from these data (Sorace and Zhan 2003; Baggerlt et al. 2003, 2004; Yasui et al. 2003a). In this chapter, I have also studied how to carry out these preprocessing procedures properly and efficiently for the prediction of patient survival times. Moreover, since the high-dimensionality of the feature set as well as some high correlations among features, in order to predict patient survival using a predictive statistical model, one needs to consider dimension reduction and important feature selection in addition to the basic pre-processing of mass spectrometry (MS) data very carefully.

A number of early attempts, mostly in the genomic data setting, use some ad hoc dimension reduction methods and incorporate the reduced set of covariates (e.g., principal components, meta-genes etc.) in a Cox proportional hazards regression model (Pawitan et al., 2004; Bovelstad et al., 2007). In proteomic studies, dimensionality of the feature set (covariates) is typically even larger comparing with gene expression data. Recently, penalized regression versions of the Cox model (Cox, 1972) have been attempted to deal with high dimensional data (Li and Luan, 2003; Gui and Li, 2005). However, the proportional hazards assumption of a Cox model itself may be too simplistic for genomic and proteomic applications. On the other hand, semi-parametric accelerated failure time (AFT) model with an unspecified error distribution is often regarded as a more flexible alternative to the Cox model in survival analysis. As far as I know, there are only a

few publications about using the AFT model in high dimensional data setting, which mostly use the microarray platforms. For example, the LASSO (Tibshirani, 1996) and the threshold-gradient-directed regularization along with AFT model are applied for estimation and variable selection by Huang et al. (2006); predicting survival times using AFT model along with PLS and LASSO is studied by Datta et al. (2007); the elastic net approach for variable selection under both the Cox proportional hazards model and the AFT model is adopted by Engler and Li (2009); Mostajabi et al. (2012) compared the performances of four relatively recent latent factor and/or penalized regression techniques (PLS, SPLS, LASSO and elastic net) to fit an AFT model based on high dimensional regressors specifically, to predict patient survival times using high dimensional MS data.

In this chapter, I focused on the two techniques performed best in the study of Mostajabi et al. (2012), SPLS and elastic net, to fit AFT models for predicting survival times of patients by using high dimensional MS data. These methods are then applied to analyze survival times generated from simulated mass spectra, as well as a real MS data set on advanced non-small cell lung caner (NSCLC) patients. To ensure the features used in analysis corresponding to real peaks, I applied a hard thresholding algorithm to remove noise signals from the MS data. However, under this denoising approach, the intensities under thresholds are considered as missing data and are usually replaced by zeroes artificially. The missing data patterns are not independent of the peak intensities of the peptides and can be considered as left censored data censored at the threshold. There are generally two basic strategies to dealing with missing values in practice. The simplest strategy is to work only with the complete intensities. That is the data used for a particular peptide/protein would be based on only the observed peak intensity, and the features containing missing data are excluded from the analysis. Alternatively, the missing values can be imputed. Tekwe et al. (2012) studied sev-

eral imputation algorithms. The imputation approach selected should be appropriate when the missing values have been censored, as they can result in biased estimates and statistical inference (Karpievitch et al., 2009). To retain sufficient true signals meanwhile reduce the bias for the subsequent predictive modeling analysis, I propose a nonparametric imputation approach based on Kaplan-Meier estimator by considering the aligned intensities on the selected m/z values across all spectra as life times. The detailed imputation scheme was explained in Section 4.2.2. I then compare the predictive performance of the patient survival times with and without the imputation of the left censored peaks. Additionally, I compare different penalized regression schemes along with the AFT models to predict the patient survival times.

4.2 Method

4.2.1 Preprocessing of MS Data

According to (Antoniadis et al., 2010; Morris et al., 2005), raw spectra acquired by TOF mass spectrometers are generally a mixture of a real signal, noise of different characteristics and a varying baseline. Statistically, a possible model for a given mass spectrum is to represent it schematically by the equation

$$y_i(t_j) = B_i(t_j) + N_i S_i(t_j) + \epsilon_{ij}, \tag{4.1}$$

where $y_i(t_j)$ represents the observed log spectral intensity for spectrum i at TOF t_j. The true signal $S_i(t)$, consists of a sum of possibly overlapping peaks, each corresponding to a particular biological molecule, e.g. a protein or a peptide. $B_i(t_j)$ is the baseline representing a relatively smooth artifact commonly seen in mass spectrometry data, N_i is a constant multiplicative factor to adjust for spectrum-specific variability, e.g. the possible different amounts of protein in each sample,

and ϵ_{ij} is an additive white noise with variance σ_i^2 arising from the measurement process. To perform feature extraction and quantification with MALDI-TOF-MS data, the observed TOFs, $t_j, j = 1, ..., J$ will be mapped to a set of inferred mass to charge ratios (m/z) $x_j, j = 1, ..., J$ by calibration. This step aligns multiple spectra and yields molecular masses that can be used to ascertain the protein identity of a peak of interest. Eventually, the data of a single spectrum is given in two columns, with the first column containing the m/z and the second column containing the corresponding intensity. Low-level preprocessing of the raw MS data are necessary to perform feature extraction and quantification with MALDI-TOF-MS data. Depending on analysis goals, the preprocessing procedures can be different and complex in different literatures (Datta et al., 2007; Antoniadis et al., 2010; Morris et al., 2005; Mostajabi et al., 2012; Ndukum et al., 2011). In our methodology, I performed three basic preprocessing steps as baseline subtraction, alignment, and denoising to maintain as much information as possible before applying the AFT models in the subsequent survival analysis.

In the baseline correction step, the baseline is subtracted from each point and rescale intensities of all spectra to positive producing a baseline corrected spectrum. This step is to remove systematic artifacts, usually attributed to clusters of ionized matrix molecules hitting the detector during early portions of the experiment, or to detector overload. The relations among raw data, baseline and processed data of one spectrum are illustrated by Figure 4.1. After baseline correction of the spectrum data, I apply a binning step to divide the m/z axis into intervals of desired length, which will help to extract meaningful peak pattern for alignment. The detailed binning scheme works as following: suppose we set the binning bandwidth as 0.5Da, we start by rounding all mass to charge values to the nearest 0.5Da. Then, moving from the lowest rounded m/z value of the spectrum to the right along the m/z axis, for each rounded m/z values say b, search for the

maximum intensity y within 0.5Da interval $(b - 0.25, b + 0.25)$. Further, select the maximum value of the intensity y on the corresponding rounded m/z value (or alternatively, bin the m/z values to the nearest 0.5Da and average over the intensity values with the same m/z value). To make sure the characteristic features occur at the same time in all spectra, the subsequent step is to align the spectra cross samples to make sure that the characteristic features occurs at the same time in all spectra. Eventually, all the binned spectra data are mapped to a matrix of common m/z values and the corresponding intensities across samples. The next step is to denoise the individual spectrum by using a hard threshold h. In hard thresholding, all intensities less than the threshold are set to zero, while all intensities no less than the threshold remain unchanged (Datta et al., 2007; Mostajabi et al., 2012; Ndukum et al., 2011). As the noise signals are usually assumed to be normally distributed, by referring the denoising scheme in Morris et al. (2005); Antoniadis et al. (2010), I propose to estimate the hard threshold for each spectrum as the median absolute deviation (MAD) of its raw intensities divided by 0.67. The hard thresholding criterion can be expressed as following:

$$\tilde{y}_i(x_j) = \begin{cases} y_i^*(x_j), & y_i^*(x_j) \geq \text{MAD}(y_i)/0.67; \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

The principle is based on keeping features with intensities greater than a certain threshold. The threshold should be large enough to eliminate initial noisy region but small enough to retain any peak that could correspond to real observable proteins or peptides.

