Contributions to Structured Variable Selection Towards Enhancing Model Interpretation and Computation Efficiency

Sumin Shen

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Xinwei Deng, Chair

Ran Jin

Inyoung Kim

Christopher T. Franck

December 19, 2019

Blacksburg, Virginia

Keywords: Nonnegative Garrote Method; Mixture Experiments; Dynamic Coefficient; Fused Lasso; Group Lasso; Alternative Direction Method of Multipliers (ADMM); Expectation-Maximization (EM).

Contributions to Structured Variable Selection Towards Enhancing Model Interpretation and Computation Efficiency

Sumin Shen

Academic Abstract

The advances in data-collecting technologies provides great opportunities to access large sample-size data sets with high dimensionality. Variable selection is an important procedure to extract useful knowledge from such complex data. While in many real-data applications, appropriate selection of variables should facilitate the model interpretation and computation efficiency. It is thus important to incorporate domain knowledge of underlying data generation mechanism to select key variables for improving the model performance. However, general variable selection techniques, such as the best subset selection and the Lasso, often do not take the underlying data generation mechanism into considerations. This thesis proposal aims to develop statistical modeling methodologies with a focus on the structured variable selection towards better model interpretation and computation efficiency. Specifically, this thesis proposal consists of three parts: an additive heredity model with coefficients incorporating the multi-level data, a regularized dynamic generalized linear model with piecewise constant functional coefficients, and a structured variable selection method within the best subset selection framework.

In Chapter 2, an additive heredity model is proposed for analyzing mixture-of-mixtures (MoM) experiments. The MoM experiment is different from the classical mixture experiment in that the mixture component in MoM experiments, known as the major component, is made up of sub-components, known as the minor components. The proposed model considers an additive structure to inherently connect the major components with the minor components. To enable a meaningful interpretation for the estimated model, we apply the hierarchical and heredity principles by using the nonnegative garrote technique for model selection. The performance of the additive heredity model was compared to several conventional methods in both unconstrained and constrained MoM experiments. The additive heredity model was then successfully applied in a real problem of optimizing the Pringles® potato crisp studied previously in the literature.

In Chapter 3, we consider the dynamic effects of variables in the generalized linear model

such as logistic regression. This work is motivated from the engineering problem with varying effects of process variables to product quality caused by equipment degradation. To address such challenge, we propose a penalized dynamic regression model which is flexible to estimate the dynamic coefficient structure. The proposed method considers modeling the functional coefficient parameter as piecewise constant functions. Specifically, under the penalized regression framework, the fused lasso penalty is adopted for detecting the changes in the dynamic coefficients. The group lasso penalty is applied to enable a sparse selection of variables. Moreover, an efficient parameter estimation algorithm is also developed based on alternating direction method of multipliers. The performance of the dynamic coefficient model is evaluated in numerical studies and three real-data examples.

In Chapter 4, we develop a structured variable selection method within the best subset selection framework. In the literature, many techniques within the LASSO framework have been developed to address structured variable selection issues. However, less attention has been spent on structured best subset selection problems. In this work, we propose a sparse Ridge regression method to address structured variable selection issues. The key idea of the proposed method is to re-construct the regression matrix in the angle of experimental designs. We employ the estimation-maximization algorithm to formulate the best subset selection problem as an iterative linear integer optimization (LIO) problem. We demonstrate the power of the proposed method in various structured variable selection problems. Moverover, the proposed method can be extended to the ridge penalized best subset selection problems. The performance of the proposed method is evaluated in numerical studies.

Contributions to Structured Variable Selection Towards Enhancing Model Interpretation and Computation Efficiency

Sumin Shen

General Audience Abstract

The advances in data-collecting technologies provides great opportunities to access large sample-size data sets with high dimensionality. Variable selection is an important procedure to extract useful knowledge from such complex data. While in many real-data applications, appropriate selection of variables should facilitate the model interpretation and computation efficiency. It is thus important to incorporate domain knowledge of underlying data generation mechanism to select key variables for improving the model performance.

However, general variable selection techniques often do not take the underlying data generation mechanism into considerations. This thesis proposal aims to develop statistical modeling methodologies with a focus on the structured variable selection towards better model interpretation and computation efficiency. The proposed approaches have been applied to real-world problems to demonstrate their model performance.

Dedication

To my parents,

Lei Xiang and Chunyan Shen,
who love hard-working

Acknowledgments

This dissertation would not have been possible without the help and support of my advisor, committee, family, and friends. I would first like to express my great gratitude to my Ph.D. advisor, Prof. Xinwei Deng, who has given me the opportunity to learn and grow in his research group. In fact, I was getting to know Dr. Deng before joining the Statistics Department through Dr. Zhiyang Zhang, who gave me suggestions and support since I was in the Chemistry Department. After being accepted as a MS graduate in Statistics by Prof. Birch in the summer of 2014, I started making contacts with Dr. Deng for suggestions on Statistics. Instead of turning me down, Dr. Deng invited us to his house for dinner and gave me the opportunity to start working on the first project, mixture-of-mixtures experiment, at the end of fall semester. The first project is fundamental but difficult for me at that time, Dr. Deng actively talked me through the ideas, allowed time for me to digest the concepts, and sharply corrected my mistakes. Dr. Deng showed his support patiently and walked me through the first project. Besides, Dr. Deng used side projects to display me the correct discipline and positive attitude to do the research job, though I did not realize this point of role models until my senior years. During the five years in the group, I was not only financially supported since 2016, but allowed enough time to explore myself and my learning. This self-learning time is important to me to rebuild my life. As an international graduate who was confused and not clear about my goal during the first years at the United States, I used these time to gain understanding and confidence, more importantly, to adjust behaviors that I developed and mistakes that I made during the very first years at U.S.. Most of my statistical learning skills and favorable working philosophy come from Dr. Deng.

I am also grateful to my committee members, Prof. Ran Jin, Prof. Inyoung Kim, Prof. Christopher T. Franck. Dr. Jin is similar in spirits to Dr. Deng in teaching students. Dr. Kim and Dr. Franck are the two professors that led me to the Statistics community in my first calendar year (Fall 2014) in the Statistics Department. They are hard-working, productive, and inspiring me to do things in earnest and enjoy life.

Thanks should also go to Prof. Alan Esker, Prof. Jeff Birch, Prof. Paul Deck, Mr. Thomas E Bell, Ms. Ruth Athanson, and Ms. Candace E. Wall. Dr. Esker brought me to Virginia

Tech in 2011, and provided support and encouragement during the first years. Dr. Esker took his time to teach me and guide me. Dr. Birch brought me to Department of Statistics, and support me in the first years in Statistics. Prof. Paul Deck, Mr. Thomas E Bell and Ms. Ruth Athanson helped me survive the visa issues in the summer, 2013. Ms. Candace E. Wall helped me edit my MS thesis in Chemistry.

I am also grateful to my friends at Virginia Tech (VT) during these years. Though it is not possible to list all your names in this limited space, I want to say thank you for all the moments, no matter happy or sad, that you brought. One of the fascinating aspects of the stay at VT is to know your lovely characters and learn from you.

Special thanks to my supportive fiancé, Dr. Xiaomin Xu, for her love. I am grateful for her support on this beautiful trip and many places that we explored together: Humboldt-Universität zu Berlin at Berlin, RIKEN at Saitama, The Chinese University of Hong Kong at HK, VT at Blacksburg, Boston, Vegas, Chicago, Indianapolis, Spartanburg, San Francisco, and DC.

Lastly, I would like to say thank you to my loving and protective parents, Lei Xiang and Chunyan Shen. I am deeply affected by their unconditional support and sacrifice.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Motivation

In modern statistical modeling and analysis, variable selection is an important technique to extract useful knowledge from data with complex structures. Proper variable selection will lead to a sparse and interpretable model. The sparsity assumption is to encourage a small number of predictors to be included in the model, which facilitates more interpretable models. Commonly used variable selection techniques include the forward selection (Efroymson 1966; Draper and Smith 1966), the best subset selection (Beale et al. 1967; Hocking and Leslie 1967), the nonnegative garrote method (Breiman 1995), the lasso (Tibshirani 1996), and the elastic net (Zou and Hastie 2005). Most of these methods consider generating a sparse model from the angle of prediction.

However, the general variable selection techniques often do not take the data generation mechanism into consideration, where the underlying structures of variables are often in concord with the original real-data application. Incorporating knowledge about the structures of variables will be very useful to accurately select important predictor variables for improving model prediction and interpretation. In the literature, there are several structured variable selection techniques proposed, such as the group lasso (Yuan and Lin 2006; Meier et al., 2008), the overlapping group lasso (Jacob et al. 2009), the fused lasso (Tibshirani et al. 2005), the sparse group lasso (Simon et al. 2013), the hierarchical selection (Zhao et al. 2006), and the tree-guided group lasso (Liu and Ye, 2010; Kim and Xing 2010), among

many others.

In this dissertation, the focus is on structured variable selection techniques, which lead to better model interpretation and computation efficiency. This dissertation aims to increase knowledge of modeling through a careful consideration on the structures of variables. Here the information on the structures of variables is originated from the real applications. From a technical perspective, we use the idea of varying-coefficients to provide a flexible framework to incorporate structures of variables into modeling.

The first motivation problem is the recent Mixture-of-mixtures (MoM) experiment studied by Kang et al. (2011). The goal of MoM experiments is to formulate a new kind of Pringles® potato crisp such that the chips do not break easily when the package form is changed from a can to a bag. It is important to consider how to incorporate the structure of variables into the modeling and analysis such that the proposed model can be better interpreted for finding the optimal formulation of products. The second motivation problem is the crystal ingot growth experiment in semiconductor manufacturing (Zhang et al. 2014). The goal of crystal ingot growth experiments is to produce high-quality crystal ingot by controlling the ingot diameter in the target range. The knowledge of dynamic effects of process variables on the quality response during the crystal ingot growth is important in the modeling and analysis of such data for reducing the product defects. The third motivation study is the recent proposed conditional main effects (CME) analysis. The CME, defined as the conditional effect of a factor at a fixed level of another factor, is intended to provide a better understanding of traditional interaction terms. The CME reformulation not only increases the dimension of predictor feature spaces, but introduces hierarchical structures and group structures within the features. It is challenging to develop a structured variable selection method to address the variable selection issues in the CME analysis.

1.2 Multilevel Structures in Mixture-of-Mixtures Experiments

In the MoM experiments where each mixture component (so-called "major component") is made up by a mixture of sub-components (so-called "minor components"), the predictor variable is the minor components' proportion with respect to the entire mixture. But this predictor variable could be decomposed into two levels of variables: the minor components' proportion with respect to the major component and the major components' proportion with respect to the entire mixture. This unique two-level structure of predictor variables is to a large extent related to the traditional multilevel model.

The traditional multilevel model, also known as the linear mixed model or the random coefficient model or the hierarchical linear model, assumes that in the model exist both fixed and random effects (Heck and Thomas 2015). The multilevel model usually has two types of variables: the individual-level variables and the group-level variables (Dedrick et al. 2009). Generally, the individual-level observations are correlated with each other within group-level variable levels. For example, when using the random intercept in the multilevel model, the intercept term will be dependent upon the group-level variable, representing the direct contribution of the group-level variable to the response of interest. The interaction between the individual-level variable and the group-level variable represents the indirect contribution of the group-level variable.

In the MoM experiments, the coefficient parameters of variables can be modeled as functional fixed effects. Kang et al. (2011) proposed a so-called major-minor model to use the Scheff*é* model to capture the relationship between the mean response and the major components, while the coefficients of major components and their interactions are modeled as a function of their respective minor components. In Chapter 2, we propose an additive heredity model to better capture the major-minor structure of MoM experiments with meaningful inter-

pretation. Specifically, we impose the coefficients of the minor components to be functions of their respective major components such that the minor components exist only when the corresponding major component is present in the model.

1.3 Varying-Coefficients in Crystal Ingot Growth Experiments

In the crystal ingot growth experiments, the produced silicon ingot would be sliced into silicon wafers for applications in solar cells and integrated circuits. The Czochralski (CZ) process is the common method to grow the crystal ingot (Scheel et al., 2003), which consists of four stages: Originally, the polycrystalline silicon is melted in a silica crucible. Next, a seed crystal is dipped into the melt to grow the ingot into the desired diameter. Then, the ingot is pulled upwards and rotated simultaneously and slowly. This step grows the majority part of the ingot and lasts more than 20 hours. At last, the ingot completes its growth. The whole process not only takes high energy consumption and long process time, but produces the crystal ingot of great value. For more details, see Zhang et al. (2014), Sun et al. (2016) and Jin et al., (2019). In this work, the focus is on stage three, i.e., the body growth stage. The objective in the body growth step is to control the ingot diameter such that the diameter is not smaller than the target and stay as close as possible to the target. If the true diameter is smaller than the target, the whole crystal ingot becomes useless, causing a huge waste. During the crystal ingot growth, the effects of process variables, such as the pulling speed and the power of the heater, on the quality response evolve over time. One of the key reasons for these dynamic effects is that the length of the crystal ingot is growing. Moreover, the equipment conditions are degrading, which is inevitable due to the deposition of byproducts on the heaters during the experiment. However, these dynamic effects of process variables on the quality response is not well studied.

In the literature, the dynamic linear model (DLM, West and Harrison 1997) and the varying coefficient model (VCM, Cleveland et al. 1991, Hastie and Tibshirani 1993, Fan and Zhang 2008) are both important tools to explore the dynamic effects of variables in the dynamic linear regression. The DLM assumes that the coefficients are not constant but rather evolving over time (Petris et al. 2009). A limitation of DLM is that it assumes a linear relationship between states in the latent state equation and thus the latent state process follows a Gaussian distribution. The VCM assumes that the coefficient functions are smooth and thus may not detect discontinuities or sudden structure changes in dynamic coefficient problems

In Chapter 3, we propose a penalized dynamic logistic regression to better understand the dynamic effects of process variables over time. We use a combination of the fused lasso penalty and the l_2-norm group lasso penalty to characterize the effects of process variables.

1.4 Structured Variable Selection in Experimental Designs

Provided a two-level experimental design where A, B, C, and D are the four factors, the main effects (MEs) and the two-factor interactions (2FIs) are defined as

$$
\begin{aligned}
ME(A) = \bar{y}(A+) - \bar{y}(A-) &= \frac{1}{2}(\bar{y}(A+|B+) + \bar{y}(A+|B-)) - \frac{1}{2}(\bar{y}(A-|B+) + \bar{y}(A-|B-)),\\
ME(B) = \bar{y}(B+) - \bar{y}(B-) &= \frac{1}{2}(\bar{y}(B+|A+) + \bar{y}(B+|A-)) - \frac{1}{2}(\bar{y}(B-|A+) + \bar{y}(B-|A-)),\\
INT(A,B) &= \frac{1}{2}(\bar{y}(A+|B+) + \bar{y}(A-|B-)) - \frac{1}{2}(\bar{y}(A+|B-) + \bar{y}(A-|B+)),\\
INT(A,B) &= \frac{1}{2}(\bar{y}(B+|A+) + \bar{y}(B-|A-)) - \frac{1}{2}(\bar{y}(B+|A-) + \bar{y}(B-|A+)),
\end{aligned}
$$

where $\bar{y}(A+)$, $\bar{y}(A-)$, $\bar{y}(B+)$, $\bar{y}(B-)$ are the averages of the response y at the level settings $A+, A-, B+$, and $B-$, respectively, and $\bar{y}(A+|B+)$, (A-|B-), (A+|B-), and (A-|B+) are the averages of y at the level settings $A+B+$, $A-B-$, $A+B-$, and $A-B+$, respectively.

The interaction terms are not straightforward for explanations in practice. For example, in the genomics, one particular question of interest is which gene is conditionally active. Wu (2018) proposed to reparametrize the effects such that the interaction terms can be represented in a more efficient way. He develops the concepts called conditional main effect (CME), which is defined as the conditional effect of one factor at a fixed level of another factor, such as CME($A|B+$) In the CME analysis (Su and Wu 2017), the CME is intended to provide a better understanding of traditional interaction terms. Given the factors A and B, their interaction term, INT(AB), can be reformulated as four CMEs, CME($A|B+$), CME($A|B-$), CME($B|A+$), and CME($B|A-$), under the relationship INT(AB) $= \frac{1}{2}\left(\text{ CME}(A|B+) - \text{CME}(A|B-)\right) = \frac{1}{2}\left(\text{ CME}(B|A+) - \text{CME}(B|A-)\right)$. Table 1.1 shows the CMEs constructed with the MEs A, B, C, and D.

It is challenging to perform variable selection within the CME framework. First, the CME reformulation increases the dimension of predictors from $p + \binom{p}{2}$ to $p + 4\binom{p}{2}$, leading to high dimension problems even when the number of main effects p is moderately large. Second, there exist two underlying group structures: (1) Siblings, defined as the CMEs with the same parent effect, such as CME($A|B+$) and CME($A|C+$). (2) Cousins, defined as the CMEs with same conditional effects, such as CME($A|C+$) and CME($B|C+$). Third, there is additional hierarchical structure between the CME and its parent effect.

Mak and Wu (2019) proposed a bi-level selection method based on the group MCP method (Zhang 2010). However, their approach treated the sibling and cousin groups separately. Given the overlapping feature among sibling and cousin groups, in Chapter 4 we propose the hierarchical sparse overlapping group selection method in the framework of best subset selection problems to perform the variable selection. We view the proposed sparse Ridge regression problem from the experimental designers' point of view. We employ the expectation-maximization algorithm to formulate the best subset selection problem as an iterative linear

Table 1.1: Experimental design with main factors A, B, C, D, and CMEs.

A	B	C	D	A|B+	A|B-	A|C+	A|C-	A|D+	A|D-	B|A+	B|A-	B|C+	B|C-	B|D+	B|D-	C|A+	C|A-	C|B+	C|B-	C|D+	C|D-	D|A+	D|A-	D|B+	D|B-	D|C+	D|C-
-1	-1	-1	-1	0	-1	0	-1	0	-1	0	-1	0	-1	0	-1	0	-1	0	-1	0	-1	0	-1	0	-1	0	-1
1	-1	-1	1	0	1	0	1	1	0	-1	0	0	-1	-1	0	-1	0	0	-1	-1	0	1	0	0	1	0	1
-1	1	-1	1	-1	0	0	-1	-1	0	0	1	0	1	1	0	0	-1	-1	0	-1	0	0	1	1	0	0	1
1	1	-1	-1	1	0	0	1	0	1	1	0	0	1	0	1	-1	0	-1	0	0	-1	-1	0	-1	0	0	-1
-1	-1	1	1	0	-1	-1	0	-1	0	0	-1	-1	0	-1	0	0	1	0	1	1	0	0	1	0	1	1	0
1	-1	1	-1	0	1	1	0	0	1	-1	0	-1	0	0	-1	1	0	0	1	0	1	-1	0	0	-1	-1	0
-1	1	1	-1	-1	0	-1	0	0	-1	0	1	1	0	0	1	0	1	1	0	0	1	0	-1	-1	0	-1	0
1	1	1	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

integer optimization (LIO) problem.

1.5 Brief Review of Variable Selection Techniques

In this section, we briefly review the variable selection techniques used in the proposal. They are the nonnegative garrote (nng) method, least absolute shrinkage and selection operator (lasso), group lasso, and fused lasso.

The idea of the nonnegative garrote technique (Breiman 1995; Yuan et al. 2007; Xiong 2010) is to scale the coefficient by a nonnegative scaler such that the variable selection is performed. Given an initial estimator $\hat{\theta}_j$ and a value for the tuning parameter s, the nng automatically selects a set of scalers from the optimization problem:

$$\begin{aligned} \underset{\boldsymbol{\theta} \in R^p}{\text{minimize}} \ & \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} c_j \hat{\theta}_j x_{ij})^2 \\ s.t. \ \ & c_j \geq 0, \ j = 1, \cdots, p, \\ & \sum_j c_j \leq s, \end{aligned}$$

where c_j is the nonnegative scaler. If $c_j = 0$, then the variable x_j is not selected. If $c_j \neq 0$, then the coefficient of the variable x_k would be scaled. The final estimator is $c_j \hat{\theta}_j$.

The nonnegative garrote problem is essentially a quadratic programming problem. Cantoni et al. (2011) applied the nng method to perform variable selection in nonparametric additive models. Sun et al. (2016) proposed a hierarchical nng to select significant functional process variables in a logistic regression.

The lasso method (Tibshirani 1996) solves the following optimization problem:

$$\underset{\boldsymbol{\theta} \in R^p}{\text{minimize}} \ -l(\boldsymbol{X}, \boldsymbol{\theta}) + \lambda ||\boldsymbol{\theta}||_1,$$

where $l(\boldsymbol{X}, \boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - X_i\theta)^2$ for the quadratic loss function, λ is the regularization parameter, and $\boldsymbol{\theta}$ is the coefficient vector. Compared to the nng, it is clear that the lasso has smaller number of constraints in the optimization problem.

The l_1-norm penalizes the loss function by the sum of the absolute values of the coefficients. The lasso can be applied in the case when $p > n$ and encourages sparse solutions with many coefficients equal to zero. The number of non-zero coefficients is at most $\min(n, p)$. Intensive work on lasso have been published with different types of loss functions $l(\boldsymbol{X}, \boldsymbol{\theta})$ including hinge loss function (Zhu et al. 2003), Gaussian graphical model loss function (Friedman et al. 2007). The lasso estimate can be efficiently solved by the coordinate descent algorithm (Friedman et al. 2007; 2009). However, there are also drawbacks for lasso. The lasso estimators are shrunk and thus biased. In addition, lasso does not necessarily yield good results in the presence of high collinearity.

The group lasso (Yuan and Lin 2006) considers cases where variables have a natural group structure such that the variable selection is performed at the group level. Group lasso solves the following optimization problem:

$$\underset{\boldsymbol{\theta} \in R^p}{\text{minimize}} \ -l(\boldsymbol{X}, \boldsymbol{\Theta}) + \lambda ||\boldsymbol{\Theta}||_2,$$

where $l(\boldsymbol{X}, \boldsymbol{\Theta}) = \sum_{i=1}^{n} (y_i - X_i\theta)^2$ for the quadratic loss function. $\boldsymbol{\Theta}$ represents a group of p_j variables. When a group is selected by the group lasso, all featurs within that group is selected. When the sizes of groups are all equal to one, the group lasso reduces to the regular lasso problem.

The fused lasso penalizes the differences between consecutive coefficients (Tibshirani et al. 2005). It encourages the sparsity of their differences. Fused lasso solves the optimization problem:

$$\underset{\boldsymbol{\theta} \in R^p}{\text{minimize}} \; -l(\boldsymbol{X}, \boldsymbol{\theta}) + \lambda \sum_{j=1}^{p} |\theta_j - \theta_{j-1}|,$$

where λ is the penalization parameter.

Fused lasso term is also known as the total variation (TV) denoising (Rudin et al. 1992). Kim et al. (2009) generalizes the fused lasso term into higher orders, i.e., the so-called trend filtering. For example, the fused lasso is the 0th order trend filtering (Tibshirani 2014) as

$$\sum_{j=1}^{p} |\theta_j - \theta_{j-1}| = ||D\theta||_1,$$

where

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ & & & & \dots\dots\dots\dots\dots & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & -1 \end{bmatrix}_{(p-1)\times p},$$

where $\boldsymbol{D}$ is known as the difference operator. And the 1st order trend filtering term is

$$\sum_{j=1}^{p-1} |\theta_{j-1} - 2\theta_j + \theta_{j+1}|.$$

The difference between trend filtering and splines are that the trend filtering is defined on

the discrete set of inputs, while the splines are defined over the continuous domains. The smoothing splines penalize the sum of squared derivatives across input points, and the locally adaptive regression splines penalize the total variation on the kth derivative.

The trend filtering has better performance when the underlying function display varying levels of smoothness at different spatial locations. Tibshirani (2014) showed that when the trend filtering order is 0 or 1, the solutions from the trend filtering and locally adaptive regression splines are the same. The computation of trend filtering is a bit slow compared with the smoothing spline, but faster than the locally adaptive regression splines. (Tibshirani 2014). The primal-dual interior point method for the fused problem with the fused term can handle problems of size on the order of n = 1,000,000.

1.6 Outline of the Dissertation

The focus of this dissertation is on statistical modeling and analysis by incorporating various effects of predictors as functional coefficients. The model performance is enhanced by appropriate variable selection techniques, especially the nonnegative garrote method, the fused lasso method, the group lasso method, and the structured best subset selection method. The proposal consists of three parts:

1) A study on how to model the mixture-of-mixtures experiments with focus on the unique two-level structure of predictor variables;

2) A proposal of the penalized dynamic logistic regression model to characterize dynamic effects of process variables;

3) An investigation of the sparse Ridge regression method to address the variable selection issues within the best subset selection framework.

The rest of this dissertation is organized as follows. In Chapter 2, an additive heredity model is proposed for analyzing the mixture-of-mixtures (MoM) experiment. The proposed model considers an additive structure to inherently connect the major components with the minor components. To enable a meaningful interpretation for the estimated model, we apply the hierarchical and heredity principles by using the nonnegative garrote technique for model selection. In Chapter 3, a regularized dynamic logistic regression model is proposed for estimating the dynamic coefficients of process variables. The group lasso penalty term is applied to enable a sparse and interpretable solution. The fused lasso penalty term is used for automatically detecting the change-points in the dynamic coefficients. In Chapter 4, a structured variable selection method within the best subset selection framework is proposed. The key idea of the proposed method is to re-construct the regression matrix in the angle of experimental designs. We employ the estimation-maximization algorithm to formulate the best subset selection problem as an iterative linear integer optimization (LIO) problem. A slightly modified version of Chapter 2 has been publised as a paper in Technometrics. Chapters 3 and 4 are based on two working manuscripts. Chapter 5 concludes the dissertation with the plan for future works.

Chapter 2 Additive Heredity Model for the Analysis of Mixture-of-Mixtures Experiments

2.1 Introduction

In the mixture experiment (Cornell 2002), the typical design variables are the proportions of mixture components in a blend, with their summation equal to unity. As a special case of mixture experiments, the mixture-of-mixtures (MoM) experiment considers the case that each mixture component (so-called major component) is also made up by a mixture of sub-components (so-called minor components). Clearly, the proportions of both major and minor components are varied (Cornell and Ramsey 1998; Piepel 1999) in a constraint experimental region. Such an inherent structure of experiments poses an intriguing challenge to appropriate modeling and analysis of MoM experiments. There can be other variables, such as process variables and the total amount of a blend, to be considered in MoM experiments. In this work, we mainly focus on the analysis of MoM experiments with design variables only including the proportions of major components and their corresponding portions of minor components.

Mixture experiment is one of the classic topics in the design and analysis of experiments area. It has been applied widely in food, medicine, and chemistry industries. MoM experiments,

as a special case of mixture experiments, also frequently appears in applications. See Piepel (1999), Dingstad et al. (2003), Borges et al. (2007), Didier et al. (2007), and Coetzer and Haines (2013) for interesting MoM case studies. One of the recent MoM examples was studied by Kang, Joseph, and Brenneman (2011). The MoM experiment was conducted to formulate a new kind of Pringles® potato crisp, whose package form is changed from a can to a bag. There are three major components A, B, and C, whose proportion are denoted by c_1, c_2, and c_3. The major component A is composed of two minor components A_1 and A_2 with proportion x_{11} and x_{12} with respect c_1, and B is composed of two minor components B_1 and B_2 with proportion x_{21} and x_{22} with respect c_2. Component C is pure material which can be considered to have only a single minor component. The hardness of the potato chip (Hardness) and the percentage of fat (% Fat) are the two response variables. More details can be found in Kang, Joseph, and Brenneman (2011).

Various regression models have been proposed in the literature to analyze mixture experiments, such as the Scheffé model (Scheffé, 1958), the Cox model (Cox, 1971), the slack-variable model (Snee and Rayner, 1982; Khuri, 2005; Kang et al., 2016), the Kronecker model (Draper and Pukelsheim, 1998; Prescott et al., 2002), the component-slope-linear model (Piepel, 2007), and a general blending model by Brown, Donev, and Bissett (2015). The idea of the Scheffé model has been extended to the multiple-Scheffé model (Lambrakis 1968, 1969; Cornell and Ramsey 1998) for analyzing MoM experiments. The multiple-Scheffé model considers a product of the Scheffé models for both major and minor components to model the major-minor interactions. However, such a modeling strategy involves a large number of interaction terms and needs a large number of experimental observations. Another limitation is that the minor components can still be in the model even when its major component is absent in the model. To address limitations of the multiple-Scheffé model, Kang et al. (2011) developed a so-called major-minor model to analyze MoM experiments.

Their key idea is to use the Scheffé model to capture the relationship between the mean response and the major components, while the coefficients of major components and their interactions are modeled as a function of their respective minor components. The major-minor model can have a smaller model size than the multiple-Scheffé model. Furthermore, when the major component is absent, all of its corresponding minor components are absent from the major-minor model. On the other hand, the complexity of the major-minor model depends on the degrees of the Scheffé models in the levels of major and minor components. When a linear Scheffé model is used for both major and minor components, the major-minor model does not have interactions between the major components and between the minor components. When a quadratic Scheffé model is assumed for both major and minor components, the resultant terms can be too complex to interpret because of some high-order interaction terms.

To address these limitations, we propose an additive heredity model to better capture the major-minor structure of MoM experiments with meaningful interpretation. Similar to the major-minor model, the additive heredity model also assumes two Scheffé models for the major and minor components respectively. But different from the major-minor model, the two models are added together to form the final heredity model. The detailed development is shown in Section 2. The additive model is simple and paves an easy way to study the contribution of each design variable in the model. Moreover, by imposing the coefficients of the minor components to be functions of their respective major components, the minor components exist only when the corresponding major component is presented in the model. The nonnegative garrote technique (Breiman 1995; Yuan et al. 2007; Xiong 2010) is used in the model estimation to enable the hierarchical and heredity principles for major and minor components. The proposed additive heredity model has several advantages. First, it provides a meaningful model interpretation of the major-minor structure such that the

coefficients of minor components are dependent upon the corresponding major components. In MoM experiments, the minor components will be presented in the model only if their corresponding major component is included in the model (Yuan and Lin 2009; Kang et al. 2011). Second, it can explicitly quantify the contributions of the individual major and minor components through the additive form. Third, the additive heredity model can control the model complexity via variable selection, which is achieved by the nonnegative garrote method. Furthermore, the additive structure is flexible to include terms of interest based on practitioners' objectives.

The rest of chapter is organized as follows. In Section 2, we detail the proposed additive heredity model. In Section 3, we present the estimation procedures by the heredity constrained nonnegative garrote method. The simulation and real case studies are conducted in Sections 4 and 5. We conclude this work with some discussion in Section 6.

2.2 Additive Heredity Model

In an MoM experiment, assume that there are q major components, and let c_k be the proportion of the kth major component such that

$$\sum_{k=1}^{q} c_k = 1, \quad 0 \leq c_k \leq 1, \quad k = 1, \ldots, q. \tag{2.1}$$

Moreover, each major component is composed of m_k minor components, whose proportions with respect to c_k are x_{kl},

$$\sum_{l=1}^{m_k} x_{kl} = 1, \quad 0 \leq x_{kl} \leq 1, \quad l = 1, \ldots, m_k. \tag{2.2}$$

To flexibly quantify the effects of major and minor components on the response y, we consider an additive modeling strategy to incorporate the major and minor structure relations in the model. We propose an additive heredity model (AHM) as:

$$y = \sum_{k=1}^{q} \gamma_k c_k + \sum_{k<j} \gamma_{kj} c_k c_j + \sum_{k=1}^{q} \sum_{l=1}^{m_k} \delta_l^{(k)}(c_k) x_{kl} + \sum_{k=1}^{q} \sum_{l<l'} \delta_{l,l'}^{(k)}(c_k) x_{kl} x_{kl'} + \epsilon, \text{ with } \epsilon \sim N(0, \sigma^2), \tag{2.3}$$

where γ_k is the coefficient for the major component proportion c_k, γ_{kj} is the coefficient for the interactions between c_k and c_j. The $\delta_l^{(k)}(c_k)$ is denoted as the coefficient for the minor component proportion x_{kl}, and $\delta_{l,l'}^{(k)}(c_k)$ is denoted as the coefficient for the interaction between x_{kl} and $x_{kl'}$. To ensure the contribution of the minor component to the response dependent upon the corresponding major component, we consider $\delta_l^{(k)}(c_k)$ to be a function of the major component c_k. A monotonic and bounded mapping from the major component c_k to $\mathbb{R}$ is a proper choice for $\delta_l^{(k)}(c_k)$ because the larger the c_k is, intuitively, the more influential the minor components of this major component should be to the response of the whole mixture. Under this consideration, we consider to use a power function as

$$\delta_l^{(k)}(c_k) = \zeta_l^{(k)} c_k^h,$$

where $\zeta_l^{(k)}$ is the coefficient and h is the power parameter. Similarly we consider $\delta_{l,l'}^{(k)}(c_k) = \zeta_{l,l'}^{(k)} c_k^{2h}$ with $\zeta_{l,l'}^{(k)}$ being the coefficient. Clearly, the power function c_k^h is bounded on the domain of c_k. The hyperparameter h is the power index of c_k and set in the range of $(0, 2)$. For the c_k in $(0, 1)$, c_k^h is decreasing with respective to h. Thus, the larger the h is, the less role the major component would play in the minor components' effects, including $\delta_l^{(k)}(c_k) x_{kl}$ and $\delta_{l,l'}^{(k)}(c_k) x_{kl} x_{kl'}$. The hyperparameter h is estimated from the data via cross-validation as described in Algorithm 1. We set the upper bound of h to be 2, which is shown to be

sufficient in our study. Readers can choose any values that are larger than zero based on their understanding of the major components' influence.