4.2.2 Imputation of Denoised MS Data

After the denoising process with hard thresholds, intensity values that are less than the noise level are replaced by zeroes in each spectrum. We can extract the

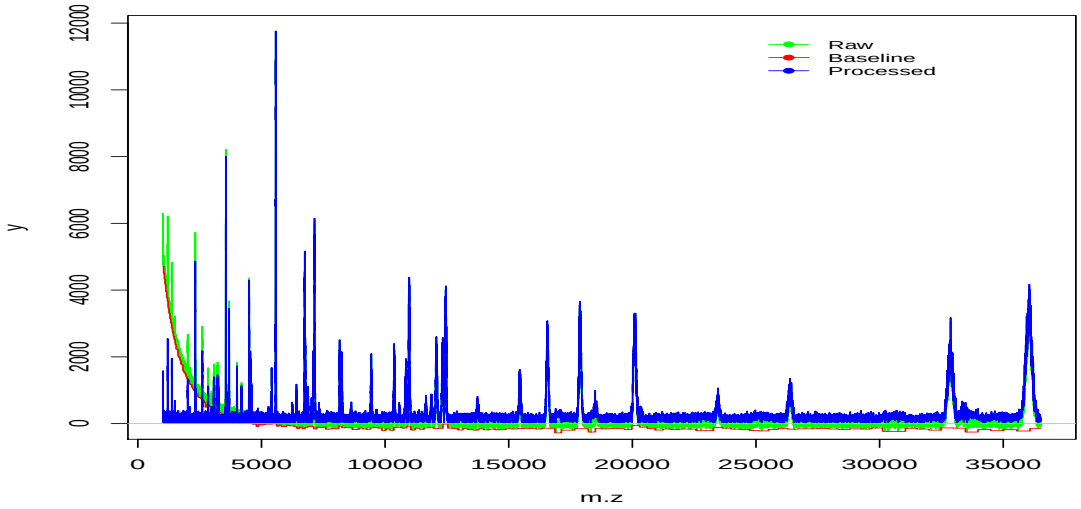


Figure 4.1: An example of baseline corrected spectra sample (Intensity vs M/Z value).

key interesting features for survival analysis by selecting the m/z's corresponding to more than a specified number of nonzero intensities across all samples. The selected intensity vectors are considered as covariates for the predicting models in the following procedure (Mostajabi et al., 2012). On the other hand, we may also want to maintain as many features as possible, and leave the AFT model to select the correlated features for survival prediction automatically. To reduce the bias caused by the zeroes induced in the denoising step, I proposed a nonparametric imputation method to impute the denoised value by its expected value given that the noise level of the spectrum was larger than the unobserved true signal.

In the proposed imputation algorithm, first for $i = 1, ..., n$, we have n spectra with n cutoffs of noise levels $h = (h_1, ..., h_n)$. After alignment, we have J selected m/z values on each spectrum. let X denote a variable for the intensities aligned on a selected m/z value across all spectra. Let T denote the variable of intensities in X that are larger than the noise levels h, and E denote the variable of intensities in X that are lower than h. In the denoising step, all values in E are replaced

by zeroes. However, If we consider X as a life time variable, X is therefore left censored with the values in E censored at the noised levels h respectively. By replacing each value in E with its expected value based on $T = \{\max(h, X); \delta = I(h \leq X))\}$, we can apply the AFT model to automatically select key related features for predicting patient survival times with reduced bias.

The detailed imputation process is as following: for each fixed j, the data is $T_{ij} = \{\max(h_i, X_{ij}); \delta = I(h_i \leq X_{ij})\}$, for $i = 1, ..., n$. The variable T_j can therefore be considered as a left censored life time variable. It is much easier to apply the Kaplan-Meier estimator to right censored variable compared with left censored one, therefore we apply a flipping approach to T_j such that $T'_j = \max(T_j) - T_j + (\max(T_j) - \min(T_j))$. The flipped data is then $T'_{ij} = \{\min(h'_i, X'_{ij}); \delta' = I(h'_i \geq X'_{ij})\}$, for $i = 1, ..., n$. When T_j is left censored, its flipped variable T'_j is right censored, and the survival function of T'_j becomes the cumulative distribution function (percentiles) of the original data T_j (Helsel and Lee, 2006).

Given a fixed j, after flipping T_j to T'_j, we can then compute the survival function for T'_j by the Kaplan-Meier estimator. Suppose the survival function of $T'_j = \{\min(h', X'_j); \delta' = I(h' \geq X'_j)\}$ is denoted as $S(t)$. Under the assumption that T'_j is independent of X'_j, $S(t)$ can be estimated by the Kaplan-Meier estimator

$$\hat{S}(t) = \prod_{\tau_{ij} \leq t} \left\{ 1 - \frac{\Delta N(\tau_{ij})}{R(\tau_{ij})} \right\}, \tag{4.3}$$

where $\tau_{(1j)}, ..., \tau_{(mj)}$ are the distinct ordered life time, $\Delta N^c(\tau_{ij})$ is the number of observations at time $\tau_{(ij)}$, and $R(\tau_{ij}) =$ Number of $\{k : T^c_{kj} \geq \tau_{ij}\}$, counts the number of individuals at risk of failing just before time $\tau_{(ij)}$.

The censored values can then be imputed from the survival distribution of T'_j. In detail, for the jth covariate X'_j, we keep the observed X'_{ij} intact; replace each of the unobserved X'_{ij} by its expected values given that the true failure time T'_{ij} was

larger than the censored time h_i'. It can be estimated from the Kaplan-Meier curve of the survival function of T_j' as

$$X_{ij}^{'*} = \{\hat{S}(h_i')\}^{-1} \sum_{\tau_{ij} > h_i'} T_{ij}' \Delta \hat{S}(\tau_{ij}),$$

where $\hat{S}$ is the Kaplan-Meier estimator of the survival function of T_j', and $\Delta\hat{S}(\tau_{ij})$ is the jump size of $\hat{S}$ at time τ_{ij}. Note that for this calculation, the largest event time τ_m is treated as a true failure no matter if $\delta_m = 0$ or not. The further explanation of this approach can be found in (Datta et al., 2007; Datta, 2005). The estimated conditional expectation $X_{ij}^{'*}$ is then flipped back to X_{ij}^* under the original scale. Thus, under this imputation approach, we let $\tilde{X}_j = X_j$, if $\delta_i = 1$, and $\tilde{X}_j = X_j^*$, if $\delta_i = 0$. Then the AFT model with SPLS and Enet can be used to fit the imputed MS data set on the log-transformed patient survival times, respectively.