The additive heredity model specifies the effects of major and minor components in an additive form. Moreover, the AHM has a flexible major-minor structure relationship by assuming the coefficient of minor components varies as a function of a major component. This varying-coefficient property is useful to flexibly accommodate certain dependence relationships between the major and minor components. The AHM thus can provide a meaningful model interpretation of the major-minor structure. Lastly, various Scheffé models of appropriate order are applicable in the AHM framework. For example, a quadratic Scheffé model for the major component and a linear Scheffé model for the minor component. We choose the quadratic Scheffé model for both major and minor components in this study because we are interested in both the main effects and the two-factor interactions of the major components, the main effects of the minor components, and the two factor interactions of the minor components from the same major component.

2.2.1 Connection to the Major-Minor Model

In this section, we make a connection between the proposed additive heredity model and the major-minor model in Kang et al. (2011). The major-minor model considers the coefficients of major components as the Scheffé model on the corresponding minor components in order to incorporate the major-minor structure relations. Take the example of having two major components c_1 and c_2, each with two minor components x_{k1}, x_{k2}, $k = 1, 2$. Assuming the quadratic Scheffé model for both the major and minor components, the major-minor model

is expressed as

$$\begin{aligned} y =&(\gamma_1 x_{11} + \gamma_2 x_{12} + \gamma_3 x_{11} x_{12})c_1 + (\gamma_4 x_{21} + \gamma_5 x_{22} + \gamma_6 x_{21} x_{22})c_2 \\ &+ (\gamma_7 x_{11} x_{21} + \gamma_8 x_{11} x_{22} + \gamma_9 x_{12} x_{21} + \gamma_{10} x_{12} x_{22} + \gamma_{11} x_{11} x_{12} x_{21} \\ &+ \gamma_{12} x_{11} x_{12} x_{22} + \gamma_{13} x_{11} x_{21} x_{22} + \gamma_{14} x_{12} x_{21} x_{22} + \gamma_{15} x_{11} x_{12} x_{21} x_{22}) c_1 c_2 + \epsilon. \end{aligned} \tag{2.4}$$

Therefore, when a major component proportion $c_k = 0$, the corresponding minor components no longer exist in the model. As a major component's proportions increase, the contribution from the corresponding minor components also increases.

Nevertheless, the use of the Scheffé model on minor components as the coefficients for the corresponding major component is not the only way to represent the major-minor structure relations. For instance, we can use a so-called minor-major model at the minor level to describe the relationship between the response, the major, and the minor components. That is, the coefficients of minor components are a function of the respective major components, indicating that the contribution from minor components is dependent upon their major components. Assuming the quadratic Scheffé model for minor components, the minor-major model can be expressed as

$$\begin{aligned} y =& (\phi_1 + \phi_2 c_1) x_{11} + (\phi_3 + \phi_4 c_1) x_{12} + (\phi_5 + \phi_6 c_1 + \phi_7 c_1^2) x_{11} x_{12} \\ &+ (\phi_8 + \phi_9 c_2) x_{21} + (\phi_{10} + \phi_{11} c_2) x_{22} + (\phi_{12} + \phi_{13} c_2 + \phi_{14} c_2^2) x_{21} x_{22} \\ &+ (\phi_{15} c_1 + \phi_{16} c_2 + \phi_{17} c_1 c_2) x_{11} x_{21} + (\phi_{18} c_1 + \phi_{19} c_2 + \phi_{20} c_1 c_2) x_{11} x_{22} \\ &+ (\phi_{21} c_1 + \phi_{22} c_2 + \phi_{23} c_1 c_2) x_{12} x_{21} + (\phi_{24} c_1 + \phi_{25} c_2 + \phi_{26} c_1 c_2) x_{12} x_{22} + \epsilon, \end{aligned} \tag{2.5}$$

where the coefficients of minor components are assumed to be dependent only upon their corresponding major component.

Note that the major-minor model in (2.4) contains terms that are complicated and difficult to interpret, such as $c_1 c_2 x_{11} x_{21} x_{22}$. In contrast, such terms do not appear in the in the AHM (2.3) when a quadratic model is used for both minor and major components. Through heredity principle incorporated in nonnegative garrote, the AHM can control the model complexity. We also like to remark that the major-minor model can be rewritten in the form of additive models. For example, if we assume the linear Scheffé model for both major and minor components, the major-minor model is equivalent to the AHM assuming $\delta_l^{(k)}(c_k) = \zeta_l^{(k)} c_k$, $\delta_{l,l'}^{(k)}(c_k) = 0$, and $\gamma_{kj} = 0$ in model (2.3). In the supplemental materials, we show that any major-minor model can be expressed in some forms of additive models.

2.3 Model Estimation

To estimate the parameters in the proposed additive heredity model, we employ the nonnegative garrote method (Breiman 1995; Yuan et al. 2007; Xiong 2010) to pursue a parsimonious and structured model. The nonnegative garrote estimate of a parameter is expressed as $\theta^{nng} = \theta^{(0)} \alpha_s$, where $\theta^{(0)}$ is the initial estimate and $\alpha_s \geq 0$ is a nonnegative scaling factor. The key idea of the nonnegative garrote method is to scale the initial estimates via scaling factors. One feature of the nonnegative garrote method is the flexibility to adapt the hierarchical and heredity principles in the form of linear constraints (Yuan and Lin, 2007). The hierarchical principle between the major and corresponding minor components is to require the minor components being present only if the corresponding major component is present in the model. The heredity principle requires that the interaction terms can appear in the model only if one of its main effects appeared in the model. By imposing such principles, it can make the proposed model more meaningful and interpretable. Besides the nonnegative garrote method, LARS method can also be modified to incorporate the heredity principle, as shown in Yuan, Joseph, and Lin (2007). But Yuan, Joseph, and Zou (2009) pointed out

that the nonnegative garrote method is more efficient in computation and much more flexible and easier to adopt any kind of constraints between the effects. For the proposed additive model, the hierarchical and heredity principle is more complex than the regular regression model and thus we choose the nonnegative garrote method.

The nonnegative garrote method with the weak heredity principle can be expressed in (2.6). The response y is a simplified notation y_i without the observation index i. Let n be the total number of observations in the MoM experiment. The γ_k^{init}, γ_{kj}^{init}, $(\zeta_l^{(k)})^{init}$, $(\zeta_{l,l'}^{(k)})^{init}$ are the initial estimates of the parameters γ_k, γ_{kj}, $\zeta_l^{(k)}$, and $\zeta_{l,l'}^{(k)}$, respectively. We denote α_k to be the scaling factor for the major component c_k, α_{kj} to be the scaling factor for the interaction between major components c_k and c_j, $\beta_l^{(k)}$ to be the scaling factor for the minor component l within the major component k, and $\beta_{l,l'}^{(k)}$ the scaling factor for the interaction between minor components l and l'. The constrained optimization of estimating parameters is given as follows:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ y - \left\{ \sum_{k=1}^{q} \gamma_k c_k + \sum_{k<j} \gamma_{kj} c_k c_j + \sum_{k=1}^{q} \sum_{l=1}^{m_k} \delta_l^{(k)}(c_k) x_{kl} + \sum_{k=1}^{q} \sum_{l<l'} \delta_{ll'}^{(k)}(c_k) x_{kl} x_{kl'} \right\} \right\}^2, \tag{2.6}$$

$$\begin{aligned}
s.t.\ & \gamma_k = \gamma_k^{init}\alpha_k,\ \gamma_{kj} = \gamma_{kj}^{init}\alpha_{kj}; \delta_l^{(k)}(c_k) = (\zeta_l^{(k)})^{init} c_k^h \beta_l^{(k)},\ \delta_{ll'}^{(k)}(c_k) = (\zeta_{ll'}^{(k)})^{init} c_k^{2h} \beta_{l,l'}^{(k)}; \\
& \alpha_k \geq 0, \alpha_{kj} \geq 0, \beta_l^{(k)} \geq 0, \beta_{l,l'}^{(k)} \geq 0; \\
& \sum_{k=1}^{q} \left\{ \sum_{j=1,k<j}^{q} (\alpha_k + \alpha_{kj}) + \sum_{l<l'}^{m_k} (\beta_l^{(k)} + \beta_{l,l'}^{(k)}) \right\} \leq M; \\
& \beta_l^{(k)} \leq \alpha_k; \beta_{l'}^{(k)} \leq \alpha_k; \\
& \alpha_{kj} \leq \alpha_k + \alpha_j, \text{ for } k \neq j; \qquad (2.6a) \\
& \beta_{l,l'}^{(k)} \leq \beta_l^{(k)} + \beta_{l'}^{(k)}, \text{ for } l \neq l', \qquad (2.6b)
\end{aligned}$$

where the constraints (2.6a) and (2.6b) are specially applied when the weak heredity principle

is assumed. The weak and strong heredity principles select the interaction terms based on their parent terms. The weak heredity only requires one of its parent terms to be significant. For example, in (6a), α_{kj} will be forced to be zero unless at least one of α_k and α_j is strictly positive. This reflects the weak heredity principle, as it indicates $c_k c_j$ would be significant if one of its parent terms c_k and c_j is significant. The strong heredity selects a two-factor interaction only if both its parent terms are significant. When the strong heredity principle is assumed, these two constraints (6a) and (6b) can be replaced by

$$\alpha_{kj} \leq \alpha_k;\ \alpha_{kj} \leq \alpha_j,\ \text{for } k \neq j;$$
$$\beta^{(k)}_{l,l'} \leq \beta^{(k)}_{l};\ \beta^{(k)}_{l,l'} \leq \beta^{(k)}_{l'},\ \text{for } l \neq l'.$$

For example, α_{kj} would be strictly positive if both α_k and α_j are larger than 0, and accordingly, $c_k c_j$ is significant if both c_k and c_j are significant. If no heredity principle is assumed, the two constraints (6a) and (6b) can be removed. Here M is a tuning parameter to control the general sparsity in the model. Note that the objective in the above optimization can be expressed as $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\eta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\eta})$, where $\boldsymbol{y} = (y_1, \ldots, y_n)$ is the response vector, $\boldsymbol{\eta} = (\alpha_1, \ldots, \beta_{q-1,q})'$ is the parameter vector containing all scaling factors α and β, and $\boldsymbol{X}$ is the corresponding regression matrix. We adopt the generalized cross-validation (GCV) for finding an optimal value of M, which is given by:

$$\text{GCV} = \frac{(\hat{\boldsymbol{y}} - \boldsymbol{y})^T(\hat{\boldsymbol{y}} - \boldsymbol{y})}{n(1 - tr(\boldsymbol{H})/n)^2},$$

where $\boldsymbol{H}$ is the hat matrix, and $tr(\boldsymbol{H}) = tr(\boldsymbol{X} diag(\hat{\boldsymbol{\eta}})(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) = tr(diag(\hat{\boldsymbol{\eta}}))$. Good justification of using GCV for tuning parameter selection in the nonnegative garrote problem can be found in Xiong (2010).

The computational algorithm for model estimation (2.6) is summarized in Algorithm 1.

Algorithm 1

Input: Data
for a sequence of h **do**
  Obtain initial estimators $\gamma_k^{init}, \gamma_{kj}^{init}, (\delta_l^{(k)})^{init}$, and $(\delta_{ll'}^{(k)})^{init}$ from ridge estimators of (2.3).
  for a sequence of M **do**
    Solve the constrained optimization problem (6) and save the scaling factors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.
    Compute the GCV value at each M.
  end for
  $M_{sel} =_M \text{GCV}(M)$
  Compute the mean squared cross validation (MSCV), defined later, at M_{sel}.
end for
$h_{sel} =_h \text{MSCV}(h)$
Output: the estimated model and MSCV at h_{sel}

Because of the nonnegative garrote method with proper constraints, the model size of the AHM can be much smaller compared to the multiple-Scheffé model. It is worth pointing out that the performance of the nonnegative garrote estimate relies on the choice of the initial estimate. Yuan and Lin (2007) argued that the nonnegative garrote method can be used with initial estimators from the ridge regression, LASSO, and the elastic net. Here we use the ridge regression estimate with the regularization parameter λ determined by the leave-one-out cross-validation for the initial estimators of the nonnegative garrote method. Other than leave-one-out cross-validation, the user can also consider using other criteria such as AICc (Draguljić et al., 2014), especially when the data are collected from designed experiments with very limited runs.

2.4 Simulation

In this section, we evaluate the performance of the proposed additive heredity model in both unconstrained and constrained MoM experiments. We consider two different types of MoM

experiments: (a) each major component contains the same number of minor components and (b) each major component contains a different number of minor components, and one major component has a single minor component. Without loss of generality, for both (a) and (b) we assume there are only three major components, i.e., c_1, c_2 and c_3. In type (a), each major component c_k has two minor components x_{k1}, x_{k2}. In type (b), the numbers of minor components corresponding to c_1,c_2 and c_3 are three, two and one, respectively. The simulation results for (a) are shown in this section. The results for (b) are in Supplement due to space limitations.

In case (a), there are five underlying models to be considered for generating the data:

$$
\begin{aligned}
I: \ y =& 10c_1 + 30c_2 + 20c_3 + 18c_1c_2 + \epsilon, \\
II: \ y =& 15c_1x_{11} + 12.5c_1x_{12} + 22.5c_2x_{21} + 20c_2x_{22} + 15c_3x_{31} + 17.5c_3x_{32} + \epsilon, \\
III: \ y =& 10c_1 + 30c_2 + 20c_3 + 15c_1^h x_{11} + 27.5c_2^h x_{21} + \epsilon, \text{ where } h = 0.5, \\
IV: \ y =& 10c_1 + 30c_2 + 20c_3 + 25c_2x_{11} + 22.5c_3x_{21} + \epsilon, \\
V: \ y =& 10c_1 + 30c_2 + 20c_3 + 7c_2c_3 + 13.75c_1^2x_{11}x_{12} + \epsilon,
\end{aligned}
$$

where the noise $\epsilon \sim N(0, \sigma^2)$ is independent of the component proportions. The noise variance is chosen such that the signal-to-noise (SN) ratio (Wu and Hamada 2009) is three.

One benchmark method for comparison is the multiple-Scheffé model. Assuming the linear Scheffé model for both major and minor components, the corresponding multiple-Scheff*é* model is

$$
y = (\alpha_1c_1 + \alpha_2c_2 + \alpha_3c_3) \times \prod_{k=1}^{3} (\beta_{k1}x_{k1} + \beta_{k2}x_{k2}) + \epsilon,
$$

which contains 24 regression coefficients. Other methods used in comparison include the

major-only linear Scheffé model (2.7) and the major-only quadratic Scheffé model (2.8),

$$y = \gamma_1 c_1 + \gamma_2 c_2 + \gamma_3 c_3 + \epsilon, \tag{2.7}$$

$$y = \gamma_1 c_1 + \gamma_2 c_2 + \gamma_3 c_3 + \gamma_4 c_1 c_2 + \gamma_5 c_1 c_3 + \gamma_6 c_2 c_3 + \epsilon. \tag{2.8}$$

Both (2.7) and (2.8) naively ignores the information on the minor components. We also make comparison with the major-minor models (Kang et al. 2011) assuming the linear Scheffé model for minor components, linear or quadratic Scheffé model for major components, defined in (2.9) and (2.10) respectively.

$$y = (\gamma_1 + \gamma_2 x_{11}) c_1 + (\gamma_3 + \gamma_4 x_{21}) c_2 + (\gamma_5 + \gamma_6 x_{31}) c_3 + \epsilon, \tag{2.9}$$

$$\begin{aligned} y = {} & \gamma_1 c_1 + \gamma_2 c_2 + \gamma_3 c_3 + \gamma_4 x_{11} c_1 + \gamma_5 x_{21} c_2 + \gamma_6 x_{31} c_3 + \gamma_7 c_1 c_2 + \gamma_8 c_1 c_3 + \gamma_9 c_2 c_3 \\ & + \gamma_{10} x_{11} c_1 c_2 + \gamma_{11} x_{11} c_1 c_3 + \gamma_{12} x_{21} c_1 c_2 + \gamma_{13} x_{21} c_2 c_3 + \gamma_{14} x_{31} c_1 c_3 + \gamma_{15} x_{31} c_2 c_3 \\ & + \gamma_{16} x_{11} x_{21} c_1 c_2 + \gamma_{17} x_{11} x_{31} c_1 c_3 + \gamma_{18} x_{21} x_{31} c_2 c_3 + \epsilon. \end{aligned} \tag{2.10}$$

A summary list of compared models is here:

a. the multiple-Scheffé model (MultipleScheffe),

b. the major-only linear Scheffé model (MajorLinear),

c. the major-only quadratic Scheffé model (MajorQuad),

d. the major-minor model assuming the linear Scheffé model for the major components (1st_MM),

e. the major-minor model assuming the quadratic Scheffé model for the major components (2nd_MM),

f. the additive heredity model with weak heredity constraints (AHM).

To evaluate the model performances, we use the metrics including the R^2, the small-sample-size corrected version of Akaike information criterion (AICc), mean squared error (MSE), mean squared cross-validation (MSCV), normalized MSCV (MSCVnorm) and model size. The R^2, AICc and MSE measure the fitting performance of models. Note that R^2 is adapted for the Scheffé models that do not contain intercept. The AICc (Hurvich and Tsai 1989; Burnham and Anderson 2002; Draguljić et al. 2014) is calculated via AICc $= n\log(\frac{RSS}{n}) + \frac{2\tilde{p}n}{n-\tilde{p}-1}$, where RSS is the residual sum of squares, $\tilde{p}$ is the number of nonzero parameters. The MSCV is calculated via MSCV $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{-i})^2$, where $\hat{y}_{-i}$ is the prediction at the i^{th} input by the model fitted without the i^{th} data point. The smaller MSCV value indicates better prediction performance. The MSCVnorm is calculated by MSCVnorm $=$ MSCV$/(\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2)$. The AHM has a varied model size because of the nonnegative garrote technique employed.

2.4.1 Unconstrained MoM Experiments

The unconstrained mixture experiment is a typical situation where each component proportion can take any value in $[0, 1]$. For the major components, we consider two different designs: the I-optimal design (Laake 1975; Goos et al. 2016) and the maximin distance design (Johnson et al. 1990). For the minor components, because each major component has two minor components, we choose three design point levels: the two endpoints and the middle point in the domain. For example, the design points for the minor components, $(x_{11},\ x_{12})$, are $(1, 0)$, $(\frac{1}{2}, \frac{1}{2})$, and $(0, 1)$. We applied the idea of crossed design to combine the designs for the major and minor components (Cornell and Ramsey 1998; Dingstad et al. 2003; Kang et al. 2011). That is, corresponding to every treatment combination of the major components, all possible settings of the minor components are included in the design.

2.4.1.1 Using I-Optimal Design for Major Components

We first use the I-optimal design for the major components. The I-optimal criterion is to minimize the average prediction variance over the experimental region. For the simulation setting described above, an I-optimal design for a quadratic Sheffé model of three major components is the simplex-centroid design (Lambrakis 1968, 1969; Cornell 2002) containing three vertices, three middle points, and the overall centroid of the triangular constrained region by $c_1+c_2+c_3 = 1$, as shown in Figure 2.1 (A). Apparently, the design for two minor components, containing $(1, 0)$, $(0, 1)$, and $(1/2, 1/2)$, is also a simplex-centroid for any mixture experiment with two components. Thus, the overall design is a crossed design of simplex-centroid for major components and simplex-centroid design for the minor components.

Table 2.1 shows the simulation results in terms of R^2, AICc, MSE, MSCV, MSCVnorm, and model size of the different models under comparison. The proposed AHM generally outperform the other models in prediction accuracy measured by MSCV and MSCVnorm for all simulation models but IV. For the simulation model I, which only contains the major components, the AHM has comparable prediction performance with the MajorQuad model. The simulation model II is essentially a linear Scheffé model disregarding the MoM structure because it is an additive model of all the terms $c_k x_{k1}$ and $c_k x_{k2}$, which are all the minor components proportions with respect to all the entire mixture. For this model, the AHM has competitive prediction performance comparable with 1st_MM, but better prediction performance than the MajorLinear and the MajorQuad. We also notice that, in the simulation model III containing the interactions between the major and corresponding minor components, both the AHM and the 2nd_MM have good prediction performance. For the simulation model IV, which contains the interaction terms between the major and non-corresponding minor components, the AHM and the 2nd_MM do not have as competitive

prediction performance as the MultipleScheffe model. One possible explanation is that the crossed-component interaction terms are not included in the proposed AHM in this study. The simulation model V contains one interaction term between c_2c_3 and one between minor components c_1x_{11} and c_1x_{12}, and the prediction performance of the AHM is best and close to that of the true model.

In terms of model fitting, the measure R^2, AICc and MSE values in Table 2.1 indicate that the AHM performs satisfactorily. The model size of the AHM varies across different settings because of the variable selection performed via the nonnegative garrote method but is often larger than that of 1st_MM but smaller than that of 2nd_MM.

Table 2.1: Performance comparisons of models under the unconstrained MoM experiment using I-optimal design for major components from 50 simulation replications (means and standard errors (in parenthesis)).

	Model	R^2	MSE	MSCV	MSCVnorm	AICc	Size
I	1st_MM	0.970 (0.002)	15.03 (0.82)	15.53 (0.85)	0.30 (0.02)	520.5 (10.4)	6.0 (0.0)
	2nd_MM	0.976 (0.001)	12.87 (0.36)	14.30 (0.47)	0.28 (0.02)	506.4 (5.3)	18.0 (0.0)
	AHM	0.975 (0.001)	12.63 (0.28)	13.19 (0.27)	**0.25** (0.02)	**490.7** (3.9)	8.4 (1.5)
	MajorLinear	0.970 (0.002)	15.06 (0.77)	15.30 (0.78)	0.30 (0.02)	517.5 (9.8)	3.0 (0.0)
	MajorQuad	0.975 (0.001)	12.86 (0.21)	13.28 (0.22)	0.26 (0.02)	**491.2** (3.1)	6.0 (0.0)
	MultipleScheffe	0.973 (0.002)	15.26 (0.91)	17.95 (1.25)	0.35 (0.02)	547.0 (11.3)	24.0 (0.0)
	TrueModel	0.974 (0.001)	12.88 (0.14)	13.15 (0.14)	0.25 (0.02)	489.2 (2.0)	4.0 (0.0)
II	1st_MM	0.988 (0.000)	16.55 (0.26)	17.09 (0.28)	**0.26** (0.02)	**538.9** (3.0)	6.0 (0.0)
	2nd_MM	0.989 (0.000)	16.46 (0.63)	18.29 (0.71)	0.28 (0.02)	552.9 (7.3)	18.0 (0.0)
	AHM	0.988 (0.000)	16.29 (0.46)	17.01 (0.46)	**0.26** (0.02)	**538.9** (4.9)	8.6 (1.4)
	MajorLinear	0.980 (0.001)	27.32 (1.81)	27.83 (1.86)	0.42 (0.03)	629.9 (12.6)	3.0 (0.0)
	MajorQuad	0.980 (0.001)	27.41 (1.85)	28.36 (1.92)	0.43 (0.03)	633.8 (12.8)	6.0 (0.0)
	MultipleScheffe	0.989 (0.000)	16.57 (0.56)	19.24 (0.79)	0.29 (0.02)	562.8 (6.4)	24.0 (0.0)
	TrueModel	0.988 (0.000)	16.55 (0.26)	17.09 (0.28)	0.26 (0.02)	538.9 (3.0)	6.0 (0.0)
III	1st_MM	0.923 (0.005)	158.35 (9.77)	163.47 (10.11)	0.30 (0.02)	965.4 (11.6)	6.0 (0.0)
	2nd_MM	0.939 (0.003)	134.00 (4.85)	148.63 (5.45)	0.28 (0.02)	949.2 (6.9)	18.0 (0.0)
	AHM	0.937 (0.003)	132.29 (3.49)	138.69 (3.22)	**0.26** (0.02)	**936.0** (4.7)	9.6 (1.3)
	MajorLinear	0.814 (0.011)	374.97 (22.70)	381.01 (23.02)	0.71 (0.03)	1125.0 (11.5)	3.0 (0.0)
	MajorQuad	0.823 (0.010)	363.30 (20.90)	375.21 (21.50)	0.70 (0.04)	1122.4 (11.0)	6.0 (0.0)
	MultipleScheffe	0.930 (0.005)	159.39 (11.02)	187.88 (14.03)	0.35 (0.03)	990.4 (13.0)	24.0 (0.0)
	TrueModel	0.935 (0.003)	133.35 (1.81)	136.92 (1.95)	0.25 (0.02)	932.1 (2.6)	5.0 (0.0)
IV	1st_MM	0.839 (0.013)	292.89 (23.77)	305.98 (24.80)	0.64 (0.03)	1081.4 (15.5)	6.0 (0.0)
	2nd_MM	0.859 (0.012)	273.34 (22.40)	300.51 (24.18)	0.62 (0.04)	1083.4 (15.7)	18.0 (0.0)
	AHM	0.843 (0.013)	285.97 (22.73)	298.10 (23.08)	0.62 (0.03)	1077.8 (14.8)	6.8 (1.3)
	MajorLinear	0.833 (0.013)	298.61 (24.06)	304.73 (24.55)	0.63 (0.03)	1081.7 (15.4)	3.0 (0.0)
	MajorQuad	0.834 (0.013)	301.35 (24.68)	312.20 (25.53)	0.65 (0.04)	1086.8 (15.6)	6.0 (0.0)
	MultipleScheffe	0.939 (0.003)	123.25 (4.68)	144.97 (5.78)	**0.30** (0.02)	**942.1** (7.2)	24.0 (0.0)
	TrueModel	0.932 (0.003)	122.56 (2.15)	125.93 (2.16)	0.26 (0.02)	916.2 (3.4)	5.0 (0.0)
V	1st_MM	0.963 (0.003)	19.05 (1.32)	19.66 (1.38)	0.44 (0.04)	565.1 (13.3)	6.0 (0.0)
	2nd_MM	0.966 (0.003)	18.89 (1.44)	20.24 (1.52)	0.46 (0.04)	578.4 (14.5)	18.0 (0.0)
	AHM	0.980 (0.001)	10.76 (0.23)	11.27 (0.25)	**0.25** (0.02)	**462.9** (4.0)	10.6 (1.3)
	MajorLinear	0.963 (0.003)	18.92 (1.31)	19.35 (1.35)	0.44 (0.04)	560.5 (13.2)	3.0 (0.0)
	MajorQuad	0.965 (0.003)	18.33 (1.32)	19.01 (1.38)	0.43 (0.04)	557.8 (13.6)	6.0 (0.0)
	MultipleScheffe	0.965 (0.003)	20.01 (1.45)	23.56 (1.94)	0.53 (0.05)	598.2 (13.8)	24.0 (0.0)
	TrueModel	0.979 (0.001)	10.82 (0.18)	11.12 (0.20)	0.25 (0.02)	457.4 (3.1)	5.0 (0.0)

2.4.1.2 Using Maximin Distance Design for Major Components

The idea of the maximin design is to spread design points in the constrained space by maximizing the minimum distance between all pairs of points. Although originally proposed for computer experiments (Johnson et al. 1990), the maximin design can be applied to mixture experiments too, except that the experimental space is confined to a polyhedron, more complicated than the typical cubic for the computer experiments. We used a stochastic search strategy to find the optimal design under the maximin distance criterion. Due to the nature of the stochastic search, the optimal design varies somewhat in each search, which is different from the simplex designs. Thus in each replicate of the simulation, the maximin design might be slightly different. The algorithm to generate the maximin distance design in the mixture experiment is available in the supplemental materials. Figure 2.1 (B) shows the maximin distance design for the three major components.

Table 2.2 compares the performances of different models in terms of the same measurements as above. The results are similar to the ones in Table 2.1. The AHM has a better prediction performance than other methods in terms of MSCV and MSCVnorm under different true simulation models but IV. The R^2, AICc and MSE values show that the AHM ranks with the best fitting models in all scenarios. For the model generating models I, IV, and V, the 2nd_MM have comparable fitting performance as the AHM. The model size of the AHM is larger than that of the 1st_MM but smaller than that of the 2nd_MM.

Table 2.2: Performance comparisons of models under the unconstrained MoM experiment using the maximin distance design for major components from 50 simulation replications (means and standard errors (in parenthesis)).