4.2.3 Survival Prediction Models

The AFT model is presented as $logT = X^T\beta + \varepsilon$, where β is an unknown $p \times 1$ parameter of interest associated with the proteomic features X and ε is an unobservable random error term that is assumed to be independent of X. Each identified protein feature will be examined as an independent covariate. The association of each feature with patient survival or time-to-recurrence will be evaluated. The latent factor and regularization techniques for fitting the AFT model of $Y = logT$ (logarithm of the patient survival time) on the proteomic features X (intensity data corresponding to selected m/z values) of patients are selected by the SPLS and elastic net methods.

The sparse partial least squares regression (SPLS) (Chun and Keleş, 2010) is an extension of partial least squares regression (PLS) (Wold, 1985) to achieve simultaneous dimension reduction and variable selection. PLS extracts latent factors or

linear combinations of the original regressors that account for most of the variation in the response while avoiding over-fitting. PLS has become a very popular tool in the field of chemometrics and bioinformatics (Datta, 2001; Pihur et al., 2008). SPLS combines the latent factor approach with regularization to obtain good performance in prediction and variable selection by producing sparse linear combinations $X\beta^T$ of the original predictors X. This technique achieves the sparsity of the coefficients on X by adding the L_1 constraints on β, and is especially applicable when p is much greater than n (Chun and Keleş, 2010).

The elastic net (ENet) (Zou and Hastie, 2005) is a widely applied regularization and variable selection method. The ENet estimator is obtained by undoing the shrinkage for the naïve elastic net estimator that is obtained by minimizing the penalized least squares

$$\mathrm{L}(\lambda, \alpha, \beta_0, \beta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda P_\alpha(\beta), \tag{4.4}$$

where

$$P_\alpha(\beta) = \alpha\|\beta\|_{L_1} + \frac{1}{2}(1-\alpha)\|\beta\|_{L_2} = \sum_{j=1}^{p}\left\{\alpha|\beta_j| + \frac{1}{2}(1-\alpha)\beta_j^2\right\}. \tag{4.5}$$

Here P_α is the elastic net penalty that is a compromise between the ridge regression penalty ($\alpha = 0$) (Hoerl and Kennard, 1970) and the LASSO penalty ($\alpha = 1$) (Tibshirani, 1996). The elastic net penalty with $\alpha = 1 - \varepsilon$, for some small $\varepsilon > 0$, performs much like the LASSO but removes any degeneracies and wild behavior caused by extreme correlations (Friedman et al., 2010). For a given λ, as α increases from 0 to 1, the sparsity of the solution to (4.4), *i.e.*, the number of coefficients being zero, increases monotonically from 0 to the sparsity of the LASSO solution. The ENet penalty is particularly useful in the cases that p is greater than n and there are many correlated predictors (Zou and Hastie, 2005).

4.3 Simulation

Using the tool developed by Coombes et al. (2005) for simulating realistic mass spectra, Morris et al. (2005) simulated hundreds of proteomic data sets. The list of corresponding true peaks (features) is also available. There are 100 data sets from this collection each containing 100 spectra. We select the first data set for our simulation. The R package pkDACLASS (Ndukum et al., 2011) was used for the preprocessing steps explained early. By binning the m/z values to the nearest 1 Da, I identified 11832 potential features across all the 100 spectra for the subsequent prediction modeling procedure. The corresponding distinct m/z values are ranged from 941 Da to 24277 Da.

In order to retain the true signals from the denoised spectra to build a predictive model, three different approached are applied for pre-selection of interesting features in (Mostajabi et al., 2012). To exam the performance of our imputation technique, we also applied three approaches referring to (Mostajabi et al., 2012). The corresponding three groups of data sets of features identified from preprocessed MALDI-TOF-MS data are then denoted as: $X(1)$: features with no less than one nonzero denoised intensities in all spectra. $X(2)$: features with no less than half of nonzero denoised intensities in all spectra. $X(3)$: features with nonzero denoised intensities in all spectra. I denote $X(4)$ as the set of same features in $X(1)$ with zeroes imputed. The first and third approaches were same as applied in (Mostajabi et al., 2012). I increased the limit number of nonzeros in the second approach from 5 to 40 compared with in (Mostajabi et al., 2012).

To simulate survival times T, I randomly select 80 spectra from the first data set. Four different scenarios for the β coefficients are considered in our simulation. These are as follows: (i) $\beta_j = \exp\{-j\}$ for $1 \leq j \leq 11832$; (ii) $\beta_j = 1/j$ for $1 \leq j \leq 11832$; (iii) for $1 \leq j \leq 1000$, $\beta_j = (j \mod 5)/10$ if $j \mod 5 > 0$,

otherwise $\beta_j = 0.5$ and for $1001 \leq j \leq 11832$, $\beta_j = 0$; and (iv) $\beta_j = 0.1$ and for $1 \leq j \leq 11832$. Note that (i) and (ii) both represent situations when all the co-variables have positive but variable effects on survival; however, due to the exponential nature of the decaying coefficients, only the first few will have a real effect on survival in scenario (i). Case (iv) denotes an extreme hypothetical scenario when all covariates have the same positive effect on survival. Presumably, (iii) denotes the most realistic scenario when the collection of covariates contains a large number of pure noise variables. In each case, the vector of coefficients is randomly sampled for computational stability. A normal distribution with variance $r\sigma^2$ is used for generating the additive errors, where $\sigma^2 = \beta^T \sum_X \beta$ is the variability in the regression model. The variance-covariance matrix $\sum_X$ is a diagonal matrix with the diagonals are all set as 1 and the off-diagonals are all 0. r denotes a noise to signal ratio. Thus, log normally distributed failure times are considered. To maintain a similar scale of error variance in each scenario, a value of $r = 1$ was used for scenarios (i) and (iii), and $r = 0.1$ was used for scenarios (ii) and (iv).

For each design choice, I simulated our training data set by sampling 40 spectra at random from the preselected 80 spectra, and the left are used as test data set. Denote the training response variable as $Y_i^t = \log T_i^t$, for $i = 1, ..., 40$. Next, denote the response variable in the test data set as $Y_i^e = \log T_i^e$, for $i = 1, ..., 40$, I calculate the estimated mean squared error of prediction (EMSEP) for testing the prediction performance of each method. EMSEP here is computed as EMSEP$= \frac{1}{40} \sum_{i=1}^{40} (\hat{Y}_i^e - Y_i^e)^2$, where $\hat{Y}_i^e$ is the estimated predicting value calculated by using the fitted model on the ith sample in the test data set. Each of these measures is computed by averaging these quantities over 50 Monte-Carlo replicates.

From the simulation results in Table 4.1, we can easily see that the proposed

Table 4.1: Estimated mean squared error of prediction (EMSEP) for the simulated data. Four simulation scenarios of simulation settings are studied.