	Model	R^2	MSE	MSCV	MSCVnorm	AICc	Size
I	1st_MM	0.973 (0.004)	13.45 (2.04)	13.84 (2.09)	0.29 (0.02)	567.2 (35.5)	6.0 (0.0)
	2nd_MM	0.978 (0.003)	11.63 (1.61)	12.73 (1.74)	0.27 (0.01)	550.7 (33.1)	18.0 (0.0)
	AHM	0.978 (0.003)	11.53 (1.58)	11.98 (1.64)	**0.25** (0.01)	**537.5** (33.1)	8.7 (1.5)
	MajorLinear	0.973 (0.004)	13.43 (2.02)	13.62 (2.04)	0.29 (0.02)	563.6 (35.3)	3.0 (0.0)
	MajorQuad	0.977 (0.003)	11.66 (1.58)	11.99 (1.62)	0.26 (0.01)	**536.8** (32.4)	6.0 (0.0)
	MultipleScheffe	0.976 (0.004)	13.54 (2.10)	15.53 (2.44)	0.33 (0.03)	591.4 (36.4)	24.0 (0.0)
	TrueModel	0.977 (0.003)	11.66 (1.58)	11.87 (1.61)	0.25 (0.01)	534.5 (32.5)	4.0 (0.0)
II	1st_MM	0.989 (0.001)	15.34 (1.91)	15.78 (1.97)	**0.26** (0.02)	**596.3** (30.7)	6.0 (0.0)
	2nd_MM	0.990 (0.001)	15.36 (1.97)	16.74 (2.14)	0.27 (0.02)	611.1 (31.5)	18.0 (0.0)
	AHM	0.989 (0.001)	15.20 (1.90)	15.76 (1.97)	**0.26** (0.02)	**597.0** (31.0)	8.4 (1.3)
	MajorLinear	0.981 (0.002)	25.62 (3.40)	26.02 (3.46)	0.42 (0.03)	703.6 (31.3)	3.0 (0.0)
	MajorQuad	0.982 (0.002)	25.77 (3.39)	26.55 (3.50)	0.43 (0.03)	708.2 (30.9)	6.0 (0.0)
	MultipleScheffe	0.990 (0.001)	15.34 (1.98)	17.45 (2.25)	0.28 (0.02)	619.1 (31.8)	24.0 (0.0)
	TrueModel	0.989 (0.001)	15.34 (1.91)	15.78 (1.97)	0.26 (0.02)	596.3 (30.7)	6.0 (0.0)
III	1st_MM	0.929 (0.007)	150.60 (15.81)	155.02 (16.32)	0.31 (0.02)	1090.4 (23.8)	6.0 (0.0)
	2nd_MM	0.943 (0.004)	128.08 (9.61)	140.05 (10.99)	0.28 (0.02)	1070.6 (17.0)	18.0 (0.0)
	AHM	0.943 (0.004)	124.16 (8.00)	129.49 (8.34)	**0.26** (0.02)	**1053.3** (14.3)	9.4 (1.6)
	MajorLinear	0.826 (0.012)	366.96 (35.22)	372.13 (35.61)	0.73 (0.03)	1279.7 (21.1)	3.0 (0.0)
	MajorQuad	0.833 (0.011)	357.67 (34.63)	367.91 (35.36)	0.72 (0.04)	1277.5 (21.0)	6.0 (0.0)
	MultipleScheffe	0.937 (0.007)	147.92 (15.40)	171.05 (18.57)	0.34 (0.03)	1109.5 (23.5)	24.0 (0.0)
	TrueModel	0.941 (0.004)	125.13 (7.85)	128.10 (8.06)	0.25 (0.02)	1050.0 (14.3)	5.0 (0.0)
IV	1st_MM	0.847 (0.011)	278.01 (27.47)	288.23 (28.36)	0.63 (0.05)	1222.9 (22.4)	6.0 (0.0)
	2nd_MM	0.872 (0.012)	246.28 (30.52)	268.71 (32.91)	0.59 (0.05)	1210.6 (29.5)	18.0 (0.0)
	AHM	0.868 (0.015)	241.95 (30.86)	253.44 (31.87)	0.56 (0.06)	1195.1 (29.4)	8.6 (1.3)
	MajorLinear	0.839 (0.009)	287.63 (24.82)	292.46 (25.18)	0.64 (0.05)	1227.3 (19.3)	3.0 (0.0)
	MajorQuad	0.840 (0.009)	289.93 (25.14)	299.03 (25.93)	0.66 (0.05)	1232.3 (19.3)	6.0 (0.0)
	MultipleScheffe	0.942 (0.006)	114.60 (11.03)	131.75 (13.08)	**0.29** (0.02)	**1054.5** (21.8)	24.0 (0.0)
	TrueModel	0.936 (0.005)	115.05 (9.66)	117.86 (9.90)	0.26 (0.01)	1031.5 (19.1)	5.0 (0.0)
V	1st_MM	0.968 (0.006)	16.78 (3.12)	17.25 (3.20)	0.43 (0.04)	613.2 (45.9)	6.0 (0.0)
	2nd_MM	0.970 (0.006)	16.65 (3.16)	17.82 (3.31)	0.44 (0.05)	626.1 (46.9)	18.0 (0.0)
	AHM	0.981 (0.002)	9.83 (1.21)	10.24 (1.25)	**0.26** (0.02)	**506.0** (28.8)	11.0 (1.4)
	MajorLinear	0.967 (0.006)	16.66 (3.06)	16.97 (3.12)	0.42 (0.04)	608.5 (45.5)	3.0 (0.0)
	MajorQuad	0.969 (0.006)	16.23 (3.03)	16.76 (3.13)	0.42 (0.04)	606.0 (46.4)	6.0 (0.0)
	MultipleScheffe	0.969 (0.006)	17.41 (3.29)	19.98 (3.82)	0.50 (0.06)	644.0 (46.6)	24.0 (0.0)
	TrueModel	0.981 (0.002)	9.88 (1.17)	10.12 (1.19)	0.25 (0.02)	500.5 (27.8)	5.0 (0.0)

2.4.2 Constrained MoM Experiments

There have been many mixture experiments with additional constraints imposed on the components. Here we also consider simulations where certain lower and upper bounds are placed on both major and minor components. Specifically, we assume that the major and minor components have to satisfy the following constraints.

$$\begin{aligned} c_1 + c_2 + c_3 = 1, \quad & 0.2 \le c_1 \le 0.45, \\ 0.4 \le c_2 \le 0.6, \quad & 0.1 \le c_3 \le 0.25, \\ x_{11} + x_{12} = 1, \quad & 0.5 \le x_{11} \le 0.85, \\ x_{21} + x_{22} = 1, \quad & 0.73 \le x_{21} \le 0.95, \\ x_{31} + x_{32} = 1, \quad & 0.68 \le x_{31} \le 0.92. \end{aligned} \tag{2.11}$$

Figure 2.1 (C) shows the I-optimal design for the second-order Scheffé model for the major components in the constrained mixture experiment. The design is generated by the AlgDesign package in R software. Figure 2.1 (D) shows the maximin distance design for the major components in the constrained mixture experiment. The comparison results for using the I-optimal design and the maximin distance design for the major components are reported in Table 2.3 and 2.4, respectively. From both tables, we can conclude that the AHM, the 1st_MM, and the 2nd_MM all have competitive prediction performance.

It is worth noting that in the simulation with data generating model IV, the AHM has a comparable prediction performance as the true model, which is an improvement from the unconstrained MoM case. This phenomenon is likely due to the additional constraints which make the design space more complicated, and thus the flexibility of the AHM is more advantageous than its counterparts for the true underlying model IV. The AHM has similar

R^2, AICc and MSE values as the 1st_MM for simulation models I-V. The model size of the AHM is larger than that of the 1st_MM but smaller than that of the 2nd_MM.

Table 2.3: Performance comparisons of models under the constrained MoM experiment using I-optimal design for major components from 50 simulation replications (means and standard errors (in parenthesis)).

	Model	R^2	MSE	MSCV	MSCVnorm	AICc	Size
I	1st_MM	0.999 (0.000)	0.47 (0.01)	0.49 (0.01)	**0.26** (0.02)	-153.1 (5.3)	6.0 (0.0)
	2nd_MM	0.999 (0.000)	0.47 (0.01)	0.51 (0.01)	0.28 (0.02)	-142.2 (6.3)	18.0 (0.0)
	AHM	0.999 (0.000)	0.46 (0.01)	0.48 (0.01)	**0.26** (0.02)	-154.1 (6.7)	9.0 (2.1)
	MajorLinear	0.999 (0.000)	0.47 (0.01)	0.48 (0.01)	**0.26** (0.02)	-156.7 (4.3)	3.0 (0.0)
	MajorQuad	0.999 (0.000)	0.46 (0.01)	0.48 (0.01)	**0.26** (0.02)	**-157.3** (4.0)	6.0 (0.0)
	MultipleScheffe	0.999 (0.000)	0.47 (0.02)	0.54 (0.02)	0.29 (0.02)	-129.8 (7.9)	24.0 (0.0)
	TrueModel	0.999 (0.000)	0.46 (0.01)	0.47 (0.01)	0.26 (0.02)	-159.2 (3.1)	4.0 (0.0)
II	1st_MM	0.999 (0.000)	1.14 (0.01)	1.17 (0.02)	**0.26** (0.02)	**36.9** (2.8)	6.0 (0.0)
	2nd_MM	0.999 (0.000)	1.14 (0.03)	1.25 (0.03)	0.27 (0.02)	50.9 (5.3)	18.0 (0.0)
	AHM	0.999 (0.000)	1.13 (0.02)	1.17 (0.02)	**0.26** (0.02)	39.0 (4.2)	9.0 (1.4)
	MajorLinear	0.999 (0.000)	1.67 (0.12)	1.70 (0.12)	0.37 (0.02)	115.8 (15.3)	3.0 (0.0)
	MajorQuad	0.999 (0.000)	1.68 (0.12)	1.73 (0.12)	0.38 (0.02)	119.8 (15.7)	6.0 (0.0)
	MultipleScheffe	0.999 (0.000)	1.14 (0.04)	1.30 (0.05)	0.29 (0.02)	58.9 (7.8)	24.0 (0.0)
	TrueModel	0.999 (0.000)	1.14 (0.01)	1.17 (0.02)	0.26 (0.02)	36.9 (2.8)	6.0 (0.0)
III	1st_MM	0.998 (0.000)	8.53 (0.19)	8.78 (0.19)	**0.26** (0.01)	**471.5** (4.8)	6.0 (0.0)
	2nd_MM	0.998 (0.000)	8.33 (0.28)	9.09 (0.31)	0.27 (0.02)	480.9 (7.5)	18.0 (0.0)
	AHM	0.998 (0.000)	8.37 (0.19)	8.53 (0.17)	**0.26** (0.01)	**471.0** (5.0)	9.3 (1.2)
	MajorLinear	0.994 (0.000)	27.01 (1.31)	27.39 (1.33)	0.82 (0.02)	716.9 (10.5)	3.0 (0.0)
	MajorQuad	0.994 (0.000)	27.24 (1.30)	28.02 (1.34)	0.84 (0.03)	722.0 (10.3)	6.0 (0.0)
	MultipleScheffe	0.998 (0.000)	8.35 (0.32)	9.46 (0.39)	0.28 (0.02)	489.7 (8.4)	24.0 (0.0)
	TrueModel	0.998 (0.000)	8.35 (0.09)	8.55 (0.09)	0.26 (0.01)	465.8 (2.4)	5.0 (0.0)
IV	1st_MM	0.994 (0.000)	12.10 (0.57)	12.47 (0.59)	0.30 (0.02)	546.8 (10.2)	6.0 (0.0)
	2nd_MM	0.995 (0.000)	10.48 (0.33)	11.46 (0.37)	0.28 (0.02)	530.6 (6.7)	18.0 (0.0)
	AHM	0.995 (0.000)	10.73 (0.33)	11.05 (0.33)	**0.27** (0.02)	**523.8** (7.1)	8.5 (1.1)
	MajorLinear	0.989 (0.001)	24.18 (1.46)	24.52 (1.48)	0.60 (0.03)	692.8 (13.0)	3.0 (0.0)
	MajorQuad	0.989 (0.001)	24.39 (1.49)	25.09 (1.53)	0.61 (0.03)	698.0 (13.1)	6.0 (0.0)
	MultipleScheffe	0.996 (0.000)	10.38 (0.29)	11.78 (0.42)	0.29 (0.02)	536.7 (6.0)	24.0 (0.0)
	TrueModel	0.995 (0.000)	10.34 (0.13)	10.58 (0.14)	0.26 (0.02)	511.8 (2.8)	5.0 (0.0)
V	1st_MM	0.999 (0.000)	0.39 (0.01)	0.41 (0.01)	0.27 (0.02)	-192.5 (7.5)	6.0 (0.0)
	2nd_MM	0.999 (0.000)	0.38 (0.02)	0.42 (0.02)	0.28 (0.02)	-185.6 (9.4)	18.0 (0.0)
	AHM	0.999 (0.000)	0.37 (0.01)	0.39 (0.01)	**0.26** (0.02)	**-200.1** (4.5)	9.7 (1.4)
	MajorLinear	0.999 (0.000)	0.50 (0.03)	0.51 (0.03)	0.34 (0.02)	-145.9 (14.0)	3.0 (0.0)
	MajorQuad	0.999 (0.000)	0.49 (0.03)	0.51 (0.03)	0.34 (0.02)	-145.4 (14.3)	6.0 (0.0)
	MultipleScheffe	0.999 (0.000)	0.39 (0.02)	0.45 (0.03)	0.30 (0.03)	-171.1 (11.1)	24.0 (0.0)
	TrueModel	0.999 (0.000)	0.37 (0.01)	0.38 (0.01)	0.26 (0.02)	-206.1 (3.1)	5.0 (0.0)

Table 2.4: Performance comparisons of models under the constrained MoM experiment using the maximin distance design for major components from 50 simulation replications (means and standard errors (in parenthesis)).

	Model	R^2	MSE	MSCV	MSCVnorm	AICc	Size
I	1st_MM	0.999 (0.000)	0.45 (0.04)	0.46 (0.05)	**0.26** (0.02)	-165.2 (22.6)	6.0 (0.0)
	2nd_MM	0.999 (0.000)	0.45 (0.05)	0.49 (0.05)	0.27 (0.02)	-152.9 (23.9)	18.0 (0.0)
	AHM	0.999 (0.000)	0.44 (0.05)	0.45 (0.05)	**0.26** (0.02)	-165.7 (24.6)	8.8 (1.7)
	MajorLinear	0.999 (0.000)	0.45 (0.04)	0.46 (0.05)	**0.26** (0.01)	**-168.8** (23.0)	3.0 (0.0)
	MajorQuad	0.999 (0.000)	0.44 (0.05)	0.46 (0.05)	**0.26** (0.02)	**-168.4** (24.4)	6.0 (0.0)
	MultipleScheffe	0.999 (0.000)	0.45 (0.05)	0.52 (0.06)	0.29 (0.02)	-138.5 (24.2)	26.5 (1.1)
	TrueModel	0.999 (0.000)	0.44 (0.05)	0.45 (0.05)	0.25 (0.02)	-170.3 (24.0)	4.0 (0.0)
II	1st_MM	0.999 (0.000)	1.08 (0.09)	1.11 (0.09)	**0.26** (0.01)	**23.8** (18.4)	6.0 (0.0)
	2nd_MM	0.999 (0.000)	1.07 (0.09)	1.17 (0.10)	0.27 (0.02)	37.3 (18.6)	18.0 (0.0)
	AHM	0.999 (0.000)	1.08 (0.09)	1.11 (0.09)	**0.26** (0.01)	26.9 (18.4)	9.0 (1.4)
	MajorLinear	0.999 (0.000)	1.61 (0.15)	1.63 (0.15)	0.38 (0.03)	106.6 (20.7)	3.0 (0.0)
	MajorQuad	0.999 (0.000)	1.61 (0.15)	1.66 (0.16)	0.39 (0.03)	111.0 (20.9)	6.0 (0.0)
	MultipleScheffe	0.999 (0.000)	1.07 (0.09)	1.23 (0.11)	0.29 (0.02)	49.1 (20.1)	26.5 (1.1)
	TrueModel	0.999 (0.000)	1.08 (0.09)	1.11 (0.09)	0.26 (0.01)	23.8 (18.4)	6.0 (0.0)
III	1st_MM	0.998 (0.000)	8.47 (0.29)	8.72 (0.30)	**0.26** (0.02)	**469.9** (7.3)	6.0 (0.0)
	2nd_MM	0.998 (0.000)	8.39 (0.27)	9.17 (0.31)	0.28 (0.02)	482.3 (7.0)	18.0 (0.0)
	AHM	0.998 (0.000)	8.38 (0.27)	8.55 (0.28)	**0.26** (0.02)	471.4 (7.1)	9.3 (1.0)
	MajorLinear	0.994 (0.000)	26.69 (1.63)	27.08 (1.65)	0.82 (0.03)	714.2 (13.1)	3.0 (0.0)
	MajorQuad	0.994 (0.000)	26.95 (1.62)	27.72 (1.67)	0.84 (0.04)	719.5 (13.0)	6.0 (0.0)
	MultipleScheffe	0.998 (0.000)	8.45 (0.34)	9.77 (0.44)	0.30 (0.02)	495.9 (9.2)	26.5 (1.1)
	TrueModel	0.998 (0.000)	8.34 (0.22)	8.54 (0.23)	0.26 (0.01)	465.4 (5.9)	5.0 (0.0)
IV	1st_MM	0.995 (0.000)	11.06 (0.82)	11.39 (0.85)	0.30 (0.02)	527.0 (16.6)	6.0 (0.0)
	2nd_MM	0.996 (0.000)	9.60 (0.54)	10.50 (0.60)	0.27 (0.02)	511.3 (12.4)	18.0 (0.0)
	AHM	0.995 (0.000)	9.82 (0.59)	10.11 (0.60)	**0.26** (0.02)	**504.7** (13.4)	8.8 (1.2)
	MajorLinear	0.989 (0.001)	23.74 (1.60)	24.07 (1.63)	0.63 (0.04)	688.8 (14.7)	3.0 (0.0)
	MajorQuad	0.989 (0.001)	23.91 (1.64)	24.60 (1.69)	0.64 (0.04)	693.6 (14.9)	6.0 (0.0)
	MultipleScheffe	0.996 (0.000)	9.55 (0.54)	11.01 (0.67)	0.29 (0.02)	522.2 (12.7)	26.5 (1.1)
	TrueModel	0.996 (0.000)	9.52 (0.48)	9.75 (0.49)	0.25 (0.02)	493.8 (11.1)	5.0 (0.0)
V	1st_MM	0.999 (0.000)	0.35 (0.03)	0.36 (0.03)	0.27 (0.02)	-219.5 (19.9)	6.0 (0.0)
	2nd_MM	0.999 (0.000)	0.34 (0.03)	0.37 (0.03)	0.28 (0.02)	-209.8 (19.5)	18.0 (0.0)
	AHM	0.999 (0.000)	0.33 (0.03)	0.34 (0.03)	**0.26** (0.01)	**-225.2** (19.9)	10.0 (1.5)
	MajorLinear	0.999 (0.000)	0.46 (0.05)	0.46 (0.05)	0.35 (0.02)	-165.2 (22.1)	3.0 (0.0)
	MajorQuad	0.999 (0.000)	0.45 (0.05)	0.47 (0.05)	0.35 (0.02)	-163.2 (22.1)	6.0 (0.0)
	MultipleScheffe	0.999 (0.000)	0.34 (0.03)	0.40 (0.04)	0.30 (0.02)	-195.8 (20.6)	26.5 (1.1)
	TrueModel	0.999 (0.000)	0.33 (0.03)	0.34 (0.03)	0.26 (0.01)	-231.3 (19.3)	5.0 (0.0)

2.5 Real-Data Analysis

In this section, we analyze two real-data problems studied previously in the literature, the Photoresist-Coating experiment (Cornell and Ramsey 1998) and the Pringles experiment (Kang et al. 2011), to evaluate the model performance of the proposed AHM.

2.5.1 Photoresist-Coating Experiment

The objective of the photoresist-coating experiment is to determine the effect of proportions of resins in the formulation on the photoresist material's characteristic of interest (Cornell and Ramsey 1998). The two major components (c_1 and c_2) are the base resin types , and the minor component are the minor resins possessing different dissolution rates (slow and fast) denoted as x_{11}, x_{12}, and x_{21}, x_{22}, respectively. The range of values for all components is $[0, 1]$. All possible settings of (c_1, c_2) are $(0.75, 0.25)$, $(0.5, 0.5)$, and $(0.25, 0.75)$. For (x_{k1}, x_{k2}) with $k = 1, 2$, three settings $(1, 0)$, $(0.5, 0.5)$ and $(0, 1)$ are chosen. The overall design is a crossed design of the three designs containing $3 \times 3 \times 3 = 27$ design points. In total, 42 measurements are observed with replications at certain design points.

Table 2.5 compares the performances of different models. The AHM has the smallest MSCV and AICc values among all. The MajorLinear and MajorQuad model have the least prediction performance, indicating that the minor components play an important role. The 2nd_MM has better prediction and model fitting than the 1st_MM and the MultipleScheffe. The model size of AHM is larger than that of the 1st_MM and 2nd_MM, but smaller than that of the MultipleScheffe.

Table 2.5: Performance comparisons of models in the Photoresist-Coating experiment

Model	R^2	MSE	MSCV	AICc	Size
AHM	0.998	1.929	**2.324**	**44.552**	9
MultipleScheffe	1.000	0.159	35.193	60.378	27
MajorLinear	0.902	90.321	95.529	193.724	2
MajorQuad	0.903	91.312	98.336	195.569	3
1st_MM	0.995	4.591	5.294	71.477	4
2nd_MM	0.998	2.425	2.860	51.952	8

The fitted AHM is

$$y = 25.919c_1 + 29.21c_2 - 6.536c_1^{1.1}x_{11} + 23.616c_1^{1.1}x_{12} - 5.58c_2^{1.1}x_{21} + 30.706c_2^{1.1}x_{22}$$
$$- 38.974c_1c_2 - 18.818(c_1^{1.1})^2x_{11}x_{12} - 19.363(c_2^{1.1})^2x_{21}x_{22}.$$

Based on the estimated parameters in the fitted AHM, the major components have significant effects on the response, and both main and interaction effects of the minor components depend on their respective major components. These results are consistent with the findings in the paper (Cornell and Ramsey 1998). However, the AHM reveals that the inter-major-component blending property exists via the interaction term between the major components, c_1c_2, which is different from the multiple Scheffé model results (Cornell and Ramsey, 1998). The multiple Scheffé model assumes that the blending properties of the minor components of one major component also depend on the presence of minor components of other major components. The reason for this different interpretation is that AHM considers all the inter-major-component interactions only at the major-component level, and all the inter-minor-component interactions are restricted within the minor components nested under the same major component. As a result, no inter-minor-component-interactions are considered for the minor components nested under different major components.

2.5.2 Pringles Experiment

The goal of Pringles® experiment is to develop a new kind of Pringles® potato crisp such that the percentage of fat and the hardness in the potato crisps are optimized. The constraints on the components are given by

$$\begin{aligned} c_1 + c_2 + c_3 = 1, &\quad 0.601 \le c_1 \le 0.643, \\ 0.34 \le c_2 \le 0.38, &\quad 0.017 \le c_3 \le 0.019, \\ x_{11} + x_{12} = 1, &\quad x_{21} + x_{22} = 1, \\ 0.835 \le x_{11} \le 0.905, &\quad 0.095 \le x_{12} \le 0.165, \\ 0.9 \le x_{21} \le 0.98, &\quad 0.02 \le x_{22} \le 0.1. \end{aligned}$$

The experimental design is illustrated in Kang et al (2011) in details.

Table 2.6 compares the model performances for the Pringles experiment. For both responses "Hardness" and "%Fat", compared to the 1st_MM, the AHM has smaller MSCV value, but larger AICc value, suggesting better prediction performance but worse fitting performance. The MajorLinear and MajorQuad model have the largest MSCV values, indicating that the minor components play an important role in this study. The MultipleScheffe model and the 2nd_MM has largest AICc values. The model size of AHM is larger than that of the 1st_MM, but smaller than that of the 2nd_MM.

We use the fitted AHM to find the optimal settings to maximize the response Hardness. The fitted AHM is

$$\begin{aligned} \hat{y}_{hardness} = &\, 9.745c_1 - 5.115c_2 + 6.916c_1^{1.3}x_{11} - 11.184c_2^{1.3}x_{21} + 27.203c_2^{1.3}x_{22} + \\ &+ 21.176(c_2^{1.3})^2 x_{21}x_{22}. \end{aligned}$$

Table 2.6: Comparison between proposed models

Response	Model	R^2	MSE	MSCV	AICc	Size
%Fat	AHM	1.000	0.261	**0.362**	10.048	8
	MultipleScheffe	1.000	0.230	0.794	101.560	12
	MajorLinear	0.999	1.426	1.757	14.068	3
	MajorQuad	0.999	1.569	3.508	22.131	5
	1st_MM	1.000	0.296	0.421	**-6.202**	5
	2nd_MM	1.000	0.102	0.590	87.772	12
Hardness	AHM	0.996	0.157	**0.174**	-12.439	6
	MultipleScheffe	0.999	0.069	0.235	81.138	12
	MajorLinear	0.985	0.487	0.600	-4.194	3
	MajorQuad	0.986	0.546	0.856	4.195	5
	1st_MM	0.997	0.128	0.183	**-20.474**	5
	2nd_MM	0.999	0.130	0.428	91.823	12

The 1st_MM proposed by Kang et al. (2011) is

$$\hat{y}_{hardness} = 8.786c_1 + 20.966c_2 + 13.506c_3 + 8.658c_1x_{11} - 37.641c_2x_{21}.$$

Compared to the 1st_MM, the fitted AHM does not contain the third major component c_3. Similarly, we can also use the fitted AHM to find the minimizer of %Fat. Table 2.7 shows the optimal settings to minimize the response %Fat and to maximize the response Hardness, respectively. The optimization can be performed using the constrained nonlinear optimization in R software. These optimal settings agree well with Kang et al. (2011). This experiment is a preliminary study, and in the follow-up experiments, larger experiments should be conducted around the optimal settings to find better formulations.

Table 2.7: Optimal settings from the AHMs

Response	c_1	c_2	c_3	x_{11}	x_{12}	x_{21}	x_{22}
%Fat	0.641	0.34	0.019	0.892	0.108	0.9	0.1
Hardness	0.643	0.34	0.017	0.905	0.095	0.9	0.1

2.6 Discussion

The intrinsic relationship between the major and minor components is a key feature in the mixture-of-mixtures (MoM) experiment. This work proposes an additive heredity model with a meaningful interpretation of the model structure for MoM experiments. The additive heredity model considers the effects of major and minor components in an additive fashion and employs the hierarchical and heredity principles by the nonnegative garrote technique for model selection. The additive heredity model incorporates the dependence between the major and minor components via the coefficients of minor components. The coefficient functions represent various types of knowledge. For example, when one major component is not included, all of its corresponding minor components are excluded from the model. According to the numerical studies, the additive heredity model provides superior prediction performances compared to the benchmark models.

It is worth remarking that the MoM experiment is closely related to the multilevel model. The multilevel model usually has two types of variables, the group-level variables and the individual-level variables (Dedrick et al. 2009). The group-level variable has a direct effect on the response, while the individual-level variable contributes to the response in both direct and indirect ways. For example of using the random intercept in the multilevel model, the intercept term will be dependent upon the individual-level variable, representing the direct contribution of the individual-level variable. The interaction between the individual-level

variable and the group-level variable represents the indirect contribution of the individual-level variable. However, in the proposed AHM, the contribution of the minor components (the individual-level variable) is made through a function depending on the major components (the group-level variable).

The proposed AHM can be extended to the general varying-coefficient models (Hastie and Tibshirani 1993; Fan and Zhang 1999). For example, one can consider the coefficients for the minor components to be nonparametric functions of the corresponding major components. In this work, we adopted a parametric power function of order h to express the structural dependence between the major and minor components. This power function is monotonic and bounded on the domain of c_k. It will be interesting to investigate how to incorporate an appropriate nonparametric form, allowing flexible structures to describe the structural dependence of minor components on their corresponding major components.

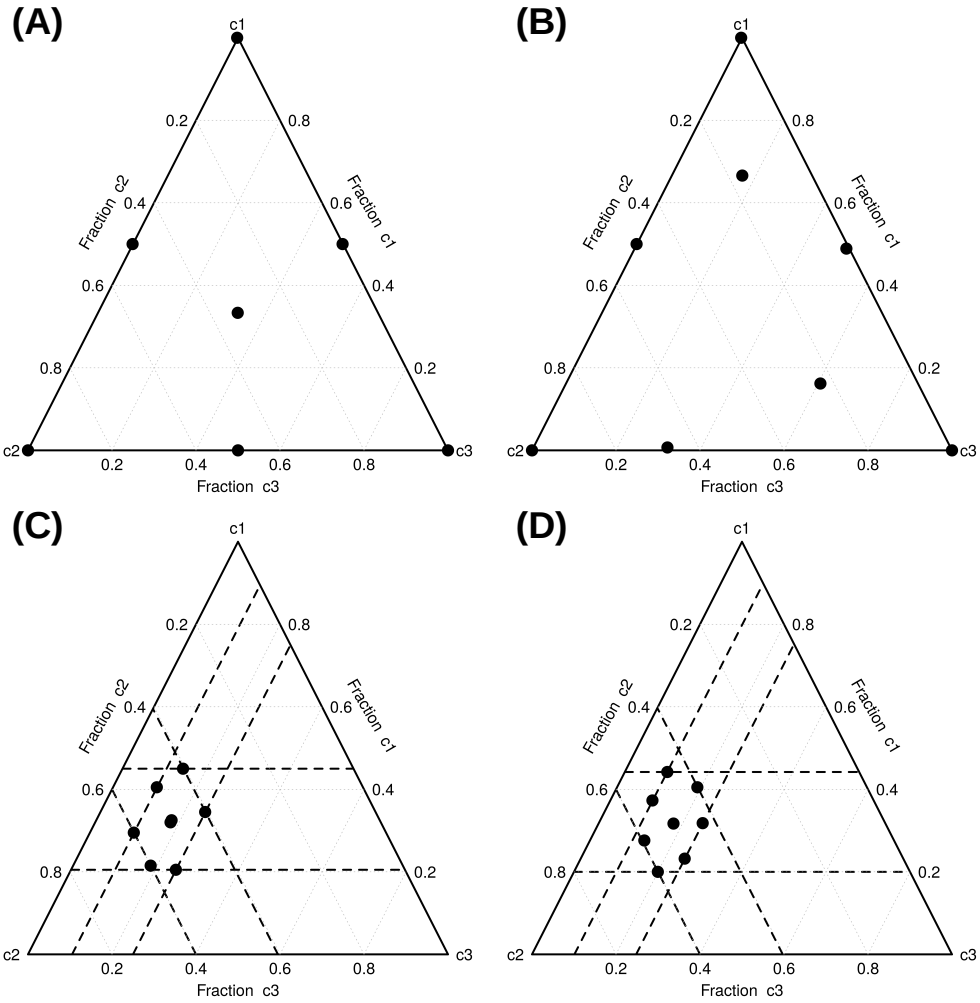

Figure 2.1: Designs for the major components: in the unconstrained mixture experiment (A) I-optimal design with 7 design points, (B) Maximin distance design with 8 design points; and in the constrained mixture experiment (C) I-optimal design with 8 design points, (D) Maximin distance design with 8 design points. In (C) and (D) the dashed lines represent the upper and lower constraints for each mixture component.

Chapter 3 Dynamic Variable Selection for Generalized Linear Models

3.1 Introduction

In various real applications of statistical modeling (Cai et al. 2000; Aguilar and West 2000; Beaulieu et al., 2012; Wang and Hastie 2012; Hong et al. 2015), the model coefficients of predictor variables are allowed to vary with certain covariates. These dynamic coefficient effects have good interpretation, but their functional forms are often implicit and complicated. For example, in the crystal ingot growth experiment (Jin et al. 2019), the effects of predictor variables (e.g., the pulling speed and the power of the heater) on the quality response are dynamic. This is because not only the length of the crystal ingot is growing, but the equipment conditions are degrading during the long-running experiment. It is difficult to specify the functional forms for the dynamic coefficients, especially considering the manufacturing operations and possible sudden changes in the effects of predictors. Therefore, there is a need to investigate on how to model and analyze the dynamic coefficients.

In the literature, the dynamic linear model (DLM, West and Harrison 1997) is an important tool to explore the dynamic effects of variables in the dynamic linear regression. The DLM assumes that the regression coefficients vary over time (Petris et al. 2009) and the model estimation is conducted by using the Kalman filter technique (Kalman 1960; Kalman and Bucy 1963). One limitation of the DLM is that it assumes a linear relationship between two

consecutive states in the latent state equation, where the latent state process follows a Gaussian distribution. Note that the Gaussian distribution cannot properly describe the abrupt changes in the state process. Another commonly used approach is the varying coefficient model (VCM, Cleveland et al. 1991, Hastie and Tibshirani 1993, Fan and Zhang 2008). The VCM explores the dynamic effects of variables by considering the model coefficients to be functions of certain variables. The estimation of varying coefficients can be obtained based on smoothing techniques, such as the kernel-local polynomial smoothing method (Fan and Zhang 1999), the polynomial spline method (Huang and Shen 2001), and the smoothing spline method (Hastie and Tibshirani 1993). However, the applications of the VCM can be limited due to its assumption that the varying coefficients are in the forms of smoothing functions. Thus, it may not be suitable for the VCM to handle problems with discontinuities or sudden structure changes in dynamic coefficients.

Several recent works demonstrate the use of the penalized likelihood approach to illustrate the dynamic effects of model coefficients. Ahmed and Xing (2009) proposed to use the fused and the l_1-norm penalty in estimating the temporal structures in time-varying Markov random field networks. Adhikari et al. (2019) developed the multinomial fused lasso regression with the fused and the l_1-norm penalty to study a longitudinal data problem. Kolar et al. (2009) introduced the so-called varying-coefficient varying-structure (VCVS) model with a quadratic loss function associated with the fused and the l_1-norm penalty to uncover the dynamic coefficient structures. These methods are based on the approach of using the total variation (TV) technique to estimate multiple change-points in coefficients (Yao 1988; Lavielle 2005; Lebrbier 2005; Harchaoui and Lévy-Leduc 2010; Bleakley and Vert 2011; Zhang et al. 2015). Their key idea is to relax the penalty term from the number of change-points to the magnitude of jumps. It is worth pointing out that the TV formulation is related to the fused lasso method (Tibshirani et al. 2005), where the consecutive variables tend to

have similar coefficient values. The focus of those works (Kolar et al. 2009; Ahmed and Xing 2009; Adhikari et al. 2019) is to achieve sparsity in the structures of coefficient parameters. That is, they select the relevant features in the estimated segments between the estimated change-points in the coefficient parameters. However, these work did not address the issue of the selection of predictor variables as a whole. Thus, they cannot appropriately identify the significant variables in the presence of noise variables.

In this work, we propose a dynamic regression model with dynamic variable selection such that the important predictor variables can be identified and their dynamic effects can be better quantified. The proposed method is designed for non-normal responses, for example, binary response, in the framework of generalized linear models. Furthermore, our proposed model considers each observation has its own coefficient parameters. The detailed development is shown in Section 2. Under the penalized regression framework, the l_2-norm group lasso penalty is adopted to select important variables for a sparse and interpretable model. The fused lasso penalty is used to encourage piecewise constant functions as a good approximation of unknown functional coefficients. We employ the alternating direction method of multipliers (ADMM, Boyd et al. 2011; Zhu 2017) coupled with the Newton-Raphson method for the parameter estimation. The proposed model has several advantages. First, the proposed dynamic model approximates the functional coefficients by piecewise constant functions, thus, it is flexible to estimate underlying complex functional coefficients. Second, the proposed method is able to handle multiple changes in the varying coefficients where both numbers and positions are detected automatically. Third, the proposed method is able to select significant variables, therefore, it is suitable for problems in high-dimensional settings.

The remainder of this Chapter is structured as follows. In Section 2, we detail the proposed regularized dynamic logistic regression model. In Section 3, we present the estimation process via the Alternating Direction Method of Multipliers (ADMM). In Sections 4 and 5, we present

the simulation studies and real-data analysis. We conclude this work with some discussion in Section 6.

3.2 Regularized Dynamic Logistic Regression

Let us consider the binary response as $y_t \in \{0, 1\}, t = 1, \ldots, n$, and the predictor variables as $\boldsymbol{x}_t = (x_{t,1}, \ldots, x_{t,p})^T$, $p \geq 1$. Without loss of generality, we assume that the p variables are continuous variables with measurement at n time points. We denote $p(\boldsymbol{x}_t) = Pr(y_t = 1|\boldsymbol{x}_t)$, that is, we have

$$y_t|\boldsymbol{x}_t = \begin{cases} 1, & \text{w.p. } p(\boldsymbol{x}_t), \\ 0, & \text{w.p. } 1 - p(\boldsymbol{x}_t). \end{cases}$$

We model the conditional probability $p(\boldsymbol{x}_t)$ with the logistic regression model $\log(p(\boldsymbol{x}_t)/(1 - p(\boldsymbol{x}_t))) = \boldsymbol{x}_t^T\boldsymbol{\beta}_t$ with $\boldsymbol{\beta}_t = (\beta_{t,1}, \ldots, \beta_{t,p})^T$. The observations have their own coefficient parameters $\boldsymbol{\beta}_t$ at each time point. Thus, the coefficient parameter $\boldsymbol{\beta}_t$ is allowed to vary over time. This is clearly an over-parameterized model since the number of unknown parameters np is larger than the sample size n. To address this issue, we consider a regularized dynamic logistic regression model, which is expressed as

$$\begin{aligned} &\text{logit}(\boldsymbol{x}_t) = \log\frac{p(\boldsymbol{x}_t)}{1 - p(\boldsymbol{x}_t)} = \boldsymbol{x}_t^T\boldsymbol{\beta}_t, \ t = 1, \cdots, n \\ &s.t. \ \sum_{j=1}^{p}\sum_{t=2}^{n} |\beta_{t,j} - \beta_{t-1,j}| \leq M_1, \\ &\sum_{j=1}^{p} \sqrt{\beta_{1,j}^2 + \cdots + \beta_{n,j}^2} \leq M_2, \end{aligned} \tag{3.1}$$

where $M_1 \geq 0$ and $M_2 \geq 0$ are the tuning parameters for the l_1-norm fuse penalty and l_2-norm group lasso penalty, respectively. Both penalties are beneficial to reduce the number

of unknown parameters in the proposed model. Furthermore, the l_1-norm fused penalty on the consecutive coefficient parameters, $\sum_{j=1}^{p}\sum_{t=2}^{n}|\beta_{t,j}-\beta_{t-1,j}|$, encourages that the adjacent coefficient parameters to have similar values than the distant coefficient parameters. That is, the l_1-norm fused penalty favors the functional coefficients being piecewise constant functions to approximate the underlying coefficients. This idea of parameter fusion is similar to those in Kolar et al. (2009) and Ahmed and Xing (2009). Moreover, the detection of locations and numbers of change-points in piecewise constant functions is data-driven.

The group lasso penalty, $\sum_{j=1}^{p}\sqrt{\beta_{1,j}^2+\cdots+\beta_{n,j}^2}$, considers the coefficient parameters at each time points for a certain variable as a group. Selecting groups is essentially equivalent to select important predictor variables. With the variable selection, we obtain models that are more interpretable and gain insight into the relationship between the response and the predictors. It is worth differentiating the variable selection feature in the proposed method from the previous works (Kolar et al. 2009; Adhikari et al. 2019). The use of the l_1-norm in their methods leads to a sparse structure in the coefficient parameters, but does not incorporate the information that the coefficient parameter from one variable at each time points are from the same variable.

The combination of the l_1-norm fused penalty and the l_2-norm group lasso penalty yields the estimated model to be sparse at the variable level but fused within the variables. This idea is similar to the sparse group lasso (Friedman et al. 2010; Simon et al. 2013), which yields the groupwise sparsity and the within group sparsity. If using only the l_1-norm fused penalty, the proposed method reduces to the benchmark method (BM1, described later). However, if using only the l_2-norm group lasso penalty, the resultant model is an over-parameterized model.

3.3 Efficient Model Estimation

To estimate the parameter matrix $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)^T$ of size $n \times p$, we minimize the logistic regression loss function combined with the l_1-norm fused penalty and the l_2-norm group lasso penalty. That is,

$$\underset{\boldsymbol{B}}{\text{minimize}} - l(\boldsymbol{B}) + \gamma_1 \sum_{j=1}^{p} \sum_{t=2}^{n} |\beta_{t,j} - \beta_{t-1,j}| + \gamma_2 \sum_{j=1}^{p} \sqrt{\beta_{1,j}^2 + \cdots + \beta_{n,j}^2} \tag{3.2}$$

with

$$\begin{aligned} l(\boldsymbol{B}) &= \log \left\{ \prod_{t=1}^{n} \left[p(\boldsymbol{x}_t)^{y_t} (1 - p(\boldsymbol{x}_t))^{1-y_t} \right] \right\} \\ &= \sum_{t=1}^{n} \left\{ y_t \boldsymbol{x}_t^T \boldsymbol{\beta}_t - \log(1 + \exp(\boldsymbol{x}_t^T \boldsymbol{\beta}_t)) \right\}, \end{aligned} \tag{3.3}$$

where $\gamma_1 \geq 0$ and $\gamma_2 \geq 0$ are tuning parameters. Note that we implicitly assume the observations are independent in the log-likelihood function, which is commonly used in the works including Kolar et al. (2009), Gibberd and Nelson (2017), and Adhikari et al. (2019).

The objective function in (3.2) is convex but the two penalties are not separable in $\boldsymbol{B}$. That is, the associated l_1-norm fused penalty and the l_2-norm group lasso penalty both contain the parameter matrix $\boldsymbol{B}$. It is well-studied to optimize a convex objective function associated with either l_1-norm fused penalty or the l_2-norm group lasso penalty, but challenging to optimize the objective function directly with the presence of both the l_1-norm fused penalty and the l_2-norm group lasso penalty. To tackle this challenge, we consider an easy-to-implement algorithm based on the alternating direction method of multipliers (ADMM, Boyd et al. 2011). The ADMM method is simple to implement and has been successfully applied to the generalized lasso problem (Wahlberg et al. 2012; Zhu 2017). The main idea

is to convert the objective function such that we are able to deal with the two penalties individually. To update variables in an alternative way, we rewrite the problem (2) in an equivalent form as

$$\begin{aligned} \underset{\boldsymbol{B}}{\text{minimize}}\ l_{\gamma_1,\gamma_2}(\boldsymbol{B}) \triangleq & -l(\boldsymbol{B}) + \gamma_1 \sum_{j=1}^{p} ||\boldsymbol{Z}_j^{(1)}||_1 + \gamma_2 \sum_{j=1}^{p} ||\boldsymbol{Z}_j^{(2)}||_2, \\ \text{subject to } & F\boldsymbol{B}_j - \boldsymbol{Z}_j^{(1)} = 0, \\ & \boldsymbol{B}_j - \boldsymbol{Z}_j^{(2)} = 0,\ j = 1, \cdots, p, \end{aligned} \tag{3.4}$$

where $||\cdot||_1$ is the l_1-norm, $||\cdot||_2$ is the l_2-norm, $\boldsymbol{Z}_j^{(1)}$ and $\boldsymbol{Z}_j^{(2)}$ are the jth columns in the matrix $\boldsymbol{Z}^{(1)}$ of size $(n-1)\times p$ and the matrix $\boldsymbol{Z}^{(2)}$ of size $n \times p$ respectively. $\boldsymbol{B}_j$ is the jth column in the matrix $\boldsymbol{B}$, and $\boldsymbol{F}$ is the first-order difference matrix of size $(n-1)\times n$, written as

$$\boldsymbol{F} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ & & & & \cdots\cdots\cdots\cdots & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & -1 \end{bmatrix}.$$

The optimization in (3.4) is different from the original problem in (3.2) since the penalty terms now are involving $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$, which are completely decoupled. Therefore, we can solve the optimization by alternating minimization of $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$. We have the augmented

Lagrangian for the problem (3.4) as

$$
\begin{aligned}
&l_{al}(\boldsymbol{B}) \triangleq l_{\gamma_1,\gamma_2}(\boldsymbol{B})+\\
&\sum_{j=1}^{p}\tilde{\lambda}_j^{(1),T}(F\boldsymbol{B}_j-\boldsymbol{Z}_j^{(1)})+\sum_{j=1}^{p}\tilde{\lambda}_j^{(2),T}(\boldsymbol{B}_j-\boldsymbol{Z}_j^{(2)})+\frac{\rho_1}{2}\sum_{j=1}^{p}||F\boldsymbol{B}_j-\boldsymbol{Z}_j^{(1)}||_2^2+\frac{\rho_2}{2}\sum_{j=1}^{p}||\boldsymbol{B}_j-\boldsymbol{Z}_j^{(2)}||_2^2\\
&=l_{\gamma_1,\gamma_2}(\boldsymbol{B})+\\
&\frac{\rho_1}{2}\sum_{j=1}^{p}||\boldsymbol{F}\boldsymbol{\beta}_j-\boldsymbol{Z}_j^{(1)}+\boldsymbol{u}_j^{(1)}||_2^2+\frac{\rho_2}{2}\sum_{j=1}^{p}||\boldsymbol{\beta}_j-\boldsymbol{Z}_j^{(2)}+\boldsymbol{u}_j^{(2)}||_2^2-\frac{\rho_1}{2}\sum_{j=1}^{p}||\boldsymbol{u}_j^{(1)}||_2^2-\frac{\rho_2}{2}\sum_{j=1}^{p}||\boldsymbol{u}_j^{(2)}||_2^2,
\end{aligned}
$$

where $\boldsymbol{\mu}_j^{(1)}=\frac{\tilde{\lambda}_j^{(1)}}{\rho_1}$ and $\boldsymbol{\mu}_j^{(2)}=\frac{\tilde{\lambda}_j^{(2)}}{\rho_2}$. Here, ρ_1 and ρ_2 are the augmented Lagrangian parameters for the l_2-norm group lasso penalty and the l_1-norm fused penalty, respectively, and $\tilde{\lambda}_j^{(1)}$ and $\tilde{\lambda}_j^{(2)}$ are the Lagrangian multipliers.

Then we can obtain the iterative updating scheme as

$$
\boldsymbol{\beta}^{k+1}=\underset{\boldsymbol{B}}{\text{argmin}}\bigg(-l(\boldsymbol{B})+\frac{\rho_1^k}{2}\sum_{j=1}^{p}||\boldsymbol{F}\boldsymbol{\beta}_j-\boldsymbol{Z}_j^{(1)}+\boldsymbol{u}_j^{(1)}||_2^2)+\frac{\rho_2^k}{2}\sum_{j=1}^{p}||\boldsymbol{\beta}_j-\boldsymbol{Z}_j^{(2)}+\boldsymbol{\mu}_j^{(2)}||_2^2\bigg), \tag{3.5}
$$

$$
\begin{aligned}
\boldsymbol{Z}_j^{(2),k+1}&=f_{ss}(\boldsymbol{B}_j^{k+1}+\boldsymbol{u}_j^{(2),k},\gamma_2/\rho_2^k),\ j=1,\cdots,p;\\
\boldsymbol{Z}_j^{(1),k+1}&=f_{ss}(F\boldsymbol{B}_j^{k+1}+\boldsymbol{u}_j^{(1),k},\gamma_1/\rho_1^k),\ j=1,\cdots,p;\\
\boldsymbol{u}_j^{(2),k+1}&=\boldsymbol{u}_j^{(2),k}+\boldsymbol{B}_j^{k+1}-\boldsymbol{Z}_j^{(2),k+1},\ j=1,\cdots,p;\\
\boldsymbol{u}_j^{(1),k+1}&=\boldsymbol{u}_j^{(1),k}+F\boldsymbol{B}_j^{k+1}-\boldsymbol{Z}_j^{(1),k+1},\ j=1,\cdots,p,
\end{aligned}
$$

with the soft-shrinkage function f_{ss} given as

$$
f_{ss}(a,b)=\text{sign}(a)\max((|a|-b),0),
$$

where ρ_1^k and ρ_2^k are the augmented Lagrangian parameters for the fused penalty and the

group penalty at iteration k, respectively.

Updating the dual variables, $\boldsymbol{u}_j^{(1)}$ and $\boldsymbol{u}_j^{(2)}$, and the primal variables, $\boldsymbol{Z}_j^{(1)}$ and $\boldsymbol{Z}_j^{(2)}$, are straightforward. Therefore, the efficiency of the algorithm depends on the minimization of $\boldsymbol{B}$ in (5), which is similar to the minimization problem of the classical logistic regression except the two additional quadratic terms. It is known that there is no analytical solution existing for the classical logistic regression problem. Thus, we apply the Newton-Raphson method solve the minimization problem. Specifically, we approximate $l(\boldsymbol{B})$ with its second-order Taylor series as

$$l(\boldsymbol{B}) \approx -\sum_{t=1}^{n}\left[l(\boldsymbol{\beta}_t^{(0)}) + \big[\frac{\partial l(\boldsymbol{\beta}_t)}{\partial \boldsymbol{\beta_t}}(\boldsymbol{\beta}_t^{(0)})\big]^T(\beta_t - \boldsymbol{\beta}_t^{(0)}) + \frac{1}{2}(\beta_t - \boldsymbol{\beta}_t^{(0)})^T H(\boldsymbol{\beta}_t^{(0)})(\beta_t - \boldsymbol{\beta}_t^{(0)})\right],$$

where

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta}_{t,j})}{\partial \boldsymbol{\beta_{t,j}}}(\boldsymbol{\beta}_t^{(0)}) =& -\boldsymbol{x_{t,j}}(y_t - p(\boldsymbol{x_{t,j}})),\\
p(\boldsymbol{x}_{t,j}) =& \frac{\exp(\boldsymbol{x}_{t,j}^T\boldsymbol{\beta}_{t,j})}{1+\exp(\boldsymbol{x}_{t,j}^T\boldsymbol{\beta}_{t,j})},\\
H(\boldsymbol{\beta}_t^{(0)}) =& \frac{\partial^2 L(\boldsymbol{\beta}_t^{(0)})}{\partial\beta_t\partial^T\beta_t} - \sum_{t=1}^{n} x_t x_t^T p(x_t|\boldsymbol{\beta}_t^{(0)})(1-p(x_t|\boldsymbol{\beta}_t^{(0)})).
\end{aligned}$$

We apply the standard ADMM stopping criterion based on primal and dual residuals, which are defined at iteration $k+1$ (Boyd et al. 2011) as:

$$\begin{aligned}
\boldsymbol{r}_j^{(2),k+1} &= \boldsymbol{B}_j^{k+1} - \boldsymbol{Z}_j^{(2),k+1},\\
\boldsymbol{s}_j^{(2),k+1} &= \rho_{gr}^k(\boldsymbol{Z}_j^{(2),k+1} - \boldsymbol{Z}_j^{(2),k}),\\
\boldsymbol{r}_j^{(1),k+1} &= F\boldsymbol{B}_j^{k+1} - \boldsymbol{Z}_j^{(1),k+1},\\
\boldsymbol{s}_j^{(1),k+1} &= \rho_{fuse}^k F^T(\boldsymbol{Z}_j^{(1),k+1} - \boldsymbol{Z}_j^{(1),k}),
\end{aligned}$$

where $\boldsymbol{r}_j^{(i)}$ and $\boldsymbol{s}_j^{(i)}$ are the jth columns in the $n \times p$ matrices $\boldsymbol{r}^{(i)}$ and $\boldsymbol{s}^{(i)}$, respectively, $i = 1$, 2. The suggested termination criterion is that the primal and dual residuals must be small as the ADMM algorithm proceeds (Boyd et al. 2011), i.e.,

$$||\boldsymbol{r}_{vec}^{(i),k}||_2 \leq \epsilon^{(i),pri} \text{ and } ||\boldsymbol{s}_{vec}^{(i),k}||_2 \leq \epsilon^{(i),dual}, \ i = 1, 2$$

with

$$\begin{aligned}
\epsilon^{(2),pri} &= \sqrt{np}\epsilon^{abs} + \epsilon^{rel} \ max\left(||vec(\boldsymbol{B}^k)||_2, ||\boldsymbol{Z}_{vec}^{(2),k}||_2\right), \\
\epsilon^{(2),dual} &= \sqrt{np}\epsilon^{abs} + \epsilon^{rel}||vec(\rho_2^k u^{(2),k})||_2, \\
\epsilon^{(1),pri} &= \sqrt{(n-1)p}\epsilon^{abs} + \epsilon^{rel} \ max\left(||vec(F\boldsymbol{B}^k)||_2, ||\boldsymbol{Z}_{vec}^{(1),k}||_2\right), \\
\epsilon^{(1),dual} &= \sqrt{np}\epsilon^{abs} + \epsilon^{rel}||vec(\rho_1^k F^T u^{(1),k})||_2,
\end{aligned}$$

where $\boldsymbol{Z}_{vec}^{(1)} = (\boldsymbol{Z}_1^{(1),T}, \cdots, \boldsymbol{Z}_p^{(1),T},)^T$, $\boldsymbol{Z}_{vec}^{(2)} = (\boldsymbol{Z}_1^{(2),T}, \cdots, \boldsymbol{Z}_p^{(2),T},)^T$, $\boldsymbol{r}_{vec}^{(1)} = (\boldsymbol{r}_1^{(1),T}, \cdots, \boldsymbol{r}_p^{(1),T},)^T$, $\boldsymbol{r}_{vec}^{(2)} = (\boldsymbol{r}_1^{(2),T}, \cdots, \boldsymbol{r}_p^{(2),T})^T$, vec$(\cdot)$ is an operator for the vectorization of a matrix, ϵ^{abs} is the absolute tolerance, and ϵ^{abs} is the relative tolerance..

Algorithm 1 summarizes the developed computational algorithm for parameter estimation of the proposed method. The algorithm is implemented in the R package SeqADMM and is available in Bitbucket (`https://bitbucket.org/vtshen/rpackages/src/master/`).

We would like to remark that the convergence speed depends heavily on the choice of the augmented Lagrangian parameters ρ_1 and ρ_2. Zhu (2015) proposed a varying penalty strategy for updating the parameter ρ. That is,

$$\rho^{k+1} = \begin{cases} \eta\rho^k & \text{if } \boldsymbol{r}_{vec2}^k/\epsilon^{pri} \geq \mu\boldsymbol{s}_{vec2}^k/\epsilon^{dual}, \\ \eta^{-1}\rho^k & \text{if } \boldsymbol{s}_{vec2}^k/\epsilon^{dual} \geq \mu\boldsymbol{r}_{vec2}^k/\epsilon^{pri}, \\ \rho^k & \text{otherwise.} \end{cases}$$

Algorithm 2 ADMM

Input: $\boldsymbol{X}$, $\boldsymbol{y}$, γ_1 and γ_2
Initialize $\boldsymbol{B}$, $\boldsymbol{Z}^{(1)}$, $\boldsymbol{Z}^{(2)}$, and $\boldsymbol{u}^{(1)}$, $\boldsymbol{u}^{(2)} = \boldsymbol{0}$.
for iteration k **do**
  $\boldsymbol{B}$-update
    Compute the $\frac{\partial^2 \boldsymbol{L}}{\partial B \partial B^T}$ and $\frac{\partial \boldsymbol{L}}{\partial B}$.
    Find appropriate step size η_s by the backtracking line search strategy.
    $\boldsymbol{B}^{k+1} = \boldsymbol{B}^k + \eta_s \times \big[- (\frac{\partial^2 \boldsymbol{L}}{\partial B \partial B^T})^{-1} \frac{\partial \boldsymbol{L}}{\partial B} \big]$.
  $\boldsymbol{Z}^{(2)}$- and $\boldsymbol{Z}^{(1)}$-update in (5).
  $\boldsymbol{U}^{(2)}$- and $\boldsymbol{U}^{(1)}$-update in (5).
  if $||\boldsymbol{r}||_2^{(g)} \leq \epsilon^{(g),pri} \wedge ||\boldsymbol{s}||_2^{(g)} \leq \epsilon^{(g),dual}$, g=1,2 **then** stop.
  end if
end for
Return $\boldsymbol{Z}^{(2)}$.

where ρ denotes ρ_1 and ρ_2, η and μ are set to be 2 and 10 as suggested in Boyd et al. (2011). The idea of this strategy is to improve the algorithm convergence when primal and dual feasibilities are on different scales.

The ADMM algorithm has been proved to be quite flexible in many large scale statistical estimation problems (Boyd et al. 2011). Ye and Xie (2011) developed a split Bregman method, which is basically equivalent to ADMM, for large scale fused lasso problems. Furthermore, the ADMM method is flexible to extend the fused lasso problem to higher order trend filtering problems (Ramdas and Tibshirani 2016) or other different types of penalties.

Selection of Regularization Parameters

The regularization parameters, (γ_1, γ_2), are determined over a search grid by the so-called structured 5-fold cross-validation technique (Arnold and Tibshirani 2016). That is, the data set is ordered (for example, indexed by time) and divided into five folds such that every fifth point is in the same fold. We train the proposed model on all observations except those in

the kth fold, and compute the mean squared cross-validation (MSCV) error defined as

$$\text{MSCV} = \frac{1}{5}\sum_{k=1}^{5}\left[\frac{1}{n_k}\sum_{i\in F_k}(y_i - \hat{y}_{i,-k})^2\right],$$

where n_k is the number of observations in the kth fold and $\hat{y}_{i,-k}$ is the prediction at the ith point. In particular, $\hat{y}_{i,-k}$ is obtained as the weighted average of the fitted values at positions i-1 and i+1 in this study.

3.4 Simulation

In this section, we conduct the simulation studies to evaluate the performance of the proposed regularized dynamic logistic regression model. The response, y_t, follows a Bernoulli distribution with the conditional probability $Pr(y_t = 1|\boldsymbol{x}_t)$. We assume that the underlying model $\text{logit}(\boldsymbol{x}_t) = \boldsymbol{x}_t^T\boldsymbol{\beta}_t$. We fix the number of significant variables with nonzero coefficient as five. The rest insignificant variables are noise variables used for high-dimension settings. In the simulation, we consider two scenarios of generating predictor variables $\boldsymbol{X}_j$. In scenario 1 (S1), the predictor matrix $\boldsymbol{X}$ follows a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a covariance matrix of size $p \times p$. We take the absolute value of $\boldsymbol{X}_j$ such that the probability $Pr(y_t = 1|\boldsymbol{x}_t)$ is greater than 0.5 when the sign of coefficient $\boldsymbol{\beta}_t$ is positive. Note that $\boldsymbol{X}_j$ is independent over time. In scenario 2 (S2), the predictor variable $\boldsymbol{X}_j$ is an autocorrelated sequence over time. The instance $x_{t,j}$ is generated from the AR(1) model $x_{t+1,j} = 0.7x_{t,j} + w_{t,j}$, where $t = 1, \cdots, n-1$, $x_{1,j} = 0$. and $\boldsymbol{W} = (w_{t,j})_{n\times p}$ follows a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. Note that $\boldsymbol{\Sigma}$ in both scenarios are equal to $(\rho^{|i-j|})_{p\times p}$ with $\rho^{|i-j|}$ as the element in the ith row and jth column in $\boldsymbol{\Sigma}$ and the correlation parameter $\rho \geq 0$.

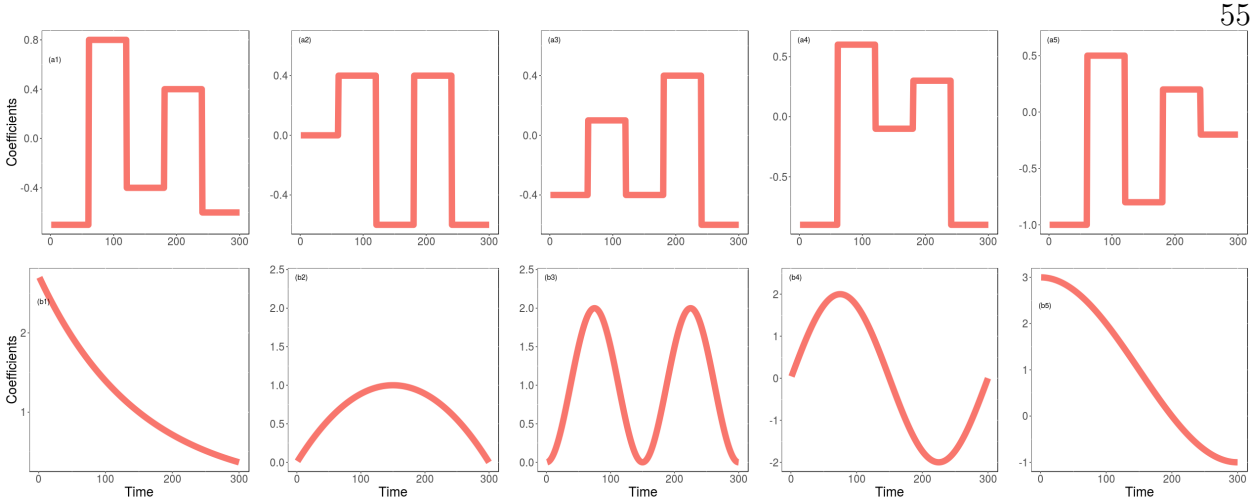

Figure 3.1: Illustrative plots for the simulated dynamic coefficients of the five significant variables when $n = 300$: (top row) the piecewise constant coefficients in case (a) and (bottom row) the smooth functional coefficients in case (b).

To conduct a comprehensive simulation study, we vary several settings, including the sample size, the number of predictor variables, and the patterns of coefficients over time. Specifically, we consider five different sample sizes, $n = 100$, 200, 300, 400, and 500; two different number of variables, $p = 20$ and 50; two different scenarios, S1 and S2; three different correlations, $\rho = 0$, 0.35, and 0.7; and four different cases for the coefficients of significant variables: (a) piecewise constant coefficients over time segments, shown in Figure 3.1 (top row); (b) smooth functional coefficients, shown in Figure 3.1 (bottom row); (c) constant coefficients over time; and (d) both smooth functional coefficients and constant coefficients. Specially, in case (a), the magnitudes of $\boldsymbol{\beta}_t$ follow a uniform distribution. The sign of $\boldsymbol{\beta_t}$ alternates on adjacent segments. In order to generate the piecewise constant $\boldsymbol{\beta}_t$, we first partition the time range into several segments of pre-defined lengths. Then, from the uniform distribution $U(0,\ 1)$ we sample values for $\beta_{t,j}$ in each segment. It is clear that the coefficient $\beta_{t,j}$ is piecewise constant over time. Note that the coefficients have alternating signs in adjacent segments. In case (b), the smooth functions for the five significant features are $f(t) = exp(-2t + 1)$,

$f(t) = 4t(1-t)$, $f(t) = 2sin^2(2\pi t)$, $f(t) = 2sin(2\pi t)$, and $f(t) = 2cos(\pi t) + 1$. In case (c), the constant magnitudes of $\boldsymbol{\beta}_t$ follow a uniform distribution. The sign of $\boldsymbol{\beta_t}$ alternates on adjacent segments. In case (d), the first three of five significant variables are the same as that in case (b), and the rest two significant variables have constant magnitudes following a uniform distribution and alternating signs.

For each simulation setting, we perform 30 replications. The proposed regularized dynamic logistic regression model (rDLR) is compared with six other models, which are: (i) least absolute shrinkage and selection operator (LASSO), (ii) multivariate adaptive regression splines (MARS), (iii) varying coefficient model with smoothing splines basis (VCM1), (iv) varying coefficient model with polynomial of degree ≤ 2 basis (VCM2), (v) dynamic logistic regression with fused penalty (BM1), and (vi) dynamic logistic regression with fused penalty and l_1-norm penalty (BM2).

The LASSO is an l_1-norm regularized method widely used in high dimensional problems where the standard linear regression fails (Tibshirani 1996). The LASSO tends to produce an interpretable model with certain sparsity structure. The LASSO solves the following optimization problem

$$\underset{\boldsymbol{\theta} \in R^p}{\text{minimize}} \; -l(\boldsymbol{X}, \boldsymbol{\theta}) + \lambda ||\boldsymbol{\theta}||_1,$$

where $l(\boldsymbol{X}, \boldsymbol{\theta}) = \sum_{t=1}^{n} \; y_t \theta^T x_t - log(1 + e^{\theta^T x_t})$ is the log-likelihood function of the logistic regression with response being either zero or one, $\lambda \geq 0$ is the tuning parameter, and $||\boldsymbol{\theta}||_1$ is the l_1-norm penalty on the coefficient vector $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^T$. Note that we only consider main effects in the logistic regression model.

The MARS is a piecewise linear regression model but allows flexible fitting by the automatic selection of spline basis functions and knots (Frideman 1991). The MARS solves the

regression problem

$$\log\frac{p(\boldsymbol{x}_t)}{1-p(\boldsymbol{x}_t)} = \sum_{j=1}^{p}\sum_{m=1}^{M_j} B_{m,j}(\boldsymbol{X}_j)a_{m,j},$$

where $B_{m,j}(\boldsymbol{X}_j)$ is the basis function for the predictor $\boldsymbol{X}_j$, $a_{m,j}$ is the mth coefficient of the basis function $B_{m,j}(\boldsymbol{X}_j)$, and M_j is the number of knots determined for the predictor $\boldsymbol{X}_j$.

The VCM1 solves the regression problem as

$$\log\frac{p(\boldsymbol{x}_t)}{1-p(\boldsymbol{x}_t)} = \sum_{j=1}^{p} f_j(t)\boldsymbol{x}_j,\ s.t.\ \int f''(u)^2 du \leq \lambda,$$

where f is represented by the smoothing spline basis. Different from the VCM1, the VCM2 uses a polynomial basis with degree ≤ 2 and solves the regression problem as

$$\log\frac{p(\boldsymbol{x}_t)}{1-p(\boldsymbol{x}_t)} = \sum_{j=1}^{p} f_j(t)\boldsymbol{x}_j.$$

The BM1 solves the optimization problem as

$$\underset{\boldsymbol{B}}{\text{minimize}}\ -l(\boldsymbol{B}) + \lambda\sum_{j=1}^{p}||F\boldsymbol{B}_j||_1,\ j=1,\cdots,p.$$

Compared to the proposed rDLR, the BM1 does not have the l_2-norm group penalty term.

The BM2 solves the optimization problem as

$$\underset{\boldsymbol{B}}{\text{minimize}}\ -l(\boldsymbol{B}) + \lambda_1\sum_{j=1}^{p}||F\boldsymbol{B}_j||_1 + \lambda_2\sum_{t=1}^{n}\sum_{j=1}^{p}|\beta_{ij}|,\ j=1,\cdots,p.$$

Compared to the proposed rDLR, the BM2 has the l_1-norm penalty on coefficient parameters at each time point. Note that we do not include the intercept term in the methods VCM1, VCM2, BM1, and BM2 in this study. We use the available R packages glmnet, earth,

and longfused to implement the methods LASSO, MARS, and BM2, respectively. We implemented the methods rDLR, VCM1, VCM2, and BM1 in the R software package SeqADMM,

To evaluate the model prediction performance and coefficient-estimation performance, we employ two sets of evaluation metrics: the prediction metrics and the parameter estimation metrics. The prediction metrics include deviance (DEV) and misclassification error rate (MER). The DEV and MER are computed in the 5-fold cross-validation. Specifically, the data set is divided into five folds, among which four folds are used to train the model, and the left-out portion is used to test the model and compute the metrics DEV and MER. Here, we point out that there are two types of cross-validation (CV) mechanisms used in this study: the normal random CV and the so-called structured CV. The normal random CV splits the partitions by random sampling. The structured CV is described in details in Section 3. We apply the normal random CV to the LASSO and MARS, and the structured CV to the rDLR, VCM1, and VCM2, BM1, and BM2.

The DEV measures the prediction performance and it is defined in terms of the difference between the negative log-likelihood of model (nll_1) and the negative log-likelihood of the saturated model (nll_2):

$$\text{DEV} = \frac{2 \times (\text{nll}_1 \text{ - } \text{nll}_2)}{m},$$

where m is the number of observations used in computing nll_1. The saturated model has a free parameter for each observation. The MER provides a high-level idea about the method's accuracy. MER is defined as the ratio of number of incorrectly-classified observation to the number of total observations.

$$\text{MER} = 1 - \frac{\text{TP} + \text{TN}}{n},$$

where TP is the number of true positive and TN is the number of true negative. For both DEV and MER, the smaller their values are, the higher the prediction accuracy is.