	Case i		Case ii		Case iii		Case iv	
	Enet	SPLS	Enet	SPLS	Enet	SPLS	Enet	SPLS
X(1) - Zeroes	0.97	1.56	2.66	4.42	1.51	11.01	1.31	2.58
X(2) - Half	0.97	1.51	2.70	4.11	2.03	10.05	1.34	2.95
X(3) - Complete	1.03	1.73	2.77	4.75	1.49	2.18	1.37	2.43
X(4) - Imputed	0.95	1.46	2.66	3.53	1.46	5.59	1.34	2.21

imputation approach (see the row $X(4)$ in Table 4.1) reduced the EMSEP in all four cases for both Enet and SPLS methods, comparing with inducing zero only method (see the row $X(1)$ in Table 4.1), except the in Case iv for ENet method. In case iv, the EMSEP for ENet in all four approaches are quite similar. There is no significant evidence to show that the pre-selection of interesting features (see the rows $X(1) - X(3)$ in Table 4.1) improved the survival prediction performance for both ENet and SPLS in all cases with only one exception (see the row $X(3)$ under case iii in Table 4.1). All these results showed that a proper imputation method should be applied to the denoised data so that we can retain as many interesting features as possible for the predictive modeling. I applied the analysis approach to Netherlands Non Small Cell Lung Cancer Data in the real case study (Voortman et al., 2009).

4.4 Netherlands Non Small Cell Lung Cancer Data Study

In this case study, I used the data set reported in Voortman et al. (2009). MALDI-TOF-MS spectra of serum samples of 27 patients with advanced non-small cell lung cancer (NSCLC), treated with chemotherapy and Bortezomib were obtained. Serum spectra of these patients are available at three time points: pre-treatment (preTx), after two cycles of treatment (post-2) and at the end of treatment (EOT). For each patient, there is an associated progression-free survival (PFS) recorded in days. No censoring exists in this data. The range of observed

survival time in this data set is 27 days to 601 days. I take EOT samples along with PFS values as major information for further analysis. Two samples are excluded due to missing EOT serum spectrum. Each spectrum consisted features with mass-to-charge ratio range of 800-4000 Dalton (Da) with the corresponding intensities. After baseline correction, it is necessary to align spectra so that characteristic features occur at the same time in all spectra. Because peak patterns are not clear in the data, to effectively select real features, I tried three scenarios by setting three different binning widths as 0.01 Da, 0.05 Da and 1 Da, separately. In different binning scenario, the number of features we selected for the regression modeling, indicated by the number of rounded m/z values are different. I selected the maximum intensities within each pre-specified intervals of m/z values. Next, I denoised all 25 spectra as described above.

To exam the performance of our imputation technique, I applied two approaches for feature selection. The corresponding two groups of data sets of features identified from preprocessed MALDI-TOF-MS data are then denoted as: $X_{(1)}$: Features with nonzero denoised intensities in all spectra. $X_{(2)}$: Features with no less than one nonzero denoised intensities in all spectra. I denote $X_{(3)}$ as the set of same features in $X_{(2)}$ with zeroes imputed. The numbers of features selected in each data set under different binning widths were summarized in Table 4.2.

Table 4.2: Number of selected features under different binning widths and feature selection approaches.

	Before Denoising	$X_{(2)}$	$X_{(1)}$
1.0 Da	3214	2757	900
0.5 Da	6427	4716	1474
0.1 Da	32131	15480	3701

In the analysis, I used each of the resulting feature sets $X_{(1)}$, $X_{(2)}$, and $X_{(3)}$,

respectively, in an AFT model to determine the relationship between progression free survival time (in days) and proteomic features for the 25 cancer samples. As mentioned before, two methods of modeling fitting SPLS and elastic net are implemented with each feature set. I compared the performance of these methods by computing their estimated mean squared error of prediction (EMSEP) which is minimized with respect to the selected values of the tuning (operational) parameters in a regression method. The EMSEP here is computed by leave-one out cross validation, EMSEP= $\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_{-i} - Y_i)^2$, where $\hat{Y}_{-i}$ is calculated by first fitting the model on the sample values other than the ith sample unit and predicting the ith value using the fitted model with the covariate X_i. SPLS regression has two key tuning parameters: the thresholding parameter (λ) and number of components (K). Following the guidelines given in Chun and Keles (2010), cross validation is computed over the grid of $K = 1, 2, ..., 20$ and $\lambda = 0.1, 0.2, ..., 0.9$. There are two tuning parameters in the elastic net as well. These are the penalty terms λ_1 and λ_2. I used the 'glmnet' R-package in the programming, and let the built-in cross validation function to decide the optimal tuning parameters automatically. LASSO is a special case of the elastic net with $\lambda_2 = 0$.

Table 4.3: Estimated mean squared error of prediction (EMSEP) for the Netherland NSCLC data. Three feature selection methods are tested; Under three different binning width 1.0 Da, 0.5 Da and 0.1 Da, $X(1)$ has 900, 1474, 3701 features and $X(2)$ has 2757, 4716, 15480 features. In each case, the minimum EMSEP value over the operational parameters is reported for each regression method.

	Elastic net			SPLS		
	1 Da	0.5 Da	0.1 Da	1 Da	0.5 Da	0.1 Da
Denoised	0.5256	0.5028	0.4885	0.6678	0.6357	0.6158
Complete	0.4957	0.4975	0.5116	0.8066	0.6233	0.7166
Imputed	0.3818	0.4172	0.4412	0.5230	0.6858	0.6949

Table 4.3 showed the measure of prediction for Netherland NSCLC data. For both SPLS and elastic net methods, the obtained prediction errors were getting smaller as the increase of width of binning (as the decrease of number of fea-

tures selected as showed in Table 4.2). Comparing SPLS and Elastic net, the later method performs better in all nine cases. For all Elastic net cases, the prediction errors from imputed data sets are smaller than other two data sets. Similar result showed for SPLS method, when the binning width is 1 Da. This showed that our imputation approach advanced the prediction performance for both two methods.

4.5 Discussion and Conclusions

The methods Elastic net and SPLS showed promise in predicting survival with properly preprocessed mass spectrometry data having large number of features versus limited sample size. Our simulation study confirmed the benefit of treating the intensity values under the noise levels as left-censored data, and the non-parametric imputation method we proposed based on Kaplan-Meier estimator effectively improved the performance of the prediction models.

It is not the primary purpose of this topic to identify the features and the corresponding proteins used for survival prediction. However, a further study in this direction can be conducted in the future research. Moreover, to explore the full effect of all preprocessing of MS data and feature selection strategies on survival prediction is beyond the scope of this article.

REFERENCES

B.L. Adam, Y. Qu, J.W. Davis, M.D. Ward, M.A. Clements, L.H. Cazares, O.J. Semmes, P.F. Schellhammer, Y. Yasui, Z. Feng, and G.L.Jr. Wright. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research*, 62:3609–3614, 2002.

R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422: 198–207, 2003.

A. Antoniadis, J. Bigot, and S. Lambert-Lacroix. Peaks detection and alignment for mass spectrometry data. *Journal de la Société Française de Statistique*, 151(1): 17–37, 2010.

M. Atlas and S. Datta. A statistical technique for monoisotopic peak detection in a mass spectrum. *Journal of Proteomis & Bioinformatics*, 2:202–216, 2009.

M. C. Berenbaum. What is synergy? *Pharmacological Reviews*, 41(2):93–141, 1989.

H.M. Bovelstad, S. Nygard, H.L. Storvold, M. Aldrin, O. Borgan, A. Frigessi, and O.C. Lingjaerde. Predicting survival from microarray data–a comparative study. *Bioinformatics*, 23:2080–2087, 2007.

J.R. Carpenter, M.G. Kenward, and I.R. White. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research*, 16(3):259–275, 2007.