The parameter estimation metrics are performance measurement (PM), correctly identified coefficient rate (CICR), number of non-zeros (NZ), and F_1 score. The PM evaluates the estimation accuracy of estimated $\hat{\beta}_{t,j}$ relative to the true $\beta^*_{t,j}$. The PM is defined as

$$\text{PM} = \frac{\sum_{t=1}^{n}\sum_{j=1}^{p}|\hat{\beta}_{t,j} - \beta^*_{t,j}|}{p\sum_{t=1}^{n}\sum_{j=1}^{p}|\beta^*_{t,j}|},$$

The smaller the PM is, the higher the estimation accuracy of coefficient parameter is. The CICR, evaluating the ability to identify correct variables, is defined as ratio of sum of the numbers of correctly identified coefficients as zero and as non-zero over the total number of coefficients. The larger the CICR is, the better the identification of correct variables is. The NZ measures the number of non-zero values in the estimated coefficients. For LASSO, the NZ is equal to the model size. For rDLR, VCM1, VCM2, BM1, and BM2, the NZ is an approximation of the model size after normalization with respect to the number of observations. The F_1 score evaluates the ability of parameter identification. The F_1 score is defined as

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}},$$

where FN is the number of false negative and FP is the number of false positive. Note that all TP, TN, FP, and FN are normalized by the total number of observations in the methods rDLR, BM1, and BM2. The larger the F_1 score is, the better the identification of variables is. The perfect F_1 score is one.

Table 3.1 shows the mean and standard deviation of DEV obtained from the different methods under comparison. The proposed rDLR has the smallest DEVs in cases (a), (b), and (d). The feature of the dynamic coefficients in the rDLR improves the model performance, considering that the LASSO and MARS have constant coefficients. The feature of the variable selection in the rDLR also contribute to the model performance, comparing the methods,

BM1 and BM2, which assume dynamic coefficients, but lack the variable selection feature. The models VCM1 and VCM2 are outperformed by the model rDLR in this study, which may be explained by their lack of the variable selection feature. The VCM1 has a better model performance than the VCM2 in all cases. In case (c), the rDLR has similar DEVs as the LASSO. In addition, the DEV values from the rDLR decreases as sample size increases.

Table 3.1: Performance comparisons of models in terms of deviance (DEV) when $\rho = 0.35$, $p = 20$, and the scenario is S2 for the four different cases of underlying functional coefficients, (a), (b), (c), (d), from 30 simulation replications (mean and standard errors (in parenthesis)).

p	Method	n	(a)	(b)	(c)	(d)
20	BM1	100	1.86 (0.24)	1.97 (0.91)	2.33 (1.13)	2.39 (1.18)
		200	1.44 (0.08)	1.21 (0.14)	1.16 (0.15)	1.09 (0.09)
		300	1.38 (0.08)	1.13 (0.25)	1.11 (0.14)	1.00 (0.09)
		400	1.30 (0.07)	1.07 (0.28)	1.06 (0.09)	0.99 (0.05)
		500	1.23 (0.07)	0.96 (0.25)	1.04 (0.11)	0.93 (0.08)
	BM2	100	1.34 (0.05)	1.25 (0.10)	1.30 (0.09)	1.22 (0.12)
		200	1.33 (0.03)	1.24 (0.08)	1.30 (0.08)	1.28 (0.07)
		300	1.33 (0.03)	1.23 (0.08)	1.32 (0.06)	1.27 (0.07)
		400	1.33 (0.03)	1.24 (0.06)	1.34 (0.03)	1.28 (0.06)
		500	1.33 (0.02)	1.22 (0.07)	1.32 (0.05)	1.24 (0.05)
	LASSO	100	1.36 (0.05)	1.14 (0.13)	1.10 (0.14)	1.05 (0.13)
		200	1.36 (0.04)	1.13 (0.09)	1.06 (0.14)	0.97 (0.09)
		300	1.37 (0.02)	1.13 (0.06)	**1.06** (0.14)	0.97 (0.09)
		400	1.35 (0.03)	1.15 (0.06)	**1.03** (0.08)	0.98 (0.07)
		500	1.37 (0.02)	1.15 (0.07)	**1.01** (0.11)	0.94 (0.09)
	MARS	100	4.15 (2.88)	5.71 (4.42)	3.98 (2.62)	4.36 (3.52)
		200	1.81 (0.21)	1.68 (0.30)	1.53 (0.26)	1.45 (0.28)
		300	1.62 (0.13)	1.40 (0.14)	1.29 (0.16)	1.20 (0.16)
		400	1.53 (0.13)	1.36 (0.12)	1.21 (0.12)	1.15 (0.10)
		500	1.50 (0.05)	1.28 (0.09)	1.16 (0.13)	1.09 (0.13)
	rDLR	100	**1.30** (0.07)	**0.97** (0.15)	1.07 (0.14)	**1.01** (0.13)
		200	**1.23** (0.07)	**0.83** (0.09)	**1.04** (0.14)	**0.94** (0.09)
		300	**1.20** (0.08)	**0.79** (0.08)	**1.05** (0.15)	**0.91** (0.07)
		400	**1.18** (0.07)	**0.78** (0.06)	**1.02** (0.09)	**0.92** (0.06)
		500	**1.15** (0.06)	**0.78** (0.05)	**1.01** (0.11)	**0.86** (0.08)
	VCM1	100	2.06 (0.24)	1.87 (0.31)	2.16 (0.36)	2.06 (0.25)
		200	1.52 (0.13)	1.27 (0.15)	1.52 (0.12)	1.47 (0.11)
		300	1.41 (0.08)	1.18 (0.09)	1.51 (0.08)	1.40 (0.07)
		400	1.35 (0.05)	1.12 (0.08)	1.44 (0.04)	1.37 (0.06)
		500	1.31 (0.06)	1.10 (0.08)	1.40 (0.05)	1.33 (0.05)
	VCM2	100	4.00 (0.96)	3.36 (1.57)	4.37 (1.26)	3.78 (1.18)
		200	1.81 (0.16)	1.69 (0.21)	1.80 (0.13)	1.78 (0.13)
		300	1.52 (0.09)	1.34 (0.10)	1.61 (0.07)	1.53 (0.08)
		400	1.44 (0.04)	1.23 (0.09)	1.50 (0.04)	1.44 (0.06)
		500	1.38 (0.06)	1.18 (0.08)	1.45 (0.04)	1.40 (0.04)

Table 3.2 shows the mean and standard deviation of MER obtained from the different models under comparison. The results are quite similar to that in Table 3.1. The rDLR generally

outperforms the methods LASSO, MARS, VCM1, VCM2, BM1, and BM2.

Table 3.2: Performance comparisons of models in terms of misclassification error rate (MER) when $\rho = 0.35$, $p = 20$, and the scenario is S2 for the four different cases of underlying functional coefficients, (a), (b), (c), (d), from 30 simulation replications (mean and standard errors (in parenthesis)).

p	Method	n	(a)	(b)	(c)	(d)
20	BM1	100	0.42 (0.07)	0.30 (0.06)	0.31 (0.06)	0.27 (0.05)
		200	0.39 (0.06)	0.35 (0.12)	0.28 (0.06)	0.25 (0.04)
		300	0.39 (0.06)	0.35 (0.16)	0.28 (0.06)	0.24 (0.03)
		400	0.36 (0.04)	0.32 (0.15)	0.26 (0.03)	0.23 (0.02)
		500	0.33 (0.03)	0.27 (0.16)	0.26 (0.04)	0.21 (0.03)
	BM2	100	0.40 (0.04)	0.35 (0.05)	0.37 (0.06)	0.31 (0.07)
		200	0.41 (0.03)	0.34 (0.04)	0.37 (0.06)	0.35 (0.05)
		300	0.43 (0.04)	0.34 (0.04)	0.38 (0.05)	0.35 (0.04)
		400	0.43 (0.03)	0.34 (0.04)	0.40 (0.04)	0.36 (0.05)
		500	0.42 (0.02)	0.33 (0.03)	0.39 (0.05)	0.34 (0.03)
	LASSO	100	0.38 (0.05)	0.27 (0.05)	0.27 (0.06)	0.24 (0.04)
		200	0.41 (0.04)	0.28 (0.04)	**0.25** (0.06)	0.23 (0.03)
		300	0.41 (0.04)	0.28 (0.03)	**0.26** (0.06)	0.23 (0.03)
		400	0.42 (0.03)	0.29 (0.03)	**0.25** (0.03)	0.23 (0.03)
		500	0.43 (0.03)	0.30 (0.03)	**0.24** (0.04)	0.21 (0.03)
	MARS	100	0.47 (0.06)	0.35 (0.08)	0.35 (0.08)	0.30 (0.06)
		200	0.44 (0.05)	0.32 (0.05)	0.32 (0.05)	0.28 (0.04)
		300	0.44 (0.04)	0.32 (0.03)	0.30 (0.06)	0.25 (0.04)
		400	0.44 (0.03)	0.32 (0.03)	0.28 (0.04)	0.26 (0.02)
		500	0.45 (0.03)	0.32 (0.03)	0.27 (0.05)	0.24 (0.03)
	rDLR	100	**0.33** (0.04)	**0.21** (0.05)	**0.25** (0.05)	**0.21** (0.05)
		200	**0.32** (0.04)	**0.18** (0.03)	**0.25** (0.05)	**0.21** (0.03)
		300	**0.31** (0.03)	**0.17** (0.03)	**0.26** (0.06)	**0.20** (0.02)
		400	**0.30** (0.03)	**0.17** (0.02)	**0.25** (0.03)	**0.21** (0.02)
		500	**0.29** (0.03)	**0.17** (0.02)	**0.24** (0.04)	**0.19** (0.02)
	VCM1	100	0.40 (0.06)	0.27 (0.07)	0.39 (0.08)	0.36 (0.06)
		200	0.38 (0.05)	0.26 (0.04)	0.38 (0.06)	0.34 (0.04)
		300	0.37 (0.04)	0.27 (0.03)	0.42 (0.04)	0.36 (0.04)
		400	0.36 (0.03)	0.26 (0.03)	0.41 (0.04)	0.36 (0.04)
		500	0.35 (0.03)	0.26 (0.03)	0.41 (0.04)	0.36 (0.03)
	VCM2	100	0.41 (0.08)	0.26 (0.10)	0.39 (0.11)	0.35 (0.10)
		200	0.42 (0.05)	0.29 (0.04)	0.41 (0.05)	0.38 (0.05)
		300	0.41 (0.04)	0.29 (0.04)	0.45 (0.04)	0.40 (0.04)
		400	0.41 (0.03)	0.28 (0.04)	0.44 (0.04)	0.39 (0.03)
		500	0.40 (0.03)	0.28 (0.03)	0.44 (0.04)	0.39 (0.03)

Figure 3.2 shows the simulation results of the coefficient-estimation performance among compared methods when the dimension p is 20, ρ is 0.35, and the scenario is S2. The rDLR and LASSO outperform the other models in PM values, and the VCM1 and VCM2 has the worst PM performance in all cases (a), (b), (c), and (d). The rDLR has the highest F_1 values and the BM2 has the lowest F_1 values. The BM1, VCM1, and VCM2 has constant F_1 values since these three methods selected all variables. This is consistent with the results of their constant CICR and number of nonzero values. The rDLR has selected a larger number of variables than the LASSO and BM2, but smaller values than the BM1, VCM1, and VCM2. When smaller number of variables were selected, considering the setting of low sparsity in the true variables, it is more likely for the methods to have larger TN values, leading to higher CICR values. This is confirmed by the CICR plots. We do not include the method MARS because MARS does not compute the estimates of coefficients for the original variables.

3.5 Real-Data Analysis

In this section, we analyze three real-data problems studied previously in the literature, the crystal ingot growth experiment (Jin et al. 2019), the Hong Kong environmental study (Fan and Chen 1997; Cai et al. 2000), and the photodegradation experiment (Gu et al. 2009; Hong et al. 2015) to evaluate the model performance of the proposed regularized dynamic logistic regression model.

3.5.1 Crystal Ingot Growth Experiments

The quality of the silicon ingot produced from the crystal ingot growth experiment is fundamental to its downstream products such as wafers, solar cells and integrated circuits. The goal of this study is to identify key process variables and characterize the dynamic effects

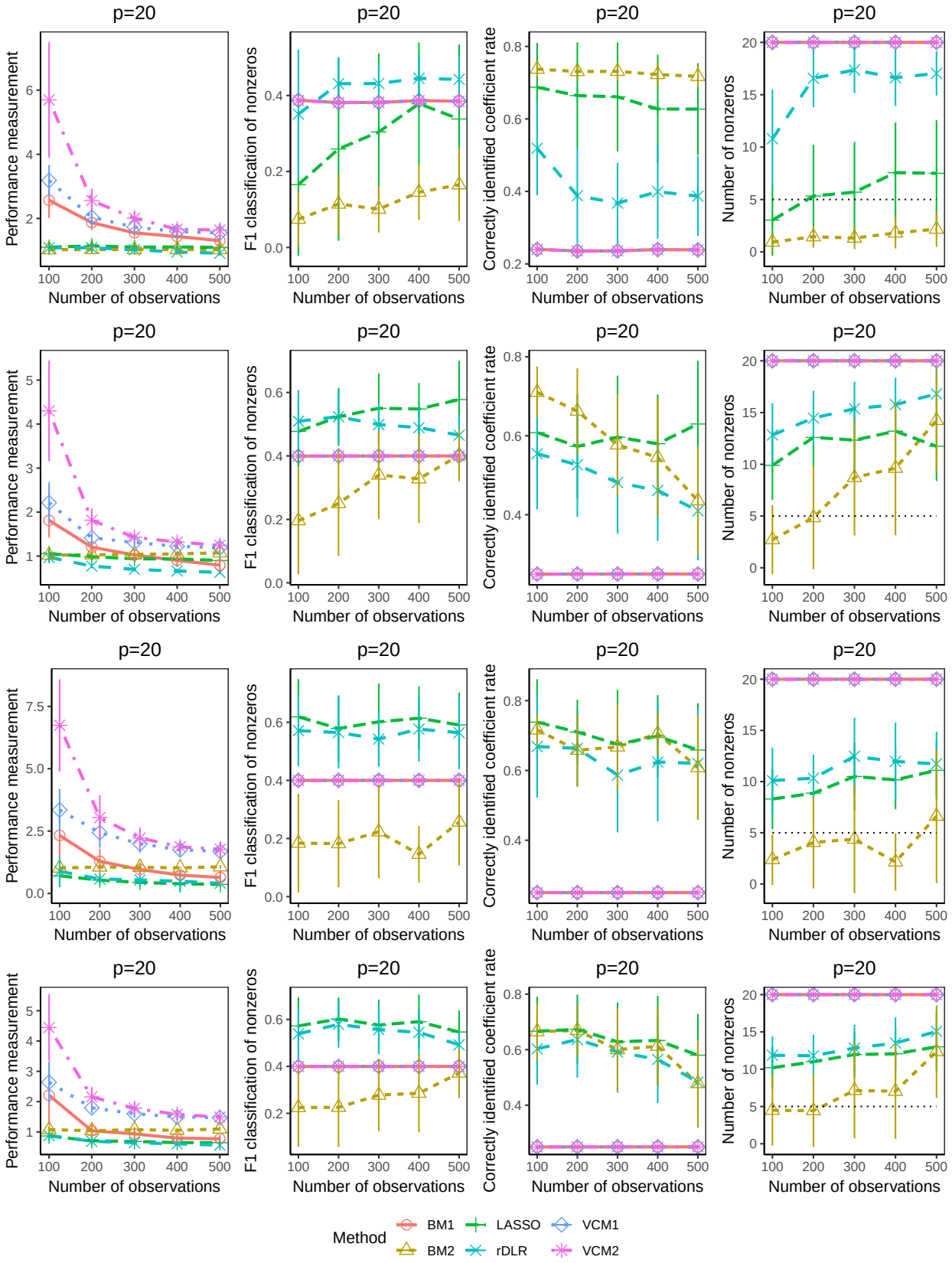

Figure 3.2: Performance comparison of models in the coefficient estimation for the cases (from top row to bottom) (a), (b), (c), and (d) when $\rho = 0.35$, $p = 20$, and the scenario is S2. The dash lines are the true values.

of those process variables on the quality response during the body growth stage in the crystal ingot growth experiment. The key criterion in the body growth stage is to control the ingot diameter such that the diameter is not smaller than the target and stay as close as possible to the target. In practice, the real diameter is to some extent larger than the target value to avoid the case that the whole ingot becomes useless (Zhang et al. 2014; Sun et al. 2016; Jin et al. 2019). In this study, the continuous response diameter is transformed into a binary variable defined as one when the quality response falls within the lower and upper limit specifications and as zero otherwise. The process variables include the power and the temperature of the heater, the pulling speed, the rotation speed, and so forth. The process variables are positive continuous and normalized between 0 and 1. The total number of process variables is 15. In total, there are over 1600 observations.

Figure 3.3 shows the estimated coefficients from different models including LASSO, rDLR, BM1, BM2, VCM1, and VCM2. The MARS is not included in the real-data study because it does not output the estimated coefficients for the original feature variables. It is clear that the estimated coefficients from the rDLR vary over time. The varying-coefficient effects of variables can be explained by the growth of ingot and the degradation of equipment (Jin et al. 2009). Taking the variable pulling speed as an example, the engineering knowledge states that the faster the pulling speed, the more shrinkage effect on the ingot diameter. When the ingot growth is at the early stage, the ingot is short and easy to be shrunk by increasing the pulling speed. At the late stage, the ingot is almost grown and the effect of the pulling speed is limited. It is harder to shrink the ingot diameter by the pulling speed. The BM1 shows that the estimated coefficients are constant over time. The BM2 has more varying estimated coefficients compared to the rDLR. The VCM methods have identified one dominant significant variables.

Figure 3.4 shows the estimated coefficient for each feature variable over time obtained from

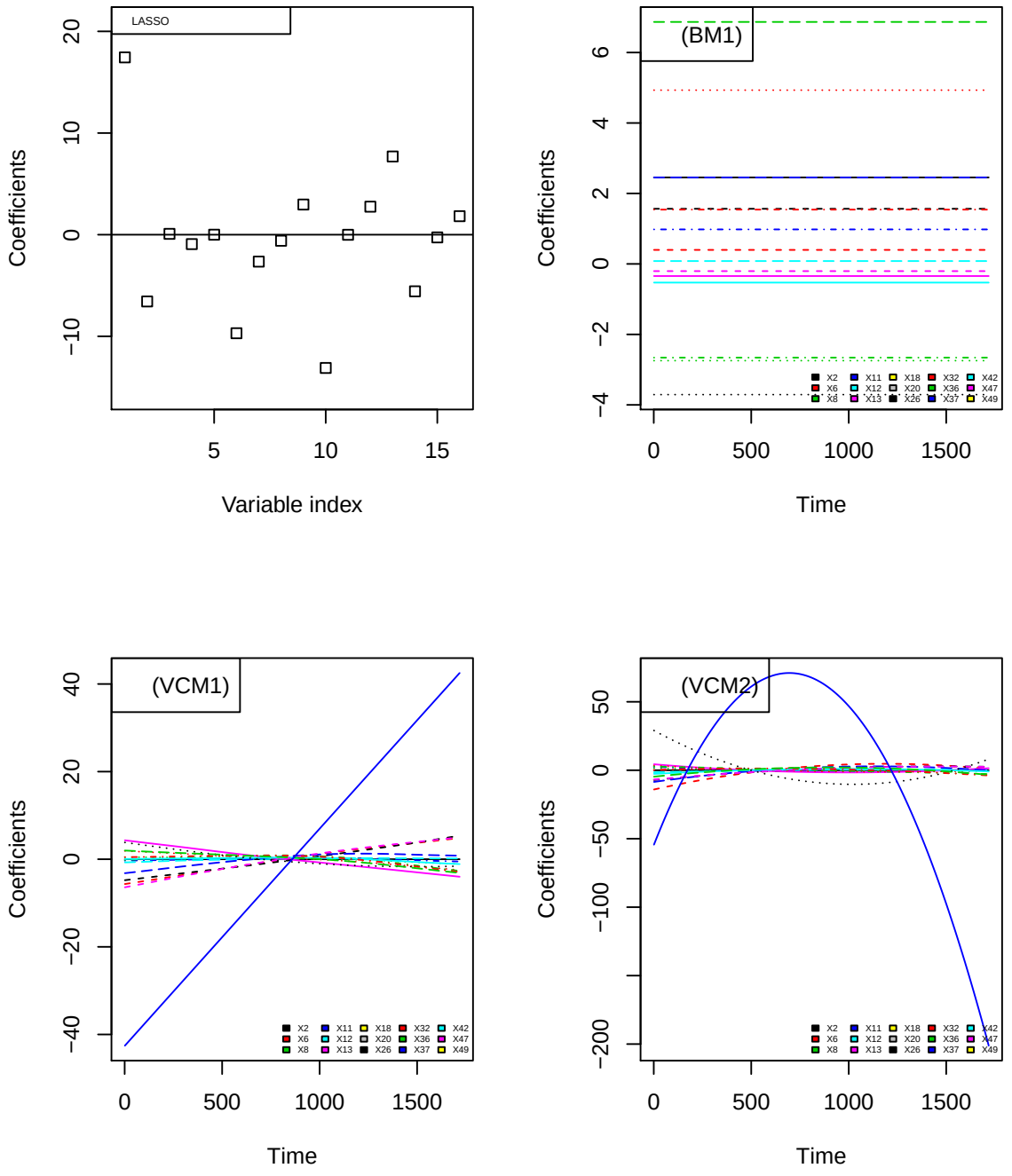

Figure 3.3: The estimated coefficients from the LASSO, the regularized dynamic logistic regression model with fused penalty and the group lasso penalty (rDLR), the regularized dynamic logistic regression model with fused penalty (BM1), the regularized dynamic logistic regression model with fused penalty and the l_1-norm penalty (BM2), the varying coefficient model with the smoothing spline basis (VCM1), and the varying coefficient model with the polynomial basis (VCM2) for the crystal ingot growth experiments.

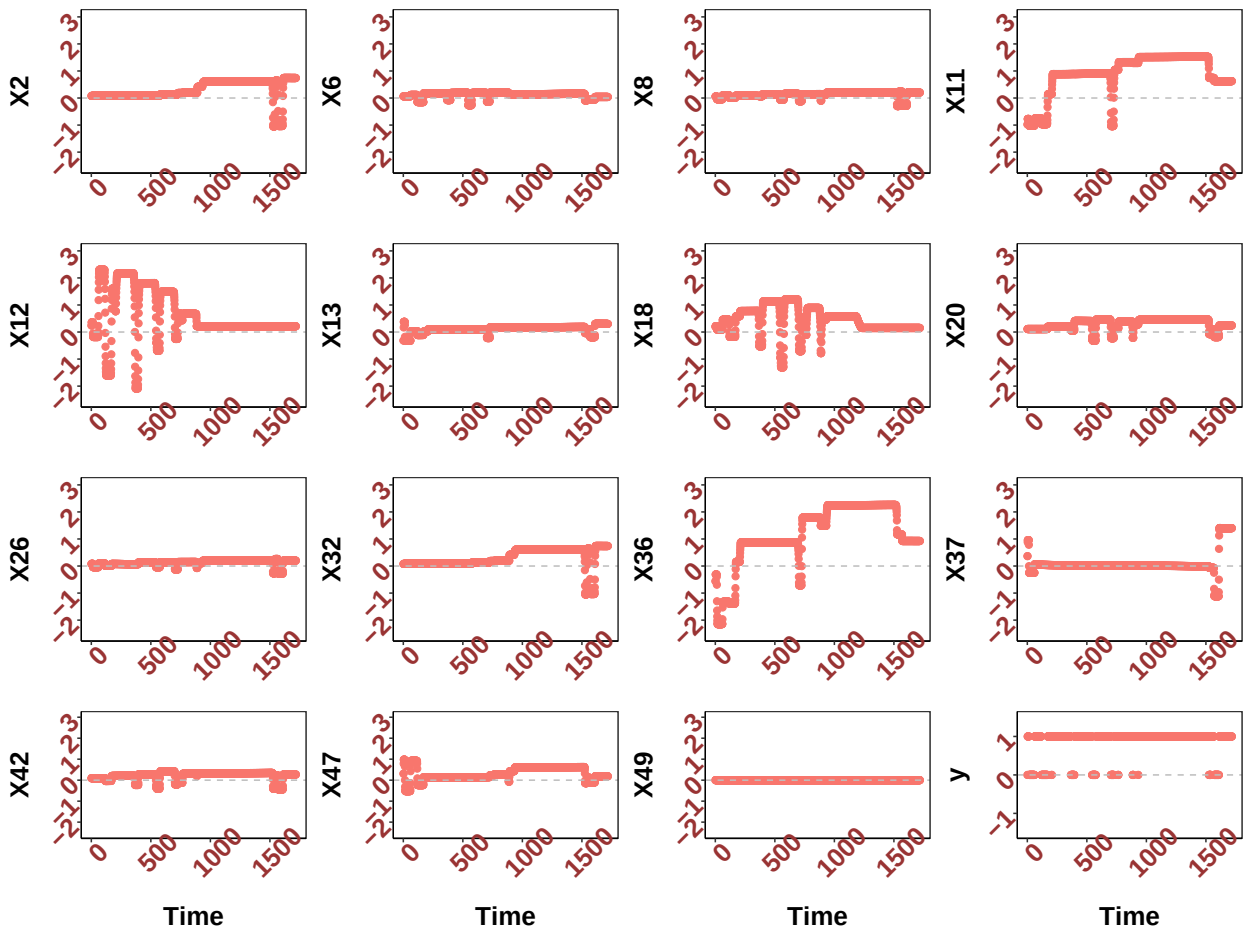

Figure 3.4: The estimated coefficients from the regularized dynamic logistic regression model with fused penalty and the group lasso penalty (rDLR) and the response for the crystal ingot growth experiments.

the rDLR. At different time segment regions, different feature variables play the main role in terms of the magnitudes of estimated coefficients. At the beginning of the experiment, the variable X12 and X18 have large coefficient estimators, but their effects are reduced to zero at the end of the experiment. On the contrary, the variable X2 has zero effects in the beginning but large effects in the end. Besides, during the whole running experiment, the variables X11 and X36 have effects, but the variable X49 has no effects.

We also study the prediction performance of compared models and report the deviance

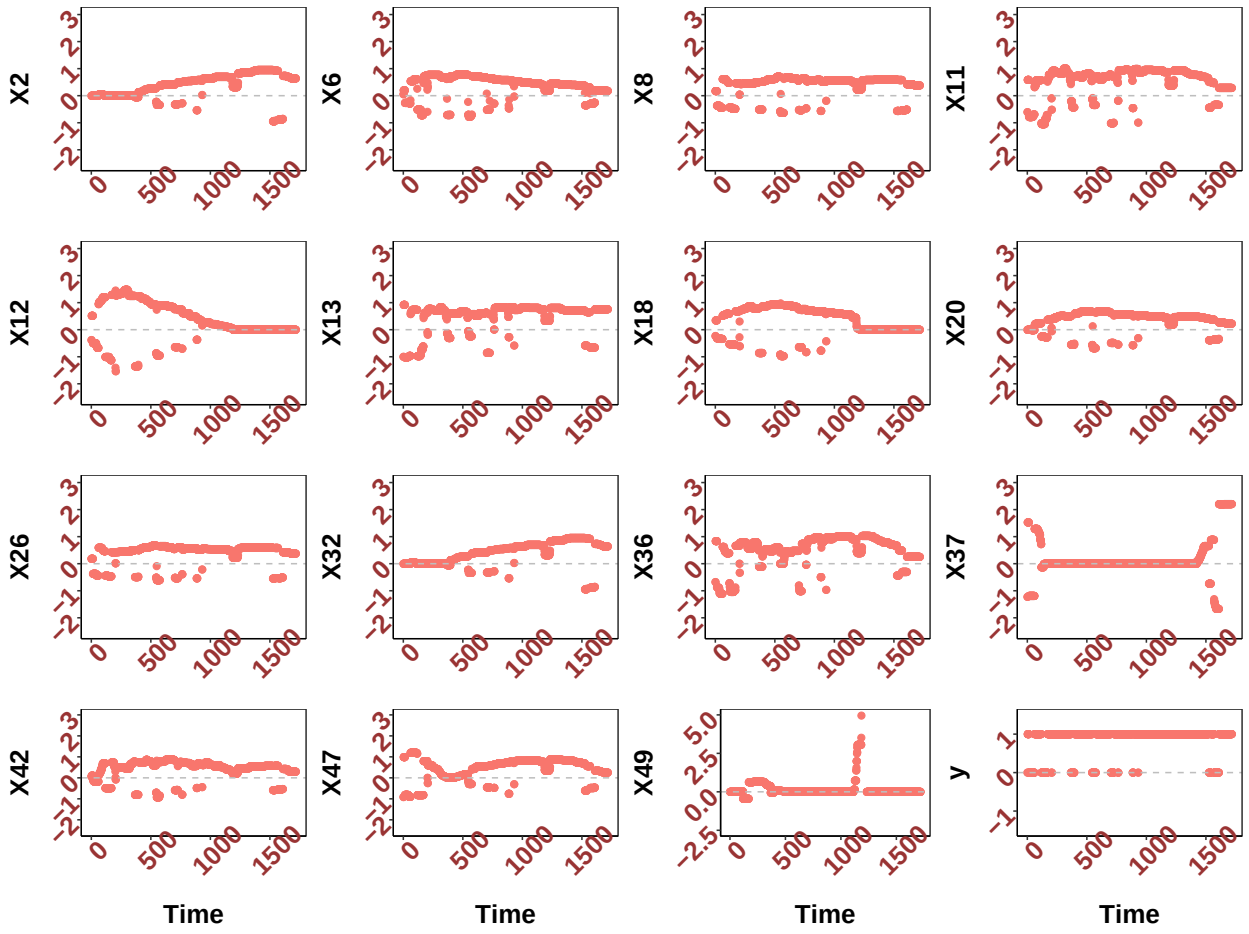

Figure 3.5: The estimated coefficients from the regularized dynamic logistic regression model with fused penalty and the l_1-norm penalty (BM2) and the response for the crystal ingot growth experiments.

Table 3.3: Prediction performance of models in the crystal ingot growth experiment.

	DEV	sd	MER	sd	Size
rDLR	0.132	0.011	0.016	0.003	14
BM1	0.675	0.014	0.133	0.005	15
BM2	**0.080**	0.007	**0.013**	0.004	15
VCM1	0.599	0.011	0.108	0.003	15
VCM2	0.543	0.009	0.096	0.006	15
LASSO	0.651	0.035	0.133	0.011	16

(DEV) and misclassification error rate (MER) from the 5-fold CV and and model size (Size) in Table 3.3. Here, for the models rDLR, VCM1, VCM2, BM1, and BM2, Size is defined as the number of predictor variables which have at least one nonzero estimated coefficient at any time point. The BM2 has the smallest DEV and MER values, however, the rDLR has the second smallest DEV and MER values.

3.5.2 Hong Kong Environmental Study

The main objective of the Hong Kong environmental study is to learn the relationship between the levels of pollutants and the number of daily hospital admission for circulation and respiration. Cai et al. (2000) reported that the relationship between the number of hospital admissions and the pollutants varies with time. In this work, we are interested in studying the impact on the hospital admission from the levels of pollutants over time. The response is a binary indicator, defined as one when the number of hospital admission is greater than the median number of daily hospital admission during each whole calendar year and as zero otherwise. The predictors are levels of pollutants including the sulfur dioxide (SO_2) and nitrogen dioxide (NO_2) in Hong Kong between January 1, 1994 and December 31, 1995. Both

predictors SO_2 and NO_2 are positive values. Moreover, 12 additional noise variables, among which six variables follow normal distribution $N(0, 1)$ and the other six variables follow the AR(1) model, are simulated and included in the analysis. The total number of observations is 730.

Figure 3.6 shows the estimated coefficients from the compared models including LASSO, rDLR, VCM1, VCM2, BM1, and BM2. The rDLR selects the expected significant variables NO_2 and SO_2. Besides, both variables have time-varying effects: more than five sudden changes in each coefficient curve. The effects of variables NO_2 and SO_2 are consistent in terms of their signs, suggesting that NO_2 and SO_2 have similar effects on the response. The LASSO identifies the significant variable as well, but the estimated coefficients of NO_2 and SO_2 have opposite signs, which is contrary to the knowledge that both SO_2 and NO_2 are hazards to the respiratory health. A closer look at the estimated functional coefficient curves (Cai et al. 2000) shows that the coefficient-signs of NO_2 and SO_2 from their model are opposite at some time points. The BM2 selects all the variables and its estimated coefficients have opposite signs. The BM2 does not identify the variable SO_2. The VCM1 and VCM2 methods identify the significant variables.

Figure 3.7 shows the estimated coefficient for each feature variable over time obtained from the rDLR. The rDLR identifies the significant variables SO_2 and NO_2 with large coefficient values, and noise variables having relatively small coefficients compared to the variables SO_2 and NO_2. The estimated coefficients for the variables SO_2 and NO_2 have the same sign at different time regions.

Table 3.4 compares the prediction performance of compared models. The rDLR has the smallest DEV and MER. The rDLR has the same model size as BM1, BM2, VCM1, and VCM2.

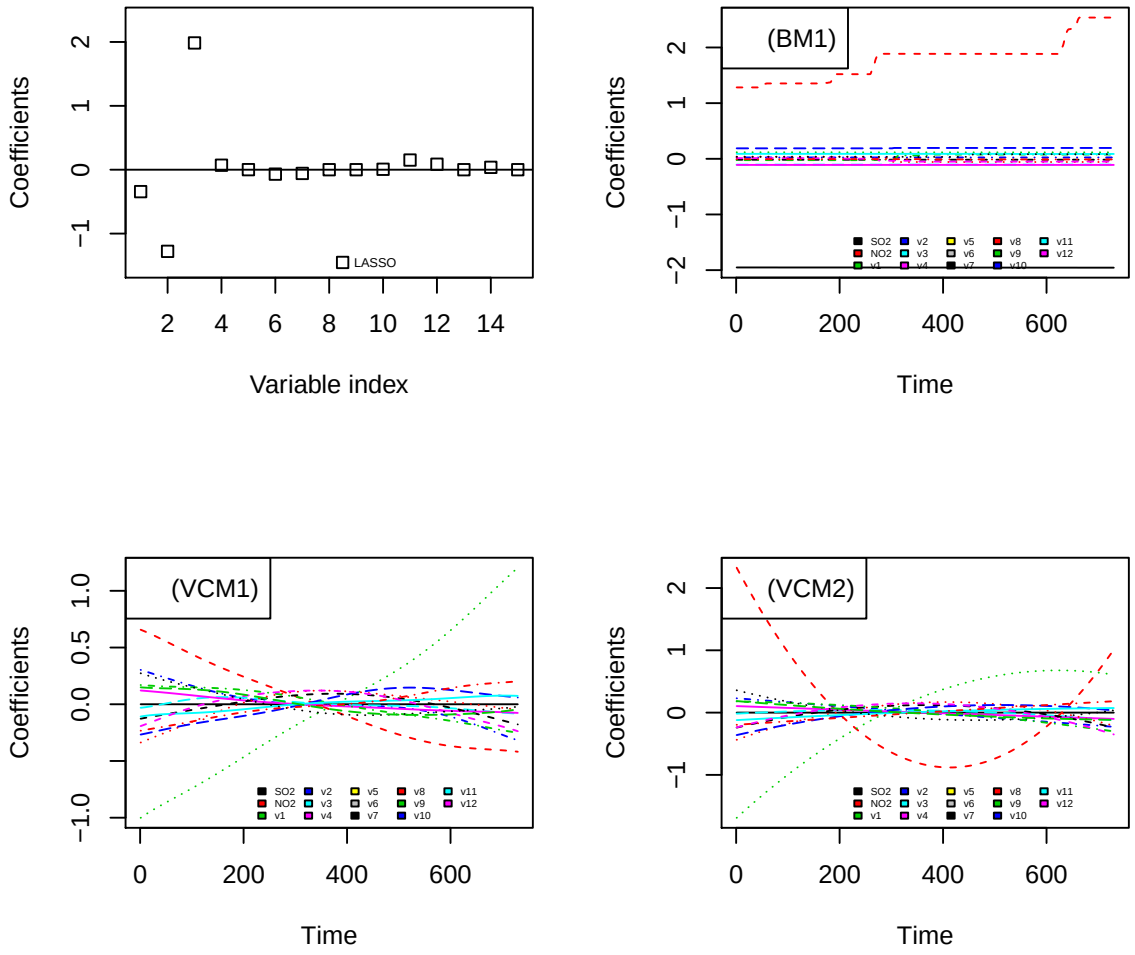

Figure 3.6: The estimated coefficients from LASSO, the regularized dynamic logistic regression model with fused penalty and the group lasso penalty (rDLR), the regularized dynamic logistic regression model with fused penalty (BM1), the regularized dynamic logistic regression model with fused penalty and the l_1-norm penalty (BM2) for the Hong Kong environmental study.

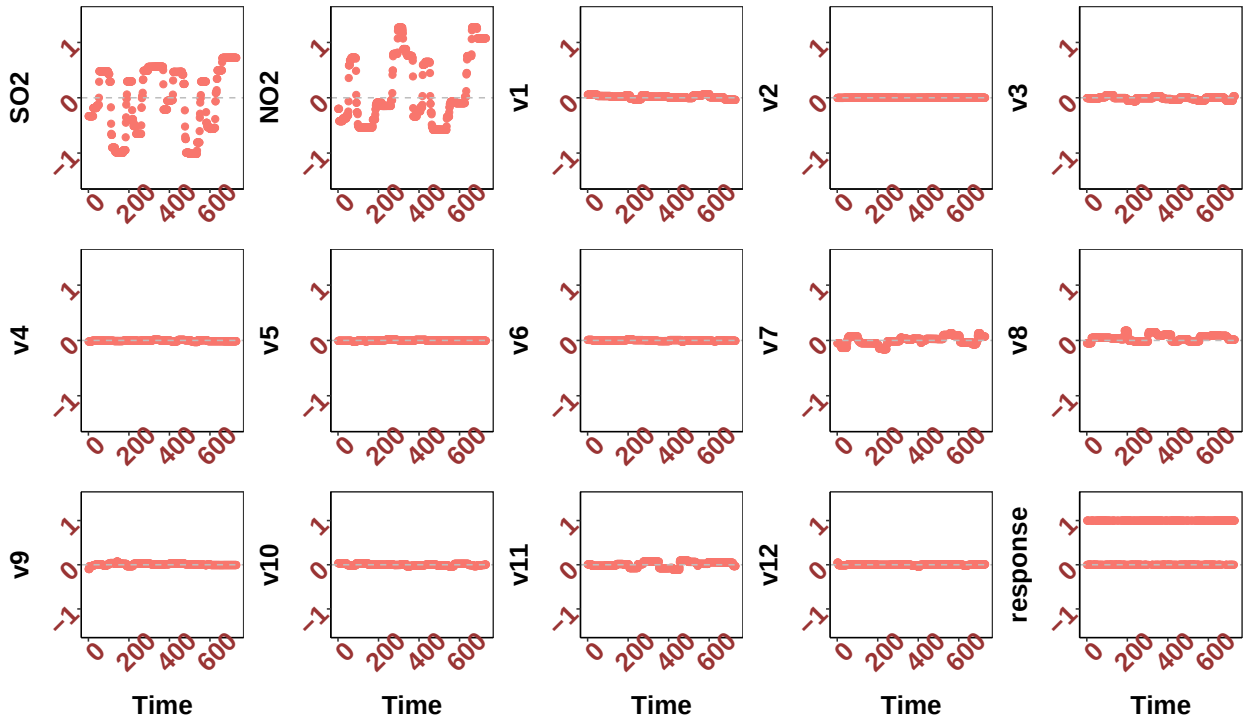

Figure 3.7: The estimated coefficients from the regularized dynamic logistic regression model with fused penalty and the group lasso penalty (rDLR) and the response for the Hong Kong environmental study.

Table 3.4: Performance comparison of models in the Hong Kong environmental study.

	DEV	sd	MER	sd	Size
rDLR	**1.237**	0.042	**0.319**	0.043	14
BM1	1.345	0.048	0.385	0.027	14
BM2	1.242	0.045	**0.312**	0.029	14
VCM1	1.394	0.029	0.464	0.026	14
VCM2	1.401	0.047	0.462	0.028	14
LASSO	1.340	0.008	0.414	0.022	11

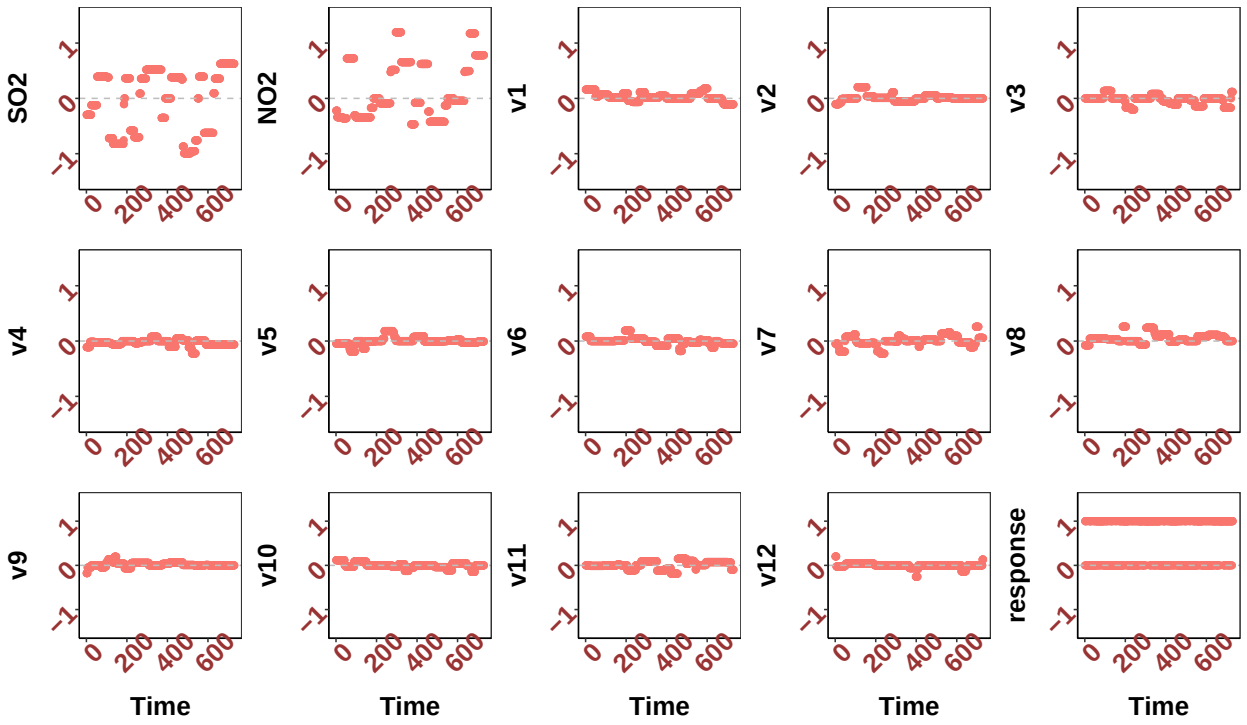

Figure 3.8: The estimated coefficients from the regularized dynamic logistic regression model with fused penalty and the l_1-norm penalty (BM2) and the response for the Hong Kong environmental study.

In addition, we use an expanding window to study the effect of number of observations on the model performance. The expanding window starts from the first observation and covers a pre-defined percent of the total observation number. The window expands by sliding forward to include additional observations in order to form a larger data sample. We control the percent as the sequence starting from 30% and ending at 100% with an interval of 10%. Figure 3.9 shows that the rDLR has the smallest DEV at all percent of total observations.

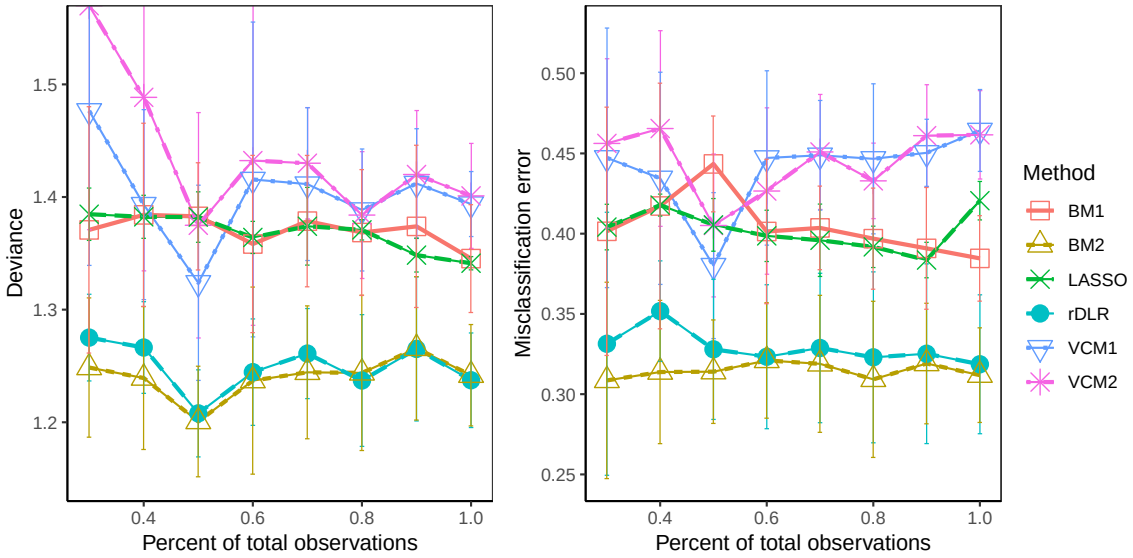

Figure 3.9: The deviance (DEV) and the misclassification error rate (MER) among compared models at various proportions of the whole data set in the Hong Kong environmental study.

3.5.3 Photodegradation Experiments

Photodegradation caused by ultraviolet (UV) radiation is the primary cause of failure for paints and coatings. Photodegradation data were collected in a multiyear research problem via scientifically-based laboratory accelerated tests by scientists at the U.S. National Institute of Standards and Technology (NIST, Gu et al. 2009). The response is the damage amounts to the material measured by Fourier transform infrared spectroscopy. For more details see Hong et al. (2015) and Duan et al. (2017). In this study, the goal is to study the impact on the damage amounts from the levels of environmental variables over time. We choose the representative "G18-8" unit as the subject. The response is transformed into a binary response defined as one when the damage amount is greater than 0.2 and as zero otherwise. The predictors are the environmental variables including the temperature (TEMP), relative humidity (RH), and UV dosage (UV). All the three environmental variables are positive except one temperature observation. Moreover, six additional variables following normal distribution $N(0, 1)$ are simulated and used as noise variables in the analysis. The total number of observations is 39.

Figure 3.10 shows the estimated coefficients from the compared models including LASSO, rDLR, BM1, and BM2. The rDLR selects the significant variables TEMP, RH, and UV. It is clear that these three variables have distinct effects on different regimes. When the damage response is zero, then the coefficients have negative effects on the degradation. However, when the response changes from zero to one, the coefficients have positive effects on the coating materials. The LASSO identifies the variables TEMP and UV as the significant variables. But the estimated coefficients are negative, suggesting the higher, for example, TEMP, the more likely the materials are not degrading.

The BM1 selects not only the three variables as significant variables, but some noise variables

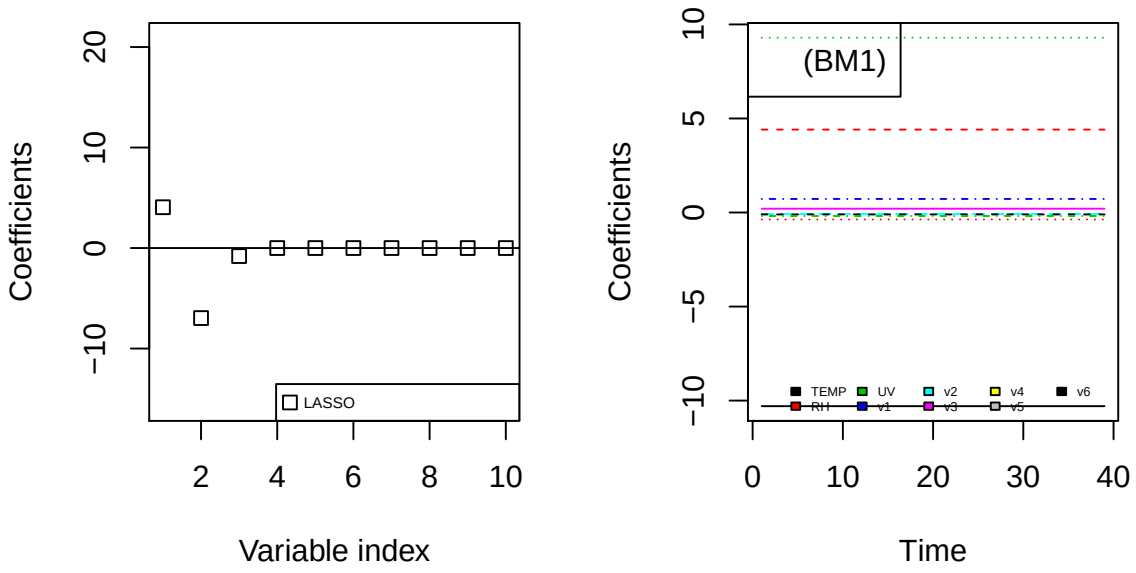

Figure 3.10: The estimated coefficients from the LASSO, the regularized dynamic logistic regression model with fused penalty and the group lasso penalty (rDLR), the regularized dynamic logistic regression model with fused penalty (BM1), the regularized dynamic logistic regression model with fused penalty and the l_1-norm penalty (BM2) for the analysis of the photodegradation data set.

with small coefficients. The estimated coefficients from BM1 do not have obvious jump change-points. This can be explained by the small size of data set, which cannot support and allow as many as nine predictor variables have dynamic coefficients. Therefore, the l_1-norm fuse penalty encourages constant coefficients. Same as LASSO, the signs of estimated TEMP and UV variables are opposite. The BM2 selects no variables before time point 19, but more than five variables after time point 19. The VCM1 and VCM2 method does not work in this case with complete separation in response, which is due to the numerical issues in their iterative reweighted least squares methods.

Figure **??** shows the estimated coefficient for each feature variable over time obtained from the rDLR. The rDLR identifies clearly the effects of significant variables TEMP, RH, and UV at different time segments with large coefficient values. In addition, the noise variables are correctly identified and excluded.

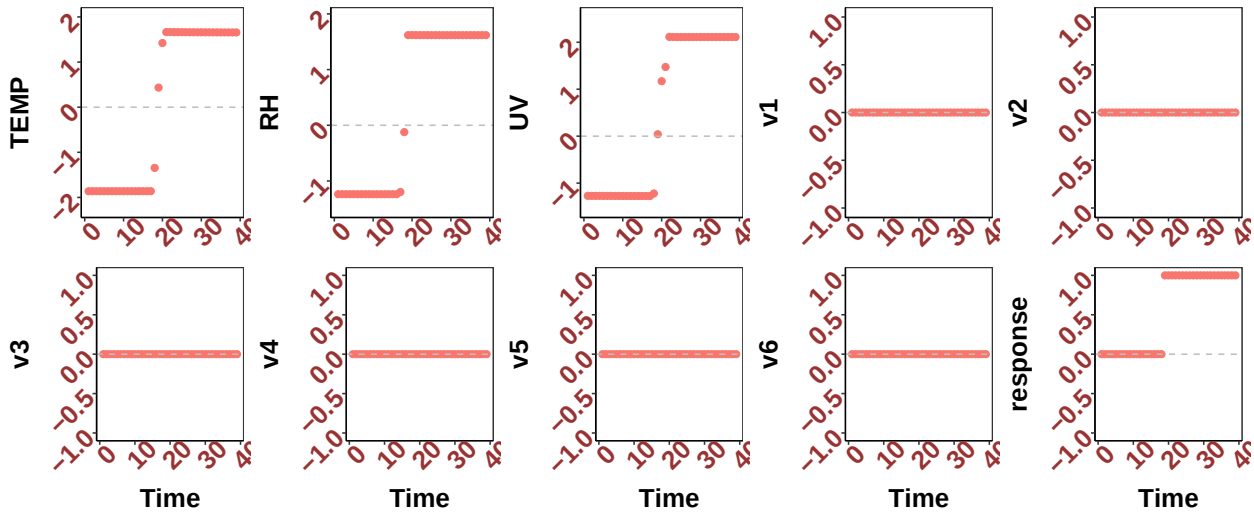

Figure 3.11: The estimated coefficients from the regularized dynamic logistic regression model with fused penalty and the group lasso penalty (rDLR) and the response for the photodegradation data set.

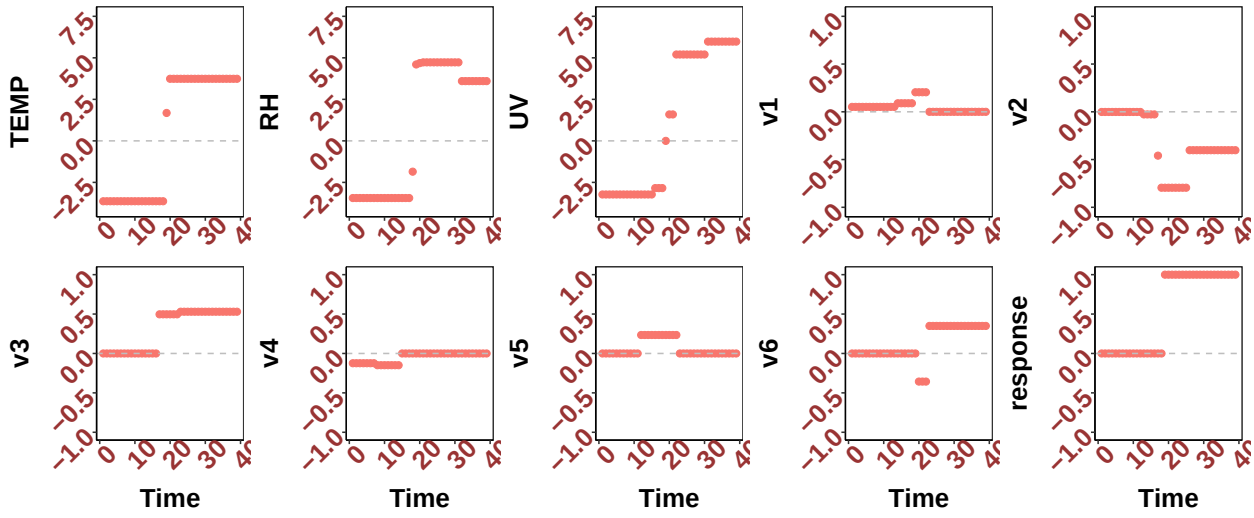

Figure 3.12: The estimated coefficients from the regularized dynamic logistic regression model with fused penalty and the l_1-norm penalty (BM2) and the response for the photodegradation data set.

Table 3.5: Performance comparison of models in the photodegradation experiment.

	DEV	sd	MER	sd	Size
rDLR	0.311	0.060	**0.000**	0.000	3
BM1	2.239	1.632	0.493	0.172	9
BM2	**0.080**	0.012	**0.000**	0.000	9
LASSO	0.931	0.167	0.205	0.066	3

Table 3.5 compares the prediction performance of compared models. The rDLR has the smallest DEV and MER.

3.6 Discussion

The quality control in the crystal ingot growth experiments is important considering the high cost of energy, long-running experimental time, and great values of ingot. However, the dynamic effects of process variables on the quality response due to the growth of the crystal ingot and the degradation of equipment conditions is not fully understood. In this work, we propose a regularized dynamic logistic regression model to learn the dynamic effects of process variables. The dynamic effects of process variables are approximated by piecewise constant functions. The selection of significant variables is achieved by the l_2-norm group lasso regularization technique. The performance of the proposed model is demonstrated in numerical studies and real-world problems.

The proposed regularized dynamic logistic regression model can be extended to a model with further regularization. For example, one can consider the l_1-norm lasso penalty on each individual coefficient β_{ij} such that sparsity is enforced within the estimated coefficients. This idea is similar to the sparse group lasso work (Friedman et al. 2010; Simon et al. 2013).

It is also interesting to consider another l_2-norm group lasso penalty on groups of variables such that the variables in one group would be selected together or not. Furthermore, Land and Friedman (1997) developed different variable fusion methods: the zero-order and the first-order fused penalty. It is interesting to investigate the effects of different degrees of variable fusion methods.

The proposed model assumes that the response at time t is only affected by current predictor variables at time t. However, the current response may also depends on the past predictor variables. It is interesting to study the impact of the past predictor variables on the current response. It is worth remarking that in this study we do not include any degradation variables in the analysis of the crystal growth experiments. One direction for future research is to consider modeling the relationships between the product quality, degradation variables, and the process variables together. Jin et al. (2019) proposed to model the coefficients of process variables as a nonlinear function of a degradation variable in order to directly reflect the effect of degradation variables on the process variables. However, their model is built on the assumption that the dynamic coefficients are closely related to the degradation variable. An interesting but challenging analysis is to model these dynamic and complex relationships in the framework of graphical models.

Chapter 4 Structured Variable Selection from an Experimental Thinking Perspective

4.1 Introduction

Variable selection is an important technique to extract useful knowledge from data with complex structures. The underlying sparsity assumption (only a few variables with non-zero model coefficients) yields meaningful models with computationally stable estimates (Tibshirani 1996). In addition to the sparsity assumption, the information on the structures of predictor variables, such as grouping (Yuan and Lin 2006) and network structures (Friedman et al. 2008), is useful for improving model interpretation and model prediction. Overlooking available structures in the modeling can result in undesired feature selection or difficulty in model interpretation.

In the literature, many structured variable selection methods have been studied especially in the context of the LASSO framework (Tibshirani 1996; Chen 1998). The commonly-explored structures include the group lasso (Bakin 1999; Yuan and Lin 2006; Meier et al., 2008), the overlapping group lasso (Jacob et al. 2009), the fused lasso (Tibshirani et al. 2005), the sparse group lasso (Simon et al. 2013), the hierarchical selection (Zhao et al. 2006), and the tree-guided group lasso (Liu and Ye, 2010; Kim and Xing 2010), among many others. Note that the optimization objective function in the Lasso is convex, which leads to many efficient algorithms such as the coordinate descent algorithm (Friedman et al. 2007; Nesterov, Y. 2012) and the alternating direction method of multipliers algorithm (Boyd et al. 2011).

On the other hand, the estimated coefficients from the LASSO can be biased because of

the well-known shrinkage effect (Zou and Hastie 2005). Moreover, it can be difficult to fully control the number of selected variables in these LASSO-related methods as the tuning parameter is to control the L_1 norm of the coefficient vector. This is attributed to the unclear relationship between the continuous regularization parameter and the number of selected variable.

To address the aforementioned concerns of LASSO-related methods, one approach is to consider the best subset selection based methods. The best subset selection problem is essentially on how to choose the best subset out of all available predictors. The corresponding optimization is non-convex and non-continuous, and known to be NP-hard (Natarajan, 1995). Furnival and Wilson (1974) proposed a branch-and-bound algorithm to solve the best subset selection problem with the dimension that is not beyond 30s. However, for problems with high-dimension data, the computation in best subset problems can be very expensive. Recently, new developments in the algorithms to solve the best subset selection problem with p in the 1000s are reported. Hazimeh and Mazumder (2019) develop algorithms based on coordinate descent and local combinatorial optimization schemes to perform fast best subset selection. Bertsimas et al. (2016) transformed the best subset problem into a mixed integer quadratic programming (MIQP) problem and solved the problem with the highly optimized Gurobi mixed integer optimization (LIO) solver. Their approach can find near-optimal solutions for n in the 100s and p in the 1000s in minutes, however, the verification of the solution optimality can take much longer time.

However, there are few works on the structured best subset selection problem in the literature. The structured best subset selection problem is how to choose the best subset out of features provided the structure knowledge among features. In this work, we propose a new method within the best subset selection framework to perform the structured variable selection. The proposed method is inspired by experimental design thinking, and borrows strengths between statistics and optimization. The key idea of the proposed sparse ridge regression method is using the experimental thinking to transform the structured best subset selection into a missing data problem with an orthogonal regression matrix. Specifically, the proposed method utilizes the expectation-maximization (EM) algorithm to formulate the best subset selection problem as an iterative linear integer optimization (LIO) problem. The LIO formulation can be applied to various structured variable selection problems. To the best of our knowledge, this is a first work that large-scale structured variable selection problems to be solved within the best subset selection problem framework. The contribution

of the proposed method unfolds in three aspects. First, the proposed method illustrates the key idea of the best subset selection: select the optimal subset and estimate their coefficients. Second, the proposed method is flexible to accommodate various structure knowledge in the modeling. The number of selected variables can be precisely controlled. Third, the proposed method is computationally efficient.

The remainder of this chapter is organized as follows. In Section 4.2, we present the proposed sparse ridge regression problem in the view of experimental designs. In Section 4.3, we detail the expectation-maximization (EM) algorithm for estimation. In Section 4.4, we discuss the application of the LIO formulation to various structured variable selection problems. In Section 4.5, we present the estimation algorithm. In Section 4.6, we present the real-data analysis. We conclude this work with some discussion in Section 4.7.

4.2 Sparse Ridge Regression

Provided the continuous response $\boldsymbol{y} \in R^p$ and the predictor matrix $X = (x_1, \cdots, x_p) \in R^{n \times p}$, the least squares regression is given as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta} \in R^p$ is the coefficient and the error term $\boldsymbol{\epsilon}$ follows a Gaussian distribution $N(\mathbf{0}, \sigma^2 I_p)$. The associated negative log-likelihood function is

$$\begin{aligned} -l(\boldsymbol{\beta}; X, \boldsymbol{y}) &= n\log(\sqrt{2\pi}\sigma) + \frac{(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta})}{2\sigma^2}, \\ &\propto (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}). \end{aligned}$$

We consider the sparse ridge regression with L_0-norm, where the negative log-likelihood function with the L_2-norm penalty term of $\boldsymbol{\beta}$ is

$$-l(\boldsymbol{\beta}; X, \boldsymbol{y}) \propto (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \gamma||\boldsymbol{\beta}||_2^2, \text{ subject to } ||\boldsymbol{\beta}||_0 \leq k, \tag{4.1}$$

Then the sparse ridge regression problem with L_0-norm is defined as the following optimization problem

$$\underset{\beta \in R^p}{\text{minimize}} \; ||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2 + \gamma||\boldsymbol{\beta}||_2^2 \text{ subject to } ||\boldsymbol{\beta}||_0 \leq k, \tag{4.2}$$

where $||\boldsymbol{\beta}||_0 = \sum_{i=1}^{p} 1\{\boldsymbol{\beta}_i \neq 0\}$ is the L_0-norm of the coefficient $\boldsymbol{\beta}$ with $1(\cdot)$ denoting the indicator function. The L_0-norm term, different from the L_1-norm penalty term in LASSO, does not depend on the magnitudes of coefficients. The turning parameter γ adjusts the effect of L_2-norm penalty, $||\boldsymbol{\beta}||_2^2$. The discrete tuning parameter k is the number of chosen predictors ranged between 0 and min{n, p}.

The sparse ridge regression can be viewed from the perspective of experimental designs. To illustrate the idea, we consider the trivial case when X is column-orthogonal. Because $X^T X$ is a diagonal matrix, the problem 4.2 becomes

$$\underset{\beta \in R^p}{\text{minimize}} \sum_{j=1}^{p} (d_j + \gamma)\beta_j^2 - 2\sum_{j=1}^{p} \beta_j r_j + \sum_{i=1}^{n} y_i^2, \text{subject to } ||\boldsymbol{\beta}||_0 \leq k, \tag{4.3}$$

where $r_j = X_j^T \boldsymbol{y}$ and d_j is the j-th diagonal element in the diagonal matrix $X^T X$.

Considering that the objective function term is decomposable in the dimension of $\boldsymbol{\beta}$, solving the problem 4.3 is not difficult. We describe one solution to the decomposable situation in Section 4.3.

In practice, X is generally not column-orthogonal. We propose to design the orthogonalization scenario by augmenting the data set. Specifically, we adopt the idea of active orthogonalization to orthogonalize an arbitrary data matrix by adding more rows, $X^\dagger \in R^{(m-n)\times p}$, resulting in the complete orthogonalized regression matrix $X_c = (X^T \ X^{\dagger,T})^T \in R^{m\times p}$. In other words, we have

$$X^T X + X^{\dagger,T} X^\dagger = D,$$

where the matrix $D \in R^{p\times p}$ is diagonal. The size of data is augmented from n to m. We treat the response $\boldsymbol{y}^*$ corresponding to the added rows as missing values. Figure 4.1 illustrates the augmented $X_c = (X^T \ X^{\dagger,T})^T \in R^{m\times p}$ and $\boldsymbol{y}_c = (\boldsymbol{y}^T, \ \boldsymbol{y}^{*,T})^T \in R^m$ after the data augmentation procedure.