Q. Chen and S. Wang. Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in medicine*, 2013.

H. Chun and S. Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

J. Cohen, P. Cohen, G.W. Stephen, and S.A. Leona. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 3 edition, 2003.

D.J. Conklin, P. Haberzettl, R.A. Prough, and A. Bhatnagar. Glutathione-s-transferase p protects against endothelial dysfunction induced by exposure to tobacco smoke. *American Journal of Physiology-Heart and Circulatory Physiology*, 296(5):H1586–H1597, 2009.

K.R. Coombes, J.M. Koomen, K.A. Baggerly, J.S. Morris, and R. Kobayashi. Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Informatics*, 1(1):41–52, 2005.

D.R. Cox. Regression models and life-tables. *Journal of Royal Statistical Scociety B*, 34:187–220, 1972.

S. Datta. Exploring relationships in gene expressions: a partial least squares approach. *Gene Expression*, 9(6):249–255, 2001.

S. Datta. Estimating the mean life time using right censored data. *Statistical Methodology*, 2(1):65–69, 2005.

S. Datta and V. Pihur. Feature selection and machine learning with mass spectrometry data. *Methods in Molecular Biology*, 593:205–229, 2010.

S. Datta, J. Le-Rademacher, and S. Datta. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and lasso. *Biometrics*, 63:259–271, 2007.

S. Datta, D. Turner, R. Singh, B. Ruset, W.M. Pierce, and T.B. Knudsen. Fetal alcohol syndrome in mice detected through proteomics screening of the amniotic fluid. *Birth Defects Research Part A: Clinical and Molecular Teratology*, 82:177–186, 2008.

J. Davignon and P. Ganz. Role of endothelial dysfunction in atherosclerosis. *Circulation*, 109(23 suppl 1):III–27, 2004.

D. Engler and Y. Li. Survival analysis with high-dimensional covariates: An application in microarray studies. *Statistical Applications in Genetics and Molecular Biology*, pages 8–14, 2009.

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

W.R. Greco, G. Bravo, and J.C. Parsons. The search for synergy: a critical review from a response surface perspective. *Pharmacological reviews*, 47(2):331–385, 1995.

J. Gui and H. Li. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21:3001–3008, 2005.

H. AR Hadi, C.S. Carr, and J. Al Suwaidi. Endothelial dysfunction: cardiovascular risk factors, therapy, and outcome. *Vascular Health and Risk Management*, 1(3): 183–198, 2005.

C. Harbron. A flexible unified approach to the analysis of pre-clinical combination studies. *Statistics in medicine*, 29(16):1746–1756, 2010.

W.M. Hardesty, M.C. Kelley, D. Mi, R.L. Low, and R.M. Caprioli. Protein signatures for survival and recurrence in metastatic melanoma. *Journal of Proteomics*, 74(7):1002–1014, 2011.

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

T. Hastie, R. Tibshirani, and J. JH Friedman. *The Elements of Statistical Learning*, volume 1. Springer, New York, 2001.

T.J. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

D. Helsel and L. Lee. Analysis of environmental data with nondetects. 2006.

M.W. Heymans, S. van Buuren, D.L. Knol, W. van Mechelen, and H. CW de Vet. Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC medical research methodology*, 7(1):33–42, 2007.

A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

J. Huang, S. Ma, and H. Xie. Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, 62:813–820, 2006.

H. Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Royal Statistical Society*, 60:271–293, 1993.

N. Jeffries. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21:3066–3073, 2005.

Y. Karpievitch, J. Stanley, T. Taverner, J. Huang, J.N. Adkins, C. Ansong, F. Heffron, T.O. Metz, W. Qian, H. Yoon, et al. A statistical framework for protein

quantitation in bottom-up ms-based proteomics. *Bioinformatics*, 25(16):2028–2034, 2009.

F. Kim, M. Pham, E. Maloney, N.O. Rizzo, G.J. Morton, B.E. Wisse, E.A. Kirk, A. Chait, and M.W. Schwartz. Vascular inflammation, insulin resistance, and reduced nitric oxide production precede the onset of peripheral insulin resistance. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 28(11):1982–1988, 2008.

M. Kong and R.L. Eubank. Monotone smoothing with application to dose-response curves. *Communication in Statistics-Computation and Simulation*, 35:991–1004, 2006.

M. Kong and J.J. Lee. A generalized response surface model with varying relative potency for assessing drug interaction. *Biometrics*, 62(4):986–995, 2006.

M. Kong and J.J. Lee. A semiparametric model for assessing drug interaction. *Biometrics*, 64(4):396–405, 2008.

J.J. Lee, M. Kong, G.D. Ayers, and R. Lotan. Interaction index and different methods for determining drug interaction in combination therapy. *Journal of biopharmaceutical statistics*, 17(3):461–480, 2007.

H. Li and Y. Luan. Kernel cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing*, 8:65–76, 2003.

R. JA Little and D.B. Rubin. *Statistical Analysis with Missing Data*, volume 539. Wiley, New York, 1987.

R. JA Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 2 edition, 2002.

J.S. Morris, K.R. Coombes, J. Koomen, K.A. Baggerly, and R. Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–1775, 2005.

F. Mostajabi, S. Datta, and S. Datta. Predicting patient survival from proteomic profile using mass spectrometry data: an empirical study. *Communications in Statistics–Simulation and Coumputation*, 42:485–498, 2012.

J. Ndukum, Atlas M., and S. Datta. pkdaclass: open source software for analysing maldi-tof data. *Bioinformation*, 6(1):45–47, 2011.

Y. Pawitan, J. Bjohle, S. Wedren, K. Humphreys, L. Skoog, F. Huang, L. Amler, P Shaw, P. Hall, and J. Bergh. Gene expression profiling for prognosis using cox regression. *Statistics in Medicine*, 23:1767–1780, 2004.

V. Pihur, S. Datta, and S. Datta. Reconstruction of genetic association networks from microarray data: a partial least squares approach. *Bioinformatics*, 24(4): 561–568, 2008.

J.L. Plummer and T.G. Short. Statistical modeling of the effects of drug combinations. *Journal of pharmacological methods*, 23(4):297–309, 1990.

T.E. Raghunathan, J.M. Lepkowski, J. Van Hoewyk, and P. Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–96, 2001.

A. J. Rai and D.W. Chan. Cancer proteomics: Serum diagnostics for tumor marker discovery. *Annals of the New York Academy of Science*, 1022:286–294, 2004.

J.O. Ramsay. Monotone regression splines in action. *Statistical Science*, 3:425–441, 1988.

J.O. Ramsay. Estimating smooth monotone functions. *Journal of Royal Statistical Society B*, 60:365–375, 1998.

J.O. Ramsay and M. Barahamowicz. Binomial regression with monotone splines: a psychometric application. *Journal of the American Statistical Association*, 84: 906–915, 1989.

B.Y. Renard, M. Kirchner, H. Steen, J.A. Steen, and F.A. Hamprecht. Nitpick: Peak dentification for mass spectrometry data. *BMC Bioinformatics*, page 9:355, 2008.