With the augmented complete response $\boldsymbol{y}_c$ and the augmented complete orthogonalized regression matrix X_c, the sparse ridge regression is reformulated as:

$$\underset{\boldsymbol{\beta}\in R^p, \ \boldsymbol{y}^* \in R^{m-n}}{\text{minimize}} \ ||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2 + ||\boldsymbol{y}^* - X^\dagger \boldsymbol{\beta}||_2^2 + \gamma||\boldsymbol{\beta}||_2^2 \text{ subject to } ||\boldsymbol{\beta}||_0 \leq k, \tag{4.4}$$

Note that when we set $\gamma = 0$, the sparse ridge regression is reduced to the classical best subset

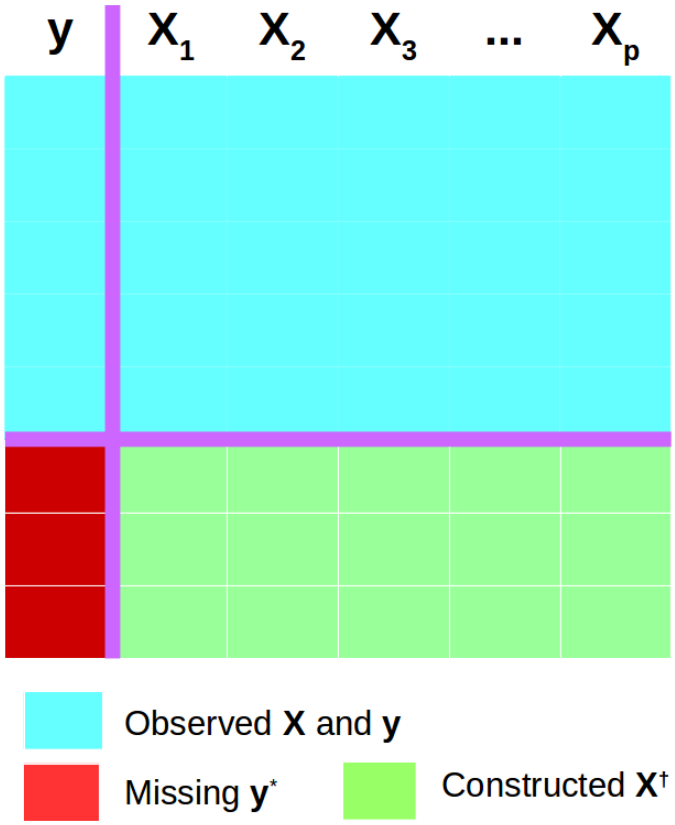

Figure 4.1: Illustration of the design for data augmentation. The cyan-colored cells represent the observed data set X with p columns and $\boldsymbol{y}$, the green-colored cells represent the augmented $X^{\dagger}$, and the red-colored cells represent the missing $\boldsymbol{y}^{*}$ corresponding to the $X^{\dagger}$. The resulting complete predictor matrix $X_c = (X^T \ X^{\dagger,T})^T$ is column-orthogonal. The resulting complete response vector $\boldsymbol{y}_c = (\boldsymbol{y}^T, \ \boldsymbol{y}^{*,T})^T$.

selection problem. Our proposed framework still works for the classical best subset selection problem, because the addition of the L_2 norm penalty in the sparse ridge regression only affects the weights of predictors in the coefficient estimator. However, the ridge penalized structured best subset model has advantages over the best subset selection model in terms of model prediction. The extra L_2-norm penalty term in the sparse ridge regression introduces the continuous and tunable shrinkage, which biased the estimated coefficients but yields a continuous fitted response, generally leading to an improvement in the model prediction. While in the best subset selection problem, the estimated coefficients are unbiased, but have a large variance due to the bias-variance tradeoff. Furthermore, the fitted response in the best subset selection model can have sudden changes in values due to the inclusion or exclusion of variables. Last, the classical best subset selection method may suffer from the singularity issue of the subset of regression matrix, though the expectation-maximization (EM) algorithm (described later in Section 4.3) is proved to converge to the Moore-Penrose generalized inverse-based least squares estimator (Xiong et al. 2016).

4.3 Expectation-Maximization Algorithm for Model Estimation

We propose to use the EM algorithm to solve the reformulated sparse ridge regression problem C.2. In the EM algorithm (Dempster et al. 1977; Horton and Laird 1998), the **E-step** is to compute the expectation of the complete log-likelihood function through imputing the missing values by their conditional expectation, and the **M-step** is to estimate parameters by maximizing the expected complete log-likelihood function.

Given the augmented complete response $\boldsymbol{y}_c = (\boldsymbol{y}^T,\ \boldsymbol{y}^{*,T})^T \in R^m$ and the augmented complete orthogonalized regression matrix $X_c = (X^T\ X^{\dagger,T})^T \in R^{m\times p}$, the complete log-likelihood function, similar to 4.1, can be written as

$$l_c(\boldsymbol{\beta}; X_c, \boldsymbol{y}_c) = -m\log(\sqrt{2\pi}\sigma) - \frac{\boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{y}^{*,T}\boldsymbol{y}^* - 2(\boldsymbol{y}^T X + \boldsymbol{y}^{*,T}X^\dagger)\boldsymbol{\beta} + \boldsymbol{\beta}^T X_c^T X_c \boldsymbol{\beta} + \gamma\boldsymbol{\beta}^T\boldsymbol{\beta}}{2\sigma^2},$$

which is subject to $||\boldsymbol{\beta}||_0 \leq k$.

In the **E-step**, we replace the missing values in response, $\boldsymbol{y}^*$, with the conditional expectation

given the observed data, $\mathrm{E}(\boldsymbol{y}^*|X^\dagger;\ \boldsymbol{\beta})$. We have

$$Q(\boldsymbol{\beta}; X_c, \boldsymbol{y}_c, \boldsymbol{\beta}^{(t)}) = E[l_c(\boldsymbol{\beta}; X_c, \boldsymbol{y}_c, \boldsymbol{\beta}^{(t)})] = -m\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\Big[\boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^{T,(t)}X^{\dagger,T}X^\dagger\boldsymbol{\beta}^{(t)} + \sigma^2 - 2(\boldsymbol{y}^TX + \boldsymbol{\beta}^{T,(t)}X^{\dagger,T}X^\dagger)\boldsymbol{\beta} + \boldsymbol{\beta}^TD\boldsymbol{\beta} + \gamma\boldsymbol{\beta}^T\boldsymbol{\beta}\Big],$$

where $\boldsymbol{\beta}^{(t)}$ is computed from the last iteration and the conditional expectations are given by

$$\begin{aligned} E(\boldsymbol{y}^*|X^\dagger, \boldsymbol{\beta}) =& X^\dagger\boldsymbol{\beta}, \\ E(\boldsymbol{y}^{*,T}\boldsymbol{y}^*|X^\dagger, \boldsymbol{\beta}) =& \boldsymbol{\beta}^TX^{\dagger,T}X^\dagger\boldsymbol{\beta} + \sigma^2. \end{aligned}$$

In the following **M-step**, we maximize $E[l_c(\boldsymbol{\beta}; X_c, \boldsymbol{y}_c, \boldsymbol{\beta}^{(t)})]$ over $\boldsymbol{\beta}$. That is

$$\underset{\boldsymbol{\beta}}{\text{maximize}} - \Big[\boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^{T,(t)}X^{\dagger,T}X^\dagger\boldsymbol{\beta}^{(t)} + \sigma^2 - 2(\boldsymbol{y}^TX + \boldsymbol{\beta}^{T,(t)}X^{\dagger,T}X^\dagger)\boldsymbol{\beta} + \boldsymbol{\beta}^TD\boldsymbol{\beta} + \gamma\boldsymbol{\beta}^T\boldsymbol{\beta}\Big],$$
$$\text{subject to } ||\boldsymbol{\beta}||_0 \leq k,$$

which can be further reduced to

$$\underset{\boldsymbol{\beta}}{\text{maximize}} - \boldsymbol{\beta}^T(D + \gamma I_p)\boldsymbol{\beta} + 2(X^T\boldsymbol{y} + X^{\dagger,T}X^\dagger\boldsymbol{\beta}^{(t)})^T\boldsymbol{\beta}, \text{ subject to } ||\boldsymbol{\beta}||_0 \leq k, \quad (4.5)$$

$$\Leftrightarrow \underset{\boldsymbol{\beta}}{\text{maximize}} \sum_{j=1}^{p}\big[-(d_j + \gamma)\beta_j^2 + 2\mu_j^{(t)}\beta_j\big], \text{ subject to } ||\boldsymbol{\beta}||_0 \leq k,$$

$$\Leftrightarrow \underset{\boldsymbol{\beta}}{\text{maximize}} \sum_{j=1}^{p}\big[-(d_j + \gamma)(\beta_j - \frac{\mu_j^{(t)}}{d_j + \gamma})^2 + \frac{\mu_j^{2,(t)}}{d_j + \gamma}\big] \text{ subject to } ||\boldsymbol{\beta}||_0 \leq k, \quad (4.6)$$

where $\mu_j^{(t)}$ is the j-th component in $\boldsymbol{\mu}^{(t)} = X_c^T\boldsymbol{y}_c = X^T\boldsymbol{y} + X^{\dagger,T}X^\dagger\boldsymbol{\beta}^{(t)}$. In the problem C.3, the objective function term is decomposable in the dimension of $\boldsymbol{\beta}$. The maximum objective function value in the dimension j of $\boldsymbol{\beta}$, $\max(OF_j)$, is

$$\max(OF_j) = \begin{cases} \frac{\mu_j^{2,(t)}}{d_j + \gamma} & \text{if } \beta_j \neq 0, \\ 0, & \text{if } \beta_j = 0. \end{cases}$$

We introduce the binary variable z_j, j-th component in $\boldsymbol{z}$, as $1\{\beta_j^{(t+1)} \neq 0\}$, then $||\boldsymbol{\beta}||_0 =$

$\sum_{i=1}^{p} z_i$. In addition, the $\max(OF_j)$ can be viewed as $z_j \frac{\mu_j^{2,(t)}}{d_j+\gamma}$. Then the problem C.3 is equivalent to

$$\begin{aligned} &\underset{\boldsymbol{z}\in R^p}{\text{maximize}} \sum_{j=1}^{p} z_j \frac{\mu_j^{2,(t)}}{d_j+\gamma}, \\ &s.t. \sum_{j=1}^{p} z_j \leq k, z_j \in \{0,1\}, j = 1, ..., p, \end{aligned} \tag{4.7}$$

Note that there are many ways to construct $X^{\dagger}$. Xiong et al. (2016) suggests to compute $X^{\dagger,T}X^{\dagger} = dI_p - X^TX$ directly, where d, the largest eigenvalue of X^TX or XX^T, can be computed by the power method or the Golub-Kahan-Lanczos bidiagonalization (GKLB) procedure (Golub and Kahan 1965). With Xiong et al.'s suggestion, we have the problem C.4 updated to

$$\begin{aligned} &\underset{\boldsymbol{z}\in R^p}{\text{maximize}} \sum_{j=1}^{p} z_j \frac{\mu_j^{2,(t)}}{d+\gamma}, \\ &s.t. \sum_{j=1}^{p} z_j \leq k, z_j \in \{0,1\}, j = 1, ..., p, \end{aligned} \tag{4.8}$$

where $\mu_j^{(t)}$ is the j-th component in $\boldsymbol{\mu}^{(t)} = X^T\boldsymbol{y} + (dI_p - X^TX)\boldsymbol{\beta}^{(t)}$.

The reformulated problem 4.8 is a linear integer optimization (LIO) problem, which use a binary variable indicating whether the corresponding objective function value is selected or not. Note that the proposed LIO formulation has a linear objective function of integer factors. This is different from Bertsimas et al.'s approach (2016), where their formulation is a mixed integer quadratic optimization (MIQO) problem with a quadratic function of coefficient $\boldsymbol{\beta}$ as the objective function. It is expected that the modern integer optimization solver such as Gurobi runs much faster on LIO problems than on MIQO problems.

The combination of the EM algorithm and the LIO formulation allows efficient computation in the sparse ridge regression problem. During the EM iteration, there is no computation for the matrix inversion. The selection process in the traditional best subset selection problem is challenging due to the 2^p combinatorial choices, but our proposed best subset selection method is fast attributed to the orthogonalization scenario constructed by the EM algorithm. The proposed LIO formulation only needs to solve the binary variable, which greatly increases

the computation efficiency of the MIO solver, Gurobi, which achieved impressive continuous improvement over the last 30 years due to the developments on the relaxation, cutting plane theory, branch and bound theory, and so forth. Besides, our LIO formulation decouples the computation between the binary variable and the continuous coefficients. There is no need for the verification of optimality on the estimated continuous coefficients in the LIO formulation. It is reported that the slow convergence in the estimated continuous coefficients is the major factor increasing the computation burden (Bertsimas et al. 2016, Hastie et al. 2017). At last, the constraints in the LIO formulation can accommodate various structure knowledge into the modeling. In the next section, we illustrate how to incorporate the structure knowledge into the constraints in the LIO formulation to solve various structured variable selection problems.

4.4 Detailed LIO Formulation for Structured Variable Selection

In this section, we illustrate the use of the LIO formulation to incorporate various structure knowledge in various problems. Specifically, We use the constraints in the LIO formulation to represent structure information. In the LIO formulation 4.8, we use the binary variable z_j to indicate the selection of the corresponding objective function value $\frac{\mu_j^{2,(t)}}{d_j+\gamma}$, which further indicates the selection of the variable X_j associated with β_j. Therefore, we can assess the structure knowledge among predictor variables by exploring constraints on the binary variables $\boldsymbol{z}$.

We first demonstrate the effectiveness of the LIO formulation in structured variable selection examples including the hierarchical selection, the group selection, the sparse group selection, the sparse overlapping group selection, and the hierarchical sparse overlapping group selection. Then we discuss the advantages of the LIO formulation for the structured variable selection. Note that in the following notations, the μ_j is a simplified notation $\mu_j^{(t)}$ without the iteration index t.

4.4.1 Hierarchical variable selection

The hierarchical variable selection enforces hierarchy among main effects and their interaction terms based on the heredity principle (Wu and Hamada 2009; Zhao et al. 2009; Bien et al. 2013), which says the selection of the interaction terms is dependent upon the presence of

their parent terms. The strong heredity selects a two-factor interaction only if both its parent terms are selected. The weak heredity selects a two-factor interaction when one of its parent terms is significant. Consider the data matrix with p main effects and their $p(p-1)/2$ pairwise interaction effects, In the following proposed formulation, the constraint 4.9 says z_{kl} will be forced to be zero if z_k or z_l is zero when it is the strong heredity case:

$$\begin{aligned} &\underset{\boldsymbol{z}}{\text{maximize}} \sum_{j=1}^{p} z_j \frac{\mu_j^2}{d+\gamma} + \sum_{k<l} z_{kl} \frac{\mu_{kl}^2}{d+\gamma} \\ &\text{subject to } \sum_{j=1}^{p} z_j + \sum_{k<l} z_{kl} \leq k, z_j \in \{0,1\}, \\ &\qquad \alpha z_{kl} \leq z_k + z_l, \ , k = 1, ..., p, \ l = 2, ..., p, \end{aligned} \tag{4.9}$$

where $\boldsymbol{z}$ is the vector of binary variables containing z_j and z_{kl}, z_{kl} is the binary variable corresponding to the interaction term X_{kl}, the constant α controls the type of the heredity constraint, $\alpha = 1$ for the weak heredity and $\alpha = 2$ for the strong heredity. The similar idea of using the constraint to control the hierarchy has been reported in the model selection in screening experiments (Vazquez-Alcocer et al. 2018).

4.4.2 Group selection

The group selection is common in problems where predictors share underlying factors. For example, one predictor variable is represented by a group of basis functions. Utilizing the group structure may yield more sensible and interpretable models. To solve the group best subset selection problem, we propose to use the following LIO formulation:

$$\begin{aligned} &\underset{\boldsymbol{z},\ \boldsymbol{s}}{\text{maximize}} \sum_{g=1}^{G} \sum_{j=1}^{|g|} z_j^{(g)} \frac{\mu_j^{2,(g)}}{d+\gamma} \\ &\quad s.t. \sum_{j=1}^{|g|} z_j^{(g)} = |g| s_g, \\ &\qquad \sum_{g=1}^{G} s_g \leq k, \ s_g, \ z_j^{(g)} \in \{0,1\}, \ j = 1, ..., |g|, \ g = 1, ..., G, \end{aligned} \tag{4.10}$$

where $\boldsymbol{z}$ is the vector of binary variables containing $z_j^{(g)}$, which is the binary variable for the variable j in the group g, $|g|$ is the number of variables in the group g, G is the total number of groups, $\boldsymbol{s}$ is the vector of binary variables containing s_g, which is the introduced slack binary factor in the constraint 4.10 indicating the selection of the whole group g or not. When $s_g = 1$, all the integer factors corresponding to the members in the group are forced to be 1. In words, all members in that group are selected. When $s_g = 0$, all the integer factors corresponding to the members in the group are forced to be 0, that is, all members in that group are excluded from the model.

4.4.3 Sparse group selection

The sparse group selection is used in cases when both sparsity of groups and within each group are desired (Simon et al. 2013), such as the identification of key genes in important pathways in biological studies. We propose a two-stage procedure to perform the sparse group selection in the framework of best subset problems. Stage 1, conduct the group selection to obtain the binary variables, $\boldsymbol{z}^*$ and $\boldsymbol{s}^*$, indicating the selection of the predictor variables and the group, respectively. Stage 2, perform the following LIO formulation to select variables within groups. The constraint 4.11 can be viewed to enforce that at least one variable is selected within the selected group. The constraint 4.12 is to inherent the results $\boldsymbol{z}^*$ and $\boldsymbol{s}^*$ from the group selection in Stage 1.

$$
\begin{aligned}
\underset{\boldsymbol{z}}{\text{maximize}} \quad & \sum_{g=1}^{G} \sum_{j=1}^{|g|} z_j^{(g)} \frac{\mu_j^{2,(g)}}{d+\gamma} \\
\text{subject to} \quad & \sum_{j=1}^{|g|} z_j^{(g)} \geq s_g^*, && (4.11) \\
& z_j^{(g)} \leq z_j^{*,(g)}, && (4.12) \\
& \sum_{g=1}^{G} \sum_{j=1}^{|g|} z_j^{(g)} \leq k, \; z_j^{(g)} \in \{0,1\}, \; j = 1, ..., |g|, \; g = 1, ..., G,
\end{aligned}
$$

where $\boldsymbol{z}$ is the vector of binary variables containing $z_j^{(g)}$.

4.4.4 Overlapping group selection

The overlapping group selection arises in many biological problems. For example, the functional groups of genes may overlap due to genes' multi-functionality (Jacob et al. 2009). Inspired by Obozinski et al. (2011), we propose to transform the overlapping groups to the non-overlapping groups by duplicating overlapping predictor columns among the overlapping groups. This procedure reduces the overlapping group selection problems to the regular group selection problems with identical overlapping predictors among overlapping groups. We call the resulting columns, their corresponding coefficients and objective values as latent columns, latent coefficients and latent objective values, respectively. Then we design the following LIO formulation with latent integer factors, latent objective values, and latent slack group integers to implement the overlapping group selection,

$$
\begin{aligned}
&\underset{\boldsymbol{z},\boldsymbol{s}}{\text{maximize}} \sum_{g=1}^{G}\sum_{j=1}^{|g|} z_j^{(g)} \frac{\mu_j^{2,(g)}}{d+\gamma} \\
&\text{subject to } \sum_{j=1}^{|g|} z_j^{(g)} \leq |g| s_g, && (4.13)\\
&\quad z_j^{(g)} + z_{j'}^{(g')} \leq 1,\ (z_j^{(g)}, z_{j'}^{(g')}) \in \{(z_j^{(g)}, z_{j'}^{(g')}) : x_j^{(g)} = x_{j'}^{(g')}, g \neq g'\} && (4.14)\\
&\quad \sum_{g=1}^{G} s_g \leq k,\ s_g \in \{0,1\},\ g = 1,...,G,\ z_j^{(g)} \in \{0,1\},\ j = 1,...,|g|,
\end{aligned}
$$

where $\boldsymbol{z}$ is the vector of binary variables containing $z_j^{(g)}$, the set $\{(z_j^{(g)}, z_{j'}^{(g')}) : x_j^{(g)} = x_{j'}^{(g')}, g \neq g'\}$ represents the integers corresponding to the overlapping columns $x_j, x_{j'}$ in different groups. $\boldsymbol{s}$ is the vector of binary variables containing s_g. The constraint 4.13 represents that when $s_g = 0$, all the latent integers corresponding to the members in the latent group are forced to be 0, that is, all members are excluded from the model. The constraint 4.14 enforces only one of overlapping columns is selected. Last, we transform the latent coefficients back to the original coefficients. In the proposed LIO formulation approach, we reveal that the optimization in the overlapping group selection is not more challenging than the regular group selection.

4.4.5 Sparse overlapping group selection

Similar to the relationship between the group selection and the sparse group selection, the sparse overlapping group selection is an extension of the overlapping group selection (Rao et al. 2013; Park et al. 2015). We propose to use the previously mentioned two-stage procedure in the sparse group selection to implement the sparse overlapping group selection. After performing the overlapping group selection, we conduct the within group selection with latent binary integers, latent objective values, and latent slack group integers, using the following LIO formulation.

$$\begin{aligned}
\underset{\boldsymbol{z}}{\text{maximize}} & \sum_{g=1}^{G}\sum_{j=1}^{|g|} z_j^{(g)} \frac{\mu_j^{2,(g)}}{d+\gamma} \\
\text{subject to } & \sum_{j=1}^{|g|} z_j^{(g)} \geq s_g^*, \\
& z_j \leq z_j^{*,(g)}, \\
& \sum_{g=1}^{G}\sum_{j=1}^{|g|} z_j^{(g)} \leq k,\ z_j^{(g)} \in \{0,1\},\ j=1,...,|g|,\ g=1,...,G,
\end{aligned}$$

where $\boldsymbol{z}$ is the vector of binary variables containing $z_j^{(g)}$. Same as the overlapping group selection, we need to transform the latent coefficients back to the original coefficients corresponding to the original columns.

4.4.6 Hierarchical sparse overlapping group selection

Motivated by the recently proposed conditional main effects (CME) analysis (Su and Wu 2017; Wu, J. 2018; Mak and Wu, 2019), we propose the hierarchical sparse overlapping group selection method in the framework of best subset selection problems to address the variable selection issues in the CME analysis. The CME, defined as the conditional effect of a factor at a fixed level of another factor, is intended to provide a better understanding of traditional interaction terms. Given the predictor variables A and B, their interaction term, INT(AB), can be reformulated as four CMEs, CME($A|B+$), CME($A|B-$), CME($B|A+$), and CME($B|A-$), under the relationship INT(AB) $= \frac{1}{2}\ (\ \text{CME}(A|B+) - \text{CME}(A|B-)\ )$ $= \frac{1}{2}\ (\ \text{CME}(B|A+) - \text{CME}(B|A-)\ )$. It is challenging to perform variable selection within

the CME framework. First, the CME reformulation increases the dimension of predictors from $p + \binom{p}{2}$ to $p + 4\binom{p}{2}$, leading to high dimension problems even when the number of main effects p is moderately large. Second, there exist two underlying group structures: (1) Siblings, defined as the CMEs with the same parent effect, such as CME($A|B+$) and CME($A|C+$). (2) Cousins, defined as the CMEs with same conditional effects, such as CME($A|C+$) and CME($B|C+$). Third, there is additional hierarchical structure between the CME and its parent effect.

Mak and Wu (2019) proposed a bi-level selection method based on the group MCP method (Zhang 2010). However, their approach treated the sibling and cousin groups separately. Given the overlapping predictor among sibling and cousin groups, we propose the hierarchical sparse overlapping group selection method in the framework of best subset selection problems to perform the variable selection in two steps. The proposed method is an extension of the sparse overlapping group selection method. We first perform the above overlapping group selection method, then we conduct the following LIO formulation with latent binary integers, latent slack group integers, and latent objective values:

$$
\begin{aligned}
\underset{\boldsymbol{z}}{\text{maximize}} \; & \sum_{g=1}^{G} \sum_{j=1}^{|g|} z_j^{(g)} \frac{\mu_j^{2,(g)}}{d+\gamma} \\
\text{subject to} \; & \sum_{i \in CME} z_{(i)}^{(g)} \leq \alpha z_{(parent)}^{(g)}, \qquad (4.15) \\
& \sum_{j=1}^{|g|} z_j^{(g)} \geq s_g^*, \\
& z_j^{(g)} \leq z_j^{*,(g)}, \\
& \sum_{g=1}^{G} \sum_{j=1}^{|g|} z_j^{(g)} \leq k, \; z_j^{(g)} \in \{0,1\}, \; j = 1, ..., |g|, \; g = 1, ..., G,
\end{aligned}
$$

where $\boldsymbol{z}$ is the vector of binary variables containing $z_j^{(g)}$, $\sum_{i \in CME} z_{(i)}^{(g)}$ is the sum of binary variables corresponding to the CME term in the group g, $z_{(parent)}^{(g)}$ is the integer factor corresponding to the parent effect of the CME term in the group g. Note that each group is composed of CME terms and their parent effect terms. The constant α in the constraint 4.15 controls the level of hierarchy within the group g. In this study, we set $\alpha = min(3, G/4)$.

We have demonstrated the power of the LIO formulation in various structured variable

selection problems. In fact, the application of the LIO formulation to incorporate structure knowledge in the variable selection process is not limited to these examples. For example, in cases where certain variables must be included in the model, we can fix the binary variable as one to account for this special requirement.

4.5 Optimization Algorithm

The computational algorithm for the sparse ridge regression method is summarized in Algorithm 3, which is implemented in the package SPRsubset in R software and is available in the Bitbucket (`https://bitbucket.org/vtshen/rpackages/src/master/`).

Algorithm 3

Input: X, $\boldsymbol{y}$, a sequence of k
Compute X^TX, $X^T\boldsymbol{y}$, $A = dI_p - X^TX$, and the largest eigenvalue d.
for each γ from ridge penalty **do**
  Compute $d_{adj} = d + \gamma$,
  for each k from the L_0 constraint **do**
    Initialize $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{r}^{(0)} = \boldsymbol{y} - X\boldsymbol{\beta}^{(0)}$
    for iteration t **do**
      if $n > p$ **then** $\boldsymbol{\mu}^{(t+1)} = X^T\boldsymbol{y} + A\boldsymbol{\beta}^{(t)}$,
      else $\boldsymbol{\mu}^{(t+1)} = X^T\boldsymbol{r}^{(t)} + d_{adj}\boldsymbol{\beta}^{(t)}$,
      end if
      $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\mu}^{(t+1)}/d_{adj}$,
      perform the structured variable selection based on the specific LIO formulation,
      if $n \leq p$ **then** $\Delta\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}$; $\boldsymbol{r}^{(t+1)} = \boldsymbol{r}^{(t)} - X\Delta\boldsymbol{\beta}^{(t+1)}$,
      end if
      Break when the maximum relative difference in $\boldsymbol{\beta}_j <$ tolerance.
    end for
  end for
end for
Output $\boldsymbol{\beta}$.

In practice, we notice that the algorithm can be very slow in convergence, which is common in other EM algorithms. We use two computational tools to speed up the tuning procedures: (1) Warm starts, which makes use of the converged solution from previous tuning parameters as the initial value for the problem at the current tuning parameter. (2) Pre-storage, which computes most constraints in the LIO formulation so that we can avoid unnecessary repeated construction for those constraints during iterations of the EM algorithm. Another approach

which performs the (ridge penalized) least squares estimation directly on the active set, is proposed in Xiong (2014) to avoid the slow iterations in achieving the convergence. This tool is similar to the so-called active set optimization, which performs coordinate descent updates over a small subset of active variables (Meier et al. 2008); Friedman, Hastie, and Tibshirani (2010)). We found such tool effective in reducing the iterations needed to achieve convergence. In the computation for the LIO formulation, when the occurrence of the nonzero component pattern in $\beta^{(t)}$ exceeds a certain threshold, we run

$$\begin{aligned} A^{(t)} =& \{j : \beta_j^{(t)} \neq 0\}, \\ \beta_{A^{(t)}}^{(t)} =& (X_{A^{(t)}}^T X_{A^{(t)}} + \gamma I_{|A^{(t)}|})^- X_{A^{(t)}}^T y, \end{aligned}$$

where $|A^{(t)}|$ is the cardinality of the support $A^{(t)}$. We run an additional iteration to verify the solution. If that support A does not change from the last iteration, we terminate the algorithm.

4.6 Real-Data Analysis

In this section, we analyze three real-data problems, the housing price study in suburbs of Boston, the infant birth weight study, and the polygenic association study on fly wing shape, to evaluate the model performance of the sparse ridge regression method with structured best subset selection.

4.6.1 Housing price in suburbs of Boston

The Boston housing price study with the aim to identify key predictors and to predict the median value of owner-occupied homes. The continuous response is the median value of owner-occupied homes. The original 13 predictor variables, listed in Table 4.1, are expanded by adding their pairwise two-way interactions. The total number of variables in the expanded regression matrix is 91 and the total number of observations is 506. The Boston data set is available from the MASS library in R software.

We perform the analysis from the methods the best subset selection (L0Learn_L0) and the ridge penalized best subset selection (L0Learn_L0L2) (Hazimeh and Mazumder 2019), the weak heredity sparse ridge regression method (SRRw), the strong heredity sparse ridge

Table 4.1: List of variable names and their acronyms in the Boston housing price study.

Variable	Acronym
per capita crime rate by town	crim
proportion of residential land zoned for lots over 25,000 sq.ft	zn
proportion of non-retail business acres per town	indus
Charles River dummy variable	chas
nitrogen oxides concentration	nox
average number of rooms per dwelling	rm
proportion of owner-occupied units built prior to 1940	age
weighted mean of distances to five Boston employment centres	dis
index of accessibility to radial highways	rad
full-value property-tax rate	tax
pupil-teacher ratio by town	ptratio
the adjusted proportion of African-Americans by town	black
lower status of the population	lstat

regression method (SRRs), the sparse ridge regression method with $\gamma = 0$ (SRR_L0) and the sparse ridge regression method (SRR).

The methods L0Learn_L0 and L0Learn_L0L2, relying on the coordinate descent algorithm and local combinatorial search method (Hazimeh and Mazumder 2019), intend to solve the following two problems

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\ ||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_0 + \gamma||\boldsymbol{\beta}||_2^2,$$

and

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\ ||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_0,$$

respectively, where λ is the tuning parameter for the L_0-norm term $||\boldsymbol{\beta}||_0$.

The method SRR_L0 solves the following problem

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\ ||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2 \text{ subject to } ||\boldsymbol{\beta}||_0 \leq k.$$

Figure 4.2 shows the selected effects from these compared models. The SRR_L0 and the SRR select more variables than the L0Learn_L0 or the L0Learn_L0L2. The SRRs selects less variables than the SRRw and the SRR.