N.O. Rizzo, E. Maloney, M. Pham, I. Luttrell, H. Wessells, S. Tateya, G. Daum, P. Handa, M.W. Schwartz, and F. Kim. Reduced no-cgmp signaling contributes to vascular inflammation and insulin resistance induced by high-fat feeding. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 30(4):758–765, 2010.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, 1987.

D. Ruppert. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757, 2002.

G.A. Satten, S. Datta, H. Moura, A.R. Woolfitt, M.D. Carvalho, G.M. Carlone, B.K. De, A. Pavlopoulos, and J.R. Barr. Standardization and denoising algorithms formass spectra to classify whole-organism bacterial specimens. *Bioinformatics*, 20:3128–3136, 2004.

S. Savelev, E. Okello, NSL Perry, R.M. Wilkins, and E.K. Perry. Synergistic and antagonistic interactions of anticholinesterase terpenoids in< i> salvia lavandulaefolia</i> essential oil. *Pharmacology Biochemistry and Behavior*, 75(3):661–668, 2003.

M. Schomaker and C. Heumann. Model selection and model averaging after multiple imputation. *Computational Statistics & Data Analysis*, 2013.

M. Stoeckli, P. Chaurand, D.E. Hallahan, and R.M. Caprioli. Imaging mass spectrometry: A new technology for the analysis of protein expression in mammalian tissues. *Nature Medicine*, 7:493–496, 2001.

T.M. Stoker. Consistent estimation of scaled coefficients. *Econometrica*, 54:1461–1481, 1986.

Y. Su, M. Yajima, A.E. Gelman, and J. Hill. Multiple imputation with diagnostics (mi) in r: opening windows into the black box. *Journal of Statistical Software*, 45 (2):1–31, 2011. URL http://www.jstatsoft.org/v45/i02/.

C.D. Tekwe, R.J. Carroll, and A.R. Dabney. Application of survival analysis methodology to the quantitative analysis of lc-ms proteomics data. *Bioinformatics*, 28(15):1998–2003, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

S. Van Buuren. *Flexible imputation of missing data*. Print ISBN: 978-1-4398-6824-9; eBook ISBN: 978-1-4398-6825-6. Chapman and Hall/CRC Press, 2012.

S. Van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. URL http://www.jstatsoft.org/v45/i03/.

S. Van Buuren, H.C. Boshuizen, D.L. Knook, et al. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694, 1999.

D. Versari, E. Daghini, A. Virdis, L. Ghiadoni, and S. Taddei. Endothelial dysfunction as a target for prevention of cardiovascular disease. *Diabetes Care*, 32(suppl 2):S314–S321, 2009.

J. Voortman, T.V. Pham, J.C. Knol, G. Giaccone, and C.R. Jimenez. Prediction of outcome of non-small cell lung cancer patients treated with chemotherapy and bortezomib by time-course maldi-tof-ms serum peptide profiling. *Proteome Sci*, 7(1):34, 2009.

H. Wold. Partial least squares. *Encyclopedia of statistical sciences*, 1985.

A.M. Wood, I.R. White, and S.G. Thompson. Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. *Clinical trials*, 1(4):368–376, 2004.

A.M. Wood, I.R. White, M. Hillsdon, and J. Carpenter. Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. *International Journal of Epidemiology*, 34(1):89–99, 2005.

A.M. Wood, I.R. White, and P. Royston. How should variable selection be performed with multiply imputed data? *Statistics in medicine*, 27(17):3227–3246, 2008.

Y. Xia and W. Härdle. Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, 97(5):1162–1184, 2006.

Y. Yu and D. Ruppert. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054, 2002. URL http://dx.doi.org/10.1198/016214502388618861.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

APPENDICES

Appendix A Specification of Penalty Matrix D

Let us denote the $K+4$ knots as $\{\tau_1, \tau_2, ..., \tau_{K+4}\}$. By formulas (3.6) and (3.7), we have

$$\begin{aligned}
M_j^3(z) =& \frac{3}{2(\tau_{j+3}-\tau_j)}\left[(z-\tau_j)M_j^2(z) + (\tau_{j+3}-z)M_{j+1}^2(z)\right] \\
=& \frac{3(z-\tau_j)^2}{(\tau_{j+3}-\tau_j)(\tau_{j+2}-\tau_j)}M_j^1(z) \\
&+ \left[\frac{3(z-\tau_j)(\tau_{j+2}-z)}{(\tau_{j+3}-\tau_j)(\tau_{j+2}-\tau_j)} + \frac{3(\tau_{j+3}-z)(z-\tau_{j+1})}{(\tau_{j+3}-\tau_j)(\tau_{j+3}-\tau_{j+1})}\right]M_{j+1}^1(z) \\
&+ \frac{3(\tau_{j+3}-z)^2}{(\tau_{j+3}-\tau_j)(\tau_{j+3}-\tau_{j+1})}M_{j+2}^1(z).
\end{aligned}$$

$$\begin{aligned}
M_j'^3(z) =& \frac{6(z-\tau_j)}{(\tau_{j+3}-\tau_j)(\tau_{j+2}-\tau_j)}M_j^1(z) \\
&+ \left[\frac{3(\tau_j+\tau_{j+2}-2z)}{(\tau_{j+3}-\tau_j)(\tau_{j+2}-\tau_j)} + \frac{3(\tau_{j+1}+\tau_{j+3}-2z)}{(\tau_{j+3}-t_j)(\tau_{j+3}-\tau_{j+1})}\right]M_{j+1}^1(z) \\
&+ \frac{6(z-\tau_{j+3})}{(\tau_{j+3}-\tau_j)(\tau_{j+3}-\tau_{j+1})}M_{j+2}^1(z) \\
=& a_{j,1}(z-\tau_j)M_j^1(z) + (a_{j,2}z + a_{j,3})M_{j+1}^1(z) + a_{j,4}(z-\tau_{j+3})M_{j+2}^1(z),
\end{aligned}$$

with

$$
\begin{aligned}
a_{j,1} &= \frac{6}{(\tau_{j+3}-\tau_j)(\tau_{j+2}-\tau_j)},\\
a_{j,2} &= \frac{-6}{\tau_{j+3}-\tau_j}\left(\frac{1}{\tau_{j+2}-\tau_j}+\frac{1}{\tau_{j+3}-\tau_{j+1}}\right),\\
a_{j,3} &= \frac{3}{\tau_{j+3}-\tau_j}\left(\frac{\tau_{j+2}+\tau_j}{\tau_{j+2}-\tau_j}+\frac{\tau_{j+3}+\tau_{j+1}}{\tau_{j+3}-\tau_{j+1}}\right),\\
a_{j,4} &= \frac{6}{(\tau_{j+3}-\tau_j)(\tau_{j+3}-\tau_{j+1})}.
\end{aligned}
\tag{4.6}
$$

Therefore, if $1 \le j \le K+1$, we have

$$
\begin{aligned}
&\int_{-\infty}^{\infty}\left[M_j'^3(z)\right]^2 dz = \int_{\tau_j}^{\tau_{j+3}}\left[M_j'^3(z)\right]^2 dz\\
=&\int_{\tau_j}^{\tau_{j+1}}\left[a_{j,1}(z-\tau_j)M_j^1(z)\right]^2 dz + \int_{\tau_{j+1}}^{\tau_{j+2}}\left[(a_{j,2}z+a_{j,3})M_{j+1}^1(z)\right]^2 dz\\
&+\int_{\tau_{j+2}}^{\tau_{j+3}}\left[a_{j,4}(z-\tau_{j+3})M_{j+2}^1(z)\right]^2 dz\\
=&\frac{a_{j,1}^2(\tau_{j+1}-\tau_j)}{3}+\frac{a_{j,2}^2(\tau_{j+2}^2+\tau_{j+1}^2+\tau_{j+2}\tau_{j+1})+3a_{j,2}a_{j,3}(\tau_{j+2}+\tau_{j+1})+3a_{j,3}^2}{3(\tau_{j+2}-\tau_{j+1})}\\
&+\frac{a_{j,4}^2(\tau_{j+3}-\tau_{j+2})}{3}.
\end{aligned}
$$

Notice that $\int_L^U \left[M_j'^3(z)\right]^2 dz = \int_{\tau_3}^{\tau_{K+2}} \left[M_j'^3(z)\right]^2 dz$, so if we set

$$a_1 = \begin{cases} \frac{a_{j,1}^2(\tau_{j+1}-\tau_j)}{3}, & 3 \le j \le K+1; \\ 0, & \text{otherwise,} \end{cases}$$

$$a_2 = \begin{cases} \frac{a_{j,2}^2(\tau_{j+2}^2+\tau_{j+1}^2+\tau_{j+2}\tau_{j+1})+3a_{j,2}a_{j,3}(\tau_{j+2}+\tau_{j+1})+3a_{j,3}^2}{3(\tau_{j+2}-\tau_{j+1})}, & 2 \le j \le K; \\ 0, & \text{otherwise,} \end{cases}$$

$$a_3 = \begin{cases} \frac{a_{j,4}^2(\tau_{j+3}-\tau_{j+2})}{3}, & 1 \le j \le K-1; \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\int_L^U \left[M_j'^3(z)\right]^2 dz = a_1 + a_2 + a_3. \tag{4.7}$$

In the similar manner, if $1 \le j \le K$, we have

$$\begin{aligned} &\int_{-\infty}^{\infty} M_j'^3(z) M_{j+1}'^3(z) dz = \int_{\tau_{j+1}}^{\tau_{j+3}} M_j'^3(z) M_{j+1}'^3 dz \\ =&\int_{\tau_{j+1}}^{\tau_{j+2}} \left[(a_{j,2}z + a_{j,3}) M_{j+1}^1(z)\right] \left[a_{j+1,1}(z-\tau_{j+1}) M_{j+1}^1(z)\right] dz \\ &+ \int_{\tau_{j+2}}^{\tau_{j+3}} \left[a_{j,4}(z-\tau_{j+3}) M_{j+2}^1(z)\right] \left[(a_{j+1,2}z + a_{j+1,3}) M_{j+2}^1(z)\right] dz \\ =&\frac{a_{j+1,1}\left[a_{j,2}(2\tau_{j+2}+\tau_{j+1})+3a_{j,3}\right]}{6} + \frac{-a_{j,4}\left[a_{j+1,2}(2\tau_{j+2}+\tau_{j+3})+3a_{j+1,3}\right]}{6}. \end{aligned}$$

Also observe that $\int_L^U M_j'^3(z) M_{j+1}'^3 dz = \int_{\tau_3}^{\tau_{K+2}} M_j'^3(z) M_{j+1}'^3 dz$, if we denote

$$a_1 = \begin{cases} \frac{a_{j+1,1}\left[a_{j,2}(2\tau_{j+2}+\tau_{j+1})+3a_{j,3}\right]}{6}, & 2 \le j \le K; \\ 0, & \text{otherwise,} \end{cases}$$

$$a_2 = \begin{cases} \frac{-a_{j,4}\left[a_{j+1,2}(2\tau_{j+2}+\tau_{j+3})+3a_{j+1,3}\right]}{6}, & 1 \le j \le K-1; \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\int_L^U M_j'^3(z) M_{j+1}'^3 dz = a_1 + a_2. \tag{4.8}$$

Finally, if $1 \leq j \leq K-1$, we have

$$\begin{aligned}
&\int_{-\infty}^{\infty} M_j'^3(z) M_{j+2}'^3(z) dz = \int_{\tau_{j+2}}^{\tau_{j+3}} M_j'^3(z) M_{j+2}'^3 dz \\
=&\int_{\tau_{j+2}}^{\tau_{j+3}} \left[a_{j,4}(z-\tau_{j+3}) M_{j+2}^1(z)\right] \left[a_{j+2,1}(z-\tau_{j+2}) M_{j+2}^1(z)\right] dz \\
=&\frac{-a_{j,4} a_{j+2,1}(\tau_{j+3}-\tau_{j+2})}{6}.
\end{aligned}$$

That is

$$\int_L^U M_j'^3(z) M_{j+2}'^3 dz = \frac{-a_{j,4} a_{j+2,1}(\tau_{j+3}-\tau_{j+2})}{6} \tag{4.9}$$

According to above calculations, we can construct the algorithm to form the matrix D as following:

For $j = 1, ..., K+1$, calculate the matrix $a_{j,1}$, $a_{j,2}$, $a_{j,3}$ and $a_{j,4}$ according to (4.6).

Set $D(j,l) = 0$ for all $j, l = 1, ..., K+1$.

For $j = 1, ..., K+1$,

set $D(j,j)$ according to (4.7),

set $D(j, j+1)$ and $D(j+1, j)$ according to (4.8), if $j+1 \leq K+1$,

set $D(j, j+2)$ and $D(j+2, j)$ according to (4.9), if $j+2 \leq K+1$.

Appendix B Bootstrap Standard Error of $g(d_1, d_2)$

The standard error for $\hat{g}(d_1, d_2)$ can be obtained via bootstrap. The detailed procedure is summarized as following:

Step 1. Fit the model based on the original observations, obtain the estimates $\hat{\alpha}$ for α, and $\hat{\beta}$ for β.

Step 2. Obtain the residuals from the final estimate of $f(*)$, i.e. $\epsilon_i = y_i - \hat{f}(z) = y_i - \sum_{j=-1}^{K-1} \hat{\beta}_j I_j^3(x_i^T\hat{\alpha}), i = 1, ..., n$.

Step 3. Generate bootstrap data set by replacing y_i with $\sum_{j=-1}^{K-1} \hat{\beta}_j I_j^3(x_i^T\hat{\alpha}) + \epsilon_i^*$, $i = 1, ..., n$, where $\{\epsilon_1^*, ..., \epsilon_n^*\}$ is a random sample from the residuals obtained in step 2.

Step 4. Fit the model using the generated data, and then obtain the estimated $\hat{\alpha}$ and $\hat{g}(d_1, d_2)$.

Step 5. Repeat step 2 to step 4 B (say, 100) times.

If we denote the estimated $g(d_1, d_2)$ in the bth ($b = 1, ..., B$) iteration as $g^{*b}(d_1, d_2)$, the standard deviation for $g(d_1, d_2)$ will be estimated by

$$\widehat{SD}^{*B}(\hat{g}(d_1, d_2)) = \left(\frac{1}{B} \sum_{b=1}^{B} (g^{*b}(d_1, d_2) - \hat{g}(d_1, d_2))^2 \right),$$

thus a $100(1-\alpha)\%$ pointwise confidence interval for $g(d_1, d_2)$ can be constructed as

$$\left[\hat{g}(d_1, d_2) - z_{\alpha/2} \times \widehat{SD}^{*B}(\hat{g}(d_1, d_2)),\ \hat{g}(d_1, d_2) + z_{\alpha/2} \times \widehat{SD}^{*B}(\hat{g}(d_1, d_2)) \right],$$

where $z_{\alpha/2}$ is the upper $\frac{\alpha}{2} \times 100\%$ percentile of the standard normal distribution, and $\hat{g}(d_1, d_2)$ is the estimate for $g(d_1, d_2)$. Our case study in Section 5 showed that the estimated variance for $g(d_1, d_2)$ can account for the carry-over errors from estimating the marginal dose-effect curves. Our simulations given in Section 4 showed that the proposed bootstrap CIs have good coverage properties.

CURRICULUM VITA

Yubing Wan

1703 S 4th St Apt 3, Louisville, KY 40208

(502) 432-2105

y0wan002@exchange.louisville.edu

EDUCATION

Ph.D. Biostatistics, University of Louisville (UofL), Louisville, KY, August 2014 (expected)

Thesis title: Penalized regressions for variable selection model, single index model and survival prediction model

Supervisors: Dr. Maiying Kong and Dr. Susmita Datta

M.S. Mathematics, University of Texas-Pan American (UTPA), Edinburg, TX, May, 2009

Thesis title: Reaction-diffusion systems with a nonlinear rate of growth

Supervisor: Dr. Zhaosheng Feng

B.S. Mathematics and Applied Mathematics, China University of Mining and Technology (CUMT), Xuzhou, Jiangsu, China, July, 2007

PROFESSIONAL EXPERIENCE

Graduate Research Assistant Department of Bioinformatics and Biostatistics, UofL, Louisville, KY, 08/2010 – Present

Assisted in an NIH/NHLBI grant supported project CAESAR for study design, data analysis, and interpretation of statistical results. I have gathered significant experience in Collecting and cleaning experimental data from multiple centers, developing suitable data format for statistical analysis, conducting statistical analysis by using R, SAS and Excel and preparing analysis reports with summary tables and figures.

Developed monotonic single-index models utilizing penalized regression splines to assess drug interaction effectively.

Developed survival prediction models with high-throughput Mass Spectrometry data for proteomic profiling of important protein signatures.

Developed a weighted elastic net method for variable selection and prediction with missing data, and applied to examine the correlated predictors for the endothelial function in an ex-vivo experiment successfully.

Developed and implemented self-designed algorithms for all above methods using R, and conducted numerical simulations to examine the performance of each method.

Assisted in providing consulting service to a clinical research team at the School of Medicine, UofL. Developed linear/nonlinear mixed-effect models for analyzing effective gene factors in warfarin metabolism with LC-MS data provided by the collaborators.

Department of Computer Science, UTPA, Edinburg, TX, 08/2009 – 08/2010

Advanced a probabilistic model and algorithm for haplotype construction from incomplete or inconsistent sequences of haplotype fragments.

Quality Enhancement Plan (QEP) Project Assistant Department of Mathematics, UTPA, Edinburg, TX, 08/2008 – 05/2009

Conducted statistical analysis on QEP data from student survey by using SPSS.

Prepared analysis reports to project coordinator and administrator regularly.

Graduate Teaching Assistant (Lecturer) Department of Mathematics, UTPA, Edinburg, TX, 08/2007 – 08/2009

Independently lectured seven sections of undergraduate level courses Elementary and Intermediate Algebra and College Algebra.

PUBLICATIONS

Journal Articles

Z. Feng, **Y. Wan**. Linearizing Transformations to a Generalized Reaction-diffusion System, 2010, *Applicable Analysis* Vol. 89, No. 7, July 2010, 1005–1021.

Y. Wan, S. Datta, D. J. Conklin, M. Kong. Variable Selection Models Based on Multiple Imputation with an Application for Predicting Median Effective Dose and Maximum Effect. Accepted by *the Journal of Statistical Computation and Simulation*.

Y. Wan, M. Kong., S. Datta (in preparation). Monotonic Single Index Models with an Application to Assessing Drug Interaction.

Y. Wan, S. Datta, M. Kong (in preparation). Survival Prediction Models for Proteomic Profiling of Protein Signatures using Mass Spectrometry Data.

Book Chapters

L. Ding, B. Fu, Y. Fu, **Y. Wan**. Application of Width-Bounded Separators to Protein Side Chain Packing Problem, Sequence and Genome Analysis: Methods and Applications, ISBN: 978-0-9807330-4-4, 2010.

CONFERENCE PRESENTATIONS

Z. Feng, **Y. Wan** (May 7-9, 2009). Reaction-diffusion Systems with A Nonlinear Rate of Growth, 8th Mississippi State - UAB Conference on Differential Equations & Computational Simulations, Department of Mathematics and Statistics, Mississippi State University Mississippi State, MS, USA.

Y. Wan, S. Datta, D. J. Conklin, M. Kong (June 2-5, 2013). Extended Variable Selection Models for Missing Data with Application to Predict Median Effective Dose and Maximum Effect, Southern Regional Council on Statistics Summer Research Conference, Montgomery Bell State Park in Burns, TN, USA.

Y. Wan, S. Datta, D. J. Conklin, M. Kong (June 10-12, 2013). Extended Variable Selection Models for Missing Data with Application to Predict Median Effective Dose and Maximum Effect, 2nd International Conference and Exhibition on Biometrics & Biostatistics, Chicago/Northbrook, IL, USA.

Y. Wan, S. Datta, D. J. Conklin, M. Kong (October 16, 2013). Variable Selection Models Based on Multiple Imputation with an Application for Predicting Median Effective Dose and Maximum Effect, ASA Kentucky Chapter Meeting, Louisville, KY, USA.

SKILLS

Strong knowledge of R, Mathematica, SPSS, C++, Windows office and Latex;

Expert in statistical inference and consulting, modeling and programming, data mining and analysis;

Basic knowledge of SAS, JAVA, Matlab, MySQL and Unix Shell

Well trained in database management

GROUP MEMBERSHIPS

American Mathematical Society (AMS)

American Statistical Association (ASA)

HONORS & AWARDS

Boyd Harshbarger Student Travel Award, Southern Regional Council on Statistics, American Statistical Association and the National Science Foundation, 06/2013

IMA Travel Support, Department of Mathematics and Statistics, Mississippi State University, 05/2009

Third Place in HESTC scientific poster competition, UTPA, 09/2008 & 2009

Outstanding Student Award, China University of Mining and Technology, 11/2005 & 2006

REFERENCES

Dr. Maiying Kong, Associate Professor, Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, Tel: (502) 852-3988, maiying.kong@louisville.edu

Dr. Susmita Datta, Professor, Graduate Director and University Scholar, Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, Tel: (502) 852-0081, susmita.datta@louisville.edu

Dr. Zhaosheng Feng, Associate Professor, Department of Mathematics, University of Texas-Pan American, Edinburg, TX 78539, Tel: (956) 665-7483, zsfeng@utpa.edu