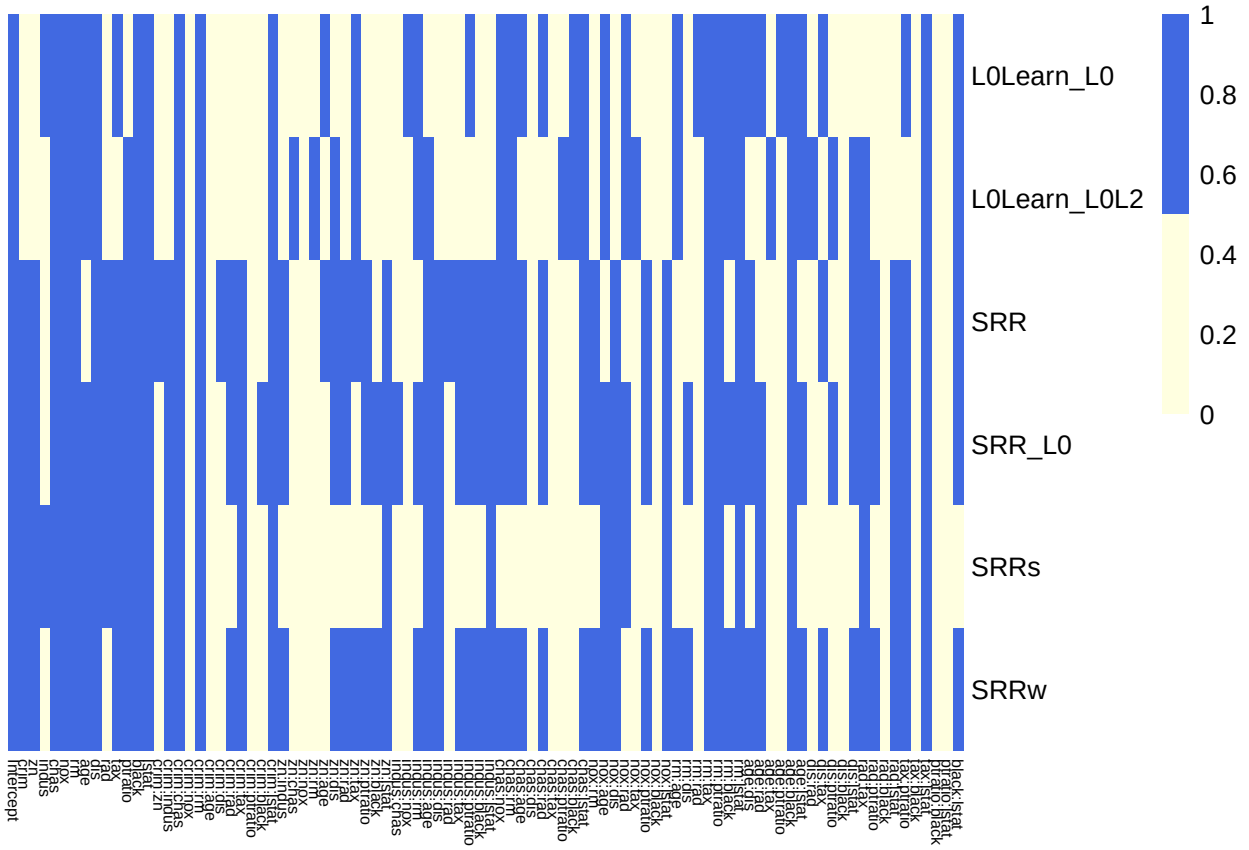

Figure 4.2: Variables selected from the compared methods in the Boston housing price study. Blue indicates selection and yellow indicates no selection.

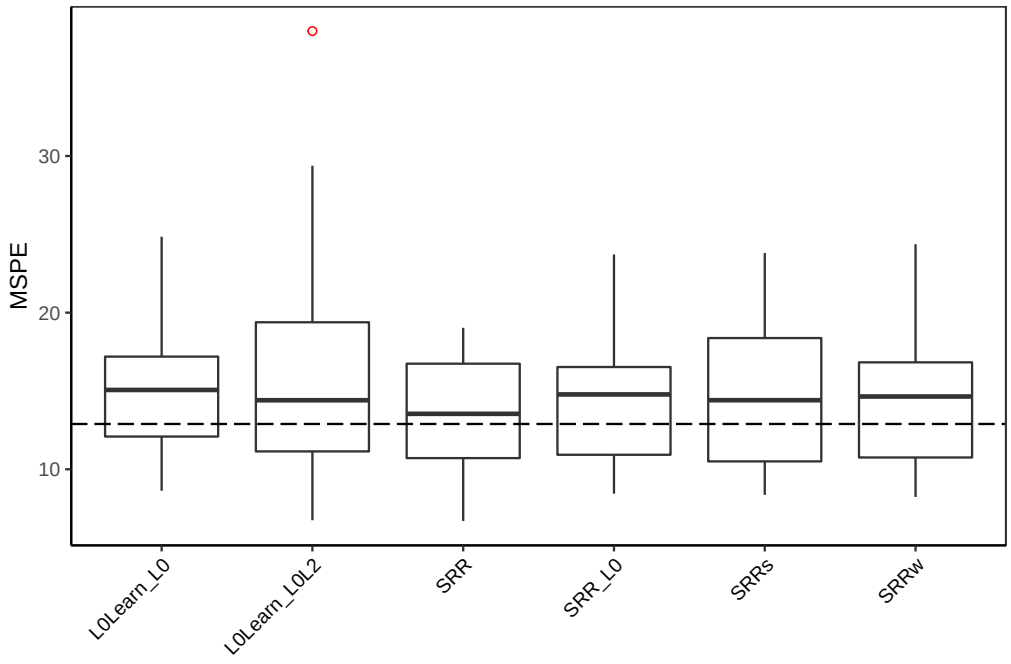

Figure 4.3: Boxplots of mean-squared prediction error (MSPE) from the compared methods in the Boston housing price study.

Figure 4.3 shows the mean-squared prediction error (MSPE) boxplots from the compared methods in predicting the median housing values. Table 4.2 shows the mean and standard error of MSPE from the compared methods in predicting the median housing values. The MSPE is estimated from the randomly sampled 20% test data after model training on the rest 80% of the data. The process is repeated 25 times to evaluate error variability. We see that the ridge penalized best subset selection methods (SRR and L0Learn_L0L2) have better prediction performance than their corresponding best subset selection method (SRR_L0 and L0Learn_L0). The SRR has the best prediction performance among the compared methods.

4.6.2 Infant birth weight study

The infant birth weight study has been analyzed in literature (Yuan and Lin 2006; Chen et al. 2016). The goal of study is to identify key predictors and to predict the infant birth weight. Particularly, we are interested in the problem of selecting one representative variable from each selected groups. The continuous response is the infant birth weight. The eight groups, including mother's age (age1, age2, age3), mother's weight (lwt1, lwt2,

Table 4.2: Performance comparisons of models from 25 replications (means and standard errors) in the Boston housing price study.

	MSPE	
method	mean	sd
L0Learn_L0	42.19	10.852
L0Learn_L0L2	16.33	7.470
SRR	13.71	3.751
SRR_L0	14.76	4.428
SRRs	14.58	4.425
SRRw	14.49	4.411

lwt3), mother's race (white, black), smoking status during pregnancy (smoke), number of previous premature labours (ptl1, ptl2m), history of hypertension (ht), presence of uterine irritability (ui), number of physician visits during the first trimester (ftv1, ftv2, ftv3m), have 16 predictor variables. In total, there are 189 observations. The data set is collected by the Baystate Medical Center, Springfield, Massachusetts, in 1986, and is available from the grpreg package in R software.

We perform the analysis from the methods group lasso (gglasso), sparse group lasso (SGL and grpreg), group best subset selection (SRRgrp), sparse group best subset selection (SRRsgrp), and the sparse group best subset selection with one representative in each selected group (SRRagent). Figure 4.4 shows the selected variables from these compared models. Among the compared methods, the SRRagent method is the only one being able to perform the selection of representative variables from each identified group. Specifically, the SRRagent selects seven groups and within each selected group the representative variables are lwt1, white, smoke, ptl1, ht, ui, and ftv1. The SRRgrp excludes the group variable number of physician visits during the first trimester (ftv1, ftv2, ftv3m). The SRRsgrp selects all the variables except the variable age1. The gglasso selects all variables as expected. The SGL yields sparse group selection while the grpreg selects all variables.

Figure 4.5 shows the mean-squared prediction error (MSPE) boxplots from the compared methods in predicting the infant birth weight values. Table 4.3 shows the mean and standard error of MSPE from the compared methods in predicting the infant birth weight values. The MSPE is estimated from the randomly sampled 20% test data after the model training on the rest 80% of the data.The process is repeated 25 times to evaluate error variability. We see that the sparse group selection methods, including SGL, grpreg, SRRsgrp, achieves better prediction results than their corresponding group selection methods gglasso and SRRgrp,

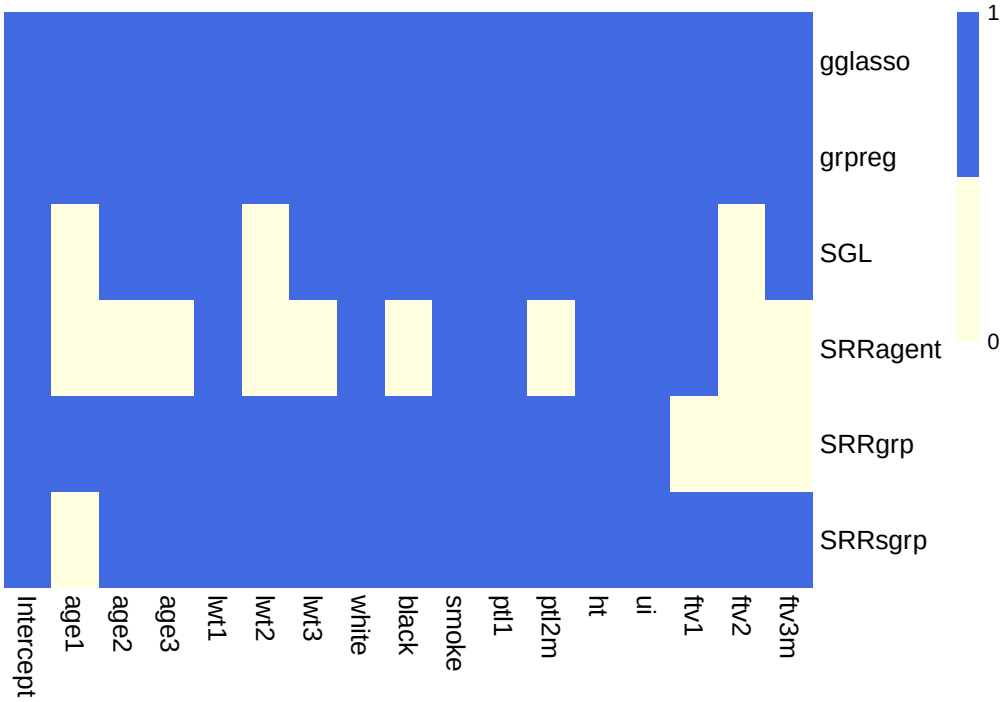

Figure 4.4: Variables selected from the compared methods in the infant birth weight study. Blue indicates selection and yellow indicates no selection.

respectively.

4.6.3 Polygenic Association Study on Fly Wing Shape

In this study, the interest is on the investigation of the polygenic association for the wing shape of one common fruit fly, known as Drosophila melanogaster. Specifically, we focus on the selection of important conditional main effects (CMEs), defined as the effects of a gene conditional on another gene being active or absent. This problem has been addressed by Mak and Wu (2019), but here we apply the proposed structured best subset selection method to the problem to demonstrate the usefulness of our proposed method. The data set is obtained from the study by Weber et al. (2001). The total number of observations is 701. The response is a continuous index for wing shape, which in this study is scaled by dividing their square root of the mean square. The 48 predictor predictors are binary. After removing the redundant 11 variables, we use the remaining 37 predictor variables in the analysis. The predictor variables are then expanded by constructing the CMEs following the cmenet method (Mak and Wu 2019). After the CME expansion, the total number of

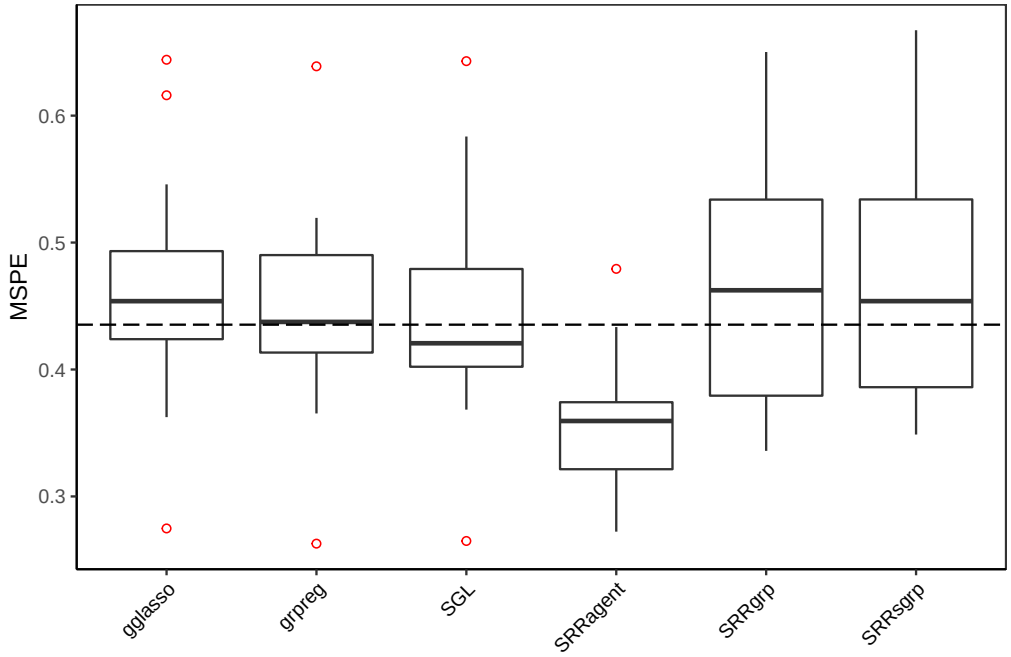

Figure 4.5: Boxplots of mean-squared prediction error (MSPE) from the compared methods in the infant birth weight study.

Table 4.3: Performance comparisons of models from 25 replications (means and standard errors) in the infant birth weight study.

	MSPE	
method	mean	sd
gglasso	0.4595	0.07671
grpreg	0.4507	0.06910
SGL	0.4469	0.07466
SRRagent	0.3529	0.04899
SRRgrp	0.4578	0.08923
SRRsgrp	0.4687	0.09125

Table 4.4: Number of variables selected from the compared methods in the fly wing shape study.

hiernet	cmenet	SRRcme_k14	SRRcme_k51
51	14	14	51

Table 4.5: Performance comparisons of models (means and standard errors) in the wing shape study.

	MSPE	
method	mean	sd
cmenet	0.0124	0.0019
hiernet	0.0124	0.0013
SRRcme_k14	0.0173	0.0020
SRRcme_k51	0.0143	0.0011

predictor variables is $37 + 4\binom{37}{2} = 2701$.

We perform the analysis from the method LASSO, hierarchical LASSO (hiernet), CME analysis (cmenet), hierarchical sparse overlapping group selection with different k values: SRRcme_k14 and SRRcme_k51. Different from other methods, the method hiernet performs the selection on the main effects and their two-way interactions. Table 4.4 shows the selected effects from these compared models.

Table 4.5 shows the mean-squared prediction error (MSPE) boxplots from the compared methods in predicting the fly wing index. The MSPE is estimated from the same procedure as in previous real-data problems. The SRRcme has comparable prediction performance as the cmenet and the hiernet, but is able to control the number of selected variables preciously.

4.7 Discussion

The structured variable selection plays an important role in extracting useful knowledge from high dimensional data set. Although many methods within the LASSO framework have been proposed, the structured variable selection methods based on the subset selection method are rare in the literature. In this work, we propose a new variable selection technique within the framework of the best subset selection problems. In the proposed sparse ridge regression method, we re-construct the regression matrix in the angle of experimental designs. We

employ the expectation-maximization (EM) algorithm to formulate the best subset selection problem as an iterative linear integer optimization (LIO) problem. The LIO formulation can be applied to various structured variable selection problems. We demonstrate the usefulness of the proposed method in various structured variable selection problems. The performance of the proposed method is evaluated in numerical studies.

Although the focus is on the sparse ridge regression in this work, the proposed framework can be extended to the problem with the L_1-norm penalty and the L_0-norm constraint. The estimator of the problem is reported to select less variables than the classical LASSO method (Mazumder et al. 2017). The extension is not difficult and we present the key procedures in the derivation for the solution in Appendix C. The proposed method can also be extended to the situations where the response of interest is binary or count. One approach is to replace the log-likelihood function in the least-squares estimation problem by the log-likelihood function in the logistic regression method or in the poisson regression method. For the parameter estimation, we can apply the iterative reweighted least squares method in the **M-step** in the EM algorithm. It is also interesting to investigate the use of zero-order fused term (Land and Friedman 1997), $||F\boldsymbol{\beta}||_0$ with F as a certain matrix, in the LIO formulation for the purpose of variable selection.

Chapter 5 Future Work

In this dissertation, we aim to incorporate underlying structure knowledge in the variable selection process for improving model interpretation and model prediction. We develop three different variable selection methods to accommodate different problems. In Chapter 2, we consider an additive structure to inherently connect the major components with the minor components in the mixture-of-mixtures experiment. In Chapter 3, we propose a regularized dynamic logistic regression model for the variable selection of dynamic process variables. In Chapter 4, we investigate a structured variable selection method within the framework of the best subset selection. We demonstrate our proposed model performance in various numerical and real-data problems.

We emphasize here several interesting extension of our proposed methods. In the proposed AHM model in Chapter 2, a monotonic and bounded function on the domain of the major components is proposed. Any appropriate nonparametric form, for example, smoothing splines, representing the relationship between the major minor components can be used in the proposed AHM method. In Chapter 3, we consider the data used in the proposed method are historical. It is worth investigating the application of the proposed method to the online data set. Additionally, we can apply the proposed method in Chapter 4 to other types of penalty terms, for example, the zero-order fused term, $||F\boldsymbol{\beta}||_0$ with F as a certain matrix, to address the structures within temporally or spatially adjacent variables.

References

(Chapter 1)

Adhikari, S., Lecci, F., Becker, J.T., Junker, B.W., Kuller, L.H., Lopez, O.L. and Tibshirani, R.J., (2019). High dimensional longitudinal classification with the multinomial fused lasso. *Statistics in medicine*, https://doi.org/10.1002/sim.8100.

Ahmed, A. and Xing, E.P. (2009). Recovering time-varying networks of dependencies in social and biological studies. *PNAS*, 106, 11878-83.

Beale, E.M.L., Kendall, M.G. and Mann, D.W. (1967). The discarding of variables in multivariateanalysis, *Biometrika*, 54(3/4), 357-66.

Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2), 813-52.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4), 373–84.

Brodsky, B.E. and Darkhovsky, B.S. (1993) Nonparametric methods in change-point problems. Kluwer Academic Publishers, The Netherlands.

Cantoni, E., Flemming, J. M., and Ronchetti, E. (2011). Variable selection in additive models by non-negative garrote. *Statistical Modeling*, 11(3), 237-52.

Chen, S., Donoho, D. L. and Saunders, M. (1998). Atomic decomposition for basis pursuit. *SIAM Journal on Scientific Computing*, 20(1), 33-61.

Chib, S. (1998) Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86, 221-41.

Cleveland, W.S., Grosse, E., and Shyu, W.M. (1991). Local regression models. In *Statistical Models in S*, 309-76.

Dedrick, R.F., Ferron, J.M., Hess, M.R., Hogarty, K.Y., Kromrey, J.D., Lang, T.R., Niles, J.D. and Lee, R.S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, 79(1), 69-102.

Draper, N. and Smith, H. (1966). Applied Regression Analysis. Wiley.

Efroymson, M. (1966). Stepwise regression-a backward and forward look. *Eastern Regional Meetings of the Institute of Mathematical Statistics.*

Fan, J. and Zhang, W. (2008). Statistcal methods with varying coefficient models. *Statistics and its interface.* 1(1), 179-95.

Friedman, J., Hastie, T., and Tibshirani, R. (2007) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432-41.

Friedman, J., and Hastie, T., and Hoefling, H, and Tibshirani, R. (2007). Pathwise coordinate optimization. *The annals of applied statistics*, 1(2), 302-32.

Friedman, J., Hastie, T. and Tibshirani, R., (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432-41.

Friedman, J. and Hastie, T., and Tibshirani, R. (2009). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), 1-22.

Guo, F. Hanneke, S., Fu, W., and Xing, E. (2007). Recovering temporally rewiring networks: A model-based approach. *In Proceedings of the 24th international conference of machine learning*, 321-28.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4) 757-96.

Heck, R.H. and Thomas, S.L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus*, Routledge, New York.

Hocking, R.R. and Leslie, R.N. (1967). Selection of the best subset in regression analysis. *Technometrics*, 9(4), 531-40.

Kang, L., Joseph, V.R. and Brenneman, W.A. (2011). Design and modeling strategies for mixture-of-mixtures experiments. *Technometrics*, 53(2), 125–36.

Kevin, B. and Jean-Philippe, V. (2011). The group fused lasso for multiple change-point detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1), 91-108.

Kim S. J., Koh,. K., Boyd, S., and Gorinevsky, D. (2009). L1 trend filtering. *SIAM review*, 51(2), 339-60.

Kim, S., and Xing, E. (2012). Tree guided group lasso for multi-task regression with structured sparsity. *The Annals of Applied Statistics*, 6(3), 1095-117.

Lavielle, M., and Lebarbier, E. (2001). An application of MCMC methods for the multiple change-points problem. *Signal Processing*, 81, 39-53.

Lebarbier, E. (2005). Detectiing multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85, 717-36.

Liu, J. and Ye, J., (2010). Moreau-Yosida regularization for grouped tree structure learning. *In Advances in neural information processing systems*, 1459-67.

Jacob, L., Obozinski, G., and Vert, J. (2009). Group lasso with overlap and graph lasso. *In proceedings of the 26th annual international conference on machine learning*, 433-40.

Jin, R., Deng, X., Chen, X., Zhu, L., and Zhang, J. (2019). Dynamic quality-process model in consideration of equipment degradation. *Journal of Quality Technology*, 1-13.

Mak, S. and Wu, J. (2019). cmenet: a new method for bi-level variable selection of conditional main effects. *Journal of American Statistician Association*, 114(526), 844-856.

Meier, S., Geer, De., and Buhlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 53-71.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic linear models with R*. Springer, New York.

Rojas, R. C. and Wahlberg, B. (2014). On change point detection using the fused lasso method. *arXiv preprint arXiv:1401.5408*,

Rudin, L. I., Osher, S., and Faterni, E., (1992). Nonlinear total variation based noise removal algorithm. *Physica D: nonlinear phenomena*, 60(1-4), 259-68.

Scheel H.J. and Fukuda, T.(2003). Theoretical and experimental solutions of the striation problem. *Crystal Growth Technology*, 69-91.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R., (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2), 231-45.

Su, H. and Wu, J. (2017). CME Analysis: A new method for unraveling aliased effects in two-level fractional factorial experiments. *Journal of Quality Technology*, 49(1), 1-10.

Sun, H., Deng, X., Wang, K., and Jin, R. (2016). Logistic regression for crystal growth process modeling through hierarchical nonnegative garrote-based variable selection. *IIE Transactions*, 48(8), 787-96.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Sta-tistical Society: Series B*, 58(1), 267-88.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1), 91-108.

Tibshirani, R.J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1), 285-323.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models, 2nd Edition*, Springer, New York.

Wu, J. (2018). A fresh look at effect aliasing and interactions: some new wine in old bottles. *Ann Inst Stat Math*, 70, 249-68.

Xiong, S. (2010). Some notes on the nonnegative garrote. *Technometrics*, 52(3), 349–61.

Xiong, S. 2014. Better Subset Regression. *Biometrika*, 101(1), 71-84.

Xiong, S., Dai, B., Huling, J., and Qian, P.Q.Z. (2016). Orthogonalizing Em: A Design-Based Least Squares Algorithm. *Technometrics*, 58(3), 285-93.

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49-67.

Yuan, M. and Lin. Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2), 143–61.

Zhang, J., Li, W., Wang, K., and Jin, R. (2014). Process adjustment with an asymmetric quality loss function. *Journal of Manufacturing Systems*, 33(1), 159-65.

Zhao, P., Rocha, G. and Yu, B., (2006). Grouped and hierarchical model selection through composite absolute penalties. Department of Statistics, UC Berkeley, *Technical Report*, 703.

Zhou S. Lafferty, J. and Wasserman, L. (2008). Time varying undirected graphs. *Machine Learning*, 80(2-3), 295-319.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2003) L1 norm support vector machines. *In Advances in neural information processing systems*, 49-56.

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

(Chapter 2)

Borges, C., Bruns, E. R., Almeida, A. A., and Scarminio, I. S. (2007). Mixture-mixture design for fingerprint optimization of chromatographic mobile phases and extraction solutions for Camellia sinensis. *Analytica Chimica Acta*, 595(1-2), 28-37.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4), 373–84.

Brown, L., Donev, A. N., and Bissett, A. C. (2015). General blending models for data from mixture experiments. *Technometrics*, 57(4), 449-56.

Coetzer, R. L. and Haines, L. M. (2013). Optimal designs for multiple-mixture by process variable experiments. In: Ucinski D., Atkinson A., Patan M. (eds) *mODa 10 - Advances in Model-Oriented Design and Analysis.* Contributions to Statistics. Springer, Heidelberg.

Cornell, J. A., and Good, I. J. (1970), The mixture problem for categorized components. *Journal of the American Statistical Association*, 65(329), 339-55.

Cornell, J.A. (1986). A comparison between two ten point designs for studying three component mixture systems. *Journal of Quality Technology*, 18(1), 1-15.

———. (2002). *Experiments with mixtures: designs, models, and the analysis of mixture data, 3rd ed,*, John Wiley & Sons, New York.

Cornell, J.A. and Ramsey, P.J. (1998). A Generalized mixture model for categorized-components problems with an application to a photoresist-coating experiment. *Technometrics*, 40(1), 48-61.

Cox, D.R. (1971). A note on polynomial response functions for mixtures. *Biometrika*, 58(1), 155–59.

Dedrick, R.F., Ferron, J.M., Hess, M.R., Hogarty, K.Y., Kromrey, J.D., Lang, T.R., Niles, J.D. and Lee, R.S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, 79(1), 69-102.

Didier, C., Etcheverrigaray, M., Kratje, R., and Goicoechea, H. C. (2007). Crossed mixture design and multiple response analysis for developing complex culture media used in recombinant protein production. *Chemometrics and Intelligent Laboratory Systems*, 86(1), 1-9.

Dingstad, G., Egelandsdal, B. and NæS, T. (2003). Modeling methods for crossed mixture experiments—a case study from sausage production. *Chemometrics and Intelligent Laboratory Systems*, 66(2), 175-90.

Draguljić, D., Woods, D.C., Dean, A.M., Lewis, S.M., and Vine, A.E. (2014). *Screening strategies in the presence of interactions. Technometrics*, 56(1), 1-16.

Draper, N.R. and Pukelsheim, F. (1998). Mixture models based on homogeneous polynomials. *Journal of Statistical Planning and Inference*, 71(1), 303–11.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Annals of Statistics*, 27(5), 1491-518.

Goos, P., Jones, B. and Syafitri, U. (2016). I-optimal design of mixture experiments. *Journal of the American Statistical Association*, 111(514), 899-911.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4) 757-96.

Johnson, M.E., Moore, L.M. and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26(2), 131–48.

Kang, L., Joseph, V.R. and Brenneman, W.A. (2011). Design and modeling strategies for mixture-of-mixtures experiments. *Technometrics*, 53(2), 125–36.

Kang, L., Salgado, J.C., and Brenneman, W.A. (2016). Comparing the slack-variable mixture model with other alternatives. *Technometrics*, 58(2), 255–68.

Khuri, A.I. (2005). Slack-variable models versus Scheffé's mixture models. *Journal of Applied Statistics*, 32(9), 887–908.

Laake, P. (1975). On the optimal allocation of observations in experiments with mixtures. *Scandinavian Journal of Statistics*, 2(3), 153-57.

Lambrakis, D.P. (1968). Experiments with mixtures: A generalization of the simplex-lattice Design. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(1), 123–36.

———. (1969). Experiments with mixtures: estimated regression function of the multiple-lattice Design. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31(2), 276–84.

Piepel, G. F. (1999). Modeling methods for mixture-of-mixtures experiments applied to a tablet formulation problem. *Pharmaceutical Development and Technology*, 4(4), 593-606.

Piepel, G.F. (2007). A component slope linear model for mixture experiments. *Quality Technology & Quantitative Management*, 4(3), 331–43.

Lawson, J. and Willden, C. (2016). Mixture experiments in R using mixexp. *Journal of Statistical Software*, 72(c02).

Prescott, P., Dean A.M., Draper, N.R. and Lewis, S.M. (2002). Mixture experiments: ill-conditioning and quadratic model specification. *Technometrics*, 44(3), 260–68.

Scheffé, H. (1958). Experiments with mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2), 344–60.

———. (1963). The simplex-centroid design for experiments with mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 25(2), 235–63.

Snee, R.D. and Rayner, A.A. (1982). Assessing the accuracy of mixture model regression calculations. *Journal of Quality Technology*, 14(2), 67-79.

Wu, C. F. J, and Hamada, M. S. (2009). *Experiments: Planning, Analysis, and Optimization, 2nd Edition*, Wiley, Hoboken, NJ.

Xiong, S. (2010). Some notes on the nonnegative garrote. *Technometrics*, 52(3), 349–61.

Yuan, M. and Lin. Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2), 143–61.

Yuan, M., Joseph, V.R. and Lin, Y. (2007). An efficient variable selection approach for analyzing designed experiments. *Technometrics*, 49(4), 430–39.

Yuan, M., Joseph, V.R. and Zou, H. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics*, 3(4), 1738–57.

(Chapter 3)

Adhikari, S., Lecci, F., Becker, J.T., Junker, B.W., Kuller, L.H., Lopez, O.L. and Tibshirani, R.J., (2019). High dimensional longitudinal classification with the multinomial fused lasso. *Statistics in medicine*, 38(12), 2184-205.

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18(3), 338-57.

Ahmed, A., and Xing, E.P. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29), 11878-83.

Arnold, T., and Tibshirani, R. (2016). Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, 25(1), 1-27.

Beaulieu, C., Chen, J., and Sarmiento, J.L. (2012). Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transactions of the Royal Society A*, 370, 1228-49.

Bleakley, K., and Vert, J. (2011). The group fused lasso for multiple change-point detection. *ArXiv Preprint* ArXiv:1106.4199.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1), 1-122.

Cai, Z.W., Fan, J.Q., and Li, R.Z. (2000). Efficient estimation and inferences for varying-coefficient models, *Journal of the American Statistical Association*. 95(451), 888-902.

Cleveland, W.S., Grosse, E., and Shyu, W.M. (1991). Local regression models. In *Statistical Models in S*, 309-76.

Desobry, F., Davy, M., and Doncarli, C., (2004). An online kernel change detection algorithm. *IEEE Trans. Signal Processing*, 53(8-2), 2961-74.

Duan, Y., Hong, Y., Meeker, W.Q., Stanley, D.L., and Gu, X. (2017). Photodegradation modeling based on laboratory accelerated test data and predictions under outdoor weathering for polymeric materials. *The Annals of Applied Statistics* 11(4), 2052-79.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The annals of Statistics*. 27(5), 1491-518

Fan, J. and Zhang, W. (2008). Statistcal methods with varying coefficient models. *Statistics and its interface.* 1(1), 179-95.

Friedman, J. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*. 19(1), 1-67.

Friedman, J., Hastie, T., and Tibshirani, T. (2010). A note on the group lasso and a sparse group lasso. *ArXiv Preprint* ArXiv:1001.0736.

Gibberd, A.J. and Nelson, J.D. (2017). Regularized estimation of piecewise constant Gaussian graphical models: the group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, 26(3), 623-34.

Gu, X., Stanley, D., Byrd, W.E., Dickens, B., Vaca-Trigo, I., Meeker, W.Q., Nguyen, T., Chin, J.W., and Martin, J.W. (2009). Linking accelerating laboratory test with outdoor performance results for a model epoxy coating system. In *Service Life Prediction of Polymeric Materials*, Springer, Boston, MA.

Harchaoui, Z. and L*É*vy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492), 1480-93.

Hastie, T.J. and Tibshirani, R.J. (1990). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4), 757-96.

Hong, Y., Duan, Y., Meeker, W.K., Stanley, D.L., and Gu. X. (2015). Statistical methods for degradation data with dynamic covariates information and an application to outdoor weathering data. *Technometrics*, 57(2), 180-93.

Kalman, R. (1960). A new approach to linear filtering and predictioin problems. *Journal of Basic Engineering (Series D)*, 82(1), 35-45.

Kalman, R. and Bucy, R. (1963). New results in linear filtering and prediction theory. *Journal of Basic Engineering (Series D)*, 83(1), 95-108.

Kolar, M., Song, L., and Xing, E.P. (2009). Sparsistent learning of varying-coefficient models with structural changes. *Advances in Neural Information Processing Systems*, 1006-14.

Land, S.R. and Friedman, J.H. (1997) Variable fusion: a new adaptive signal regression method. Technical Report 656, Department of Statistics, Carnegie Mellon University Pittsburgh.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8), 1501-10.

Lebarbier, *É*. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing*, 85(4), 717-36.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic linear models with R*. Springer, New York.

Ramdas, A. and Tibshirani, R. (2016). Fast and flexible ADMM algorithms for trend filtering. *Journal of Computational and Graphical Statistics*, 25(3), 839-58.

Scheel H.J. and Fukuda, T.(2003). Theoretical and experimental solutions of the striation problem. *Crystal Growth Technology*, 69-91.

Simon, N., Friedman, F., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2), 231-45.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267-88.

Wahlberg, B., Boyd, S., Annergren, M., and Wang, Y. (2012). An ADMM algorithm for a class of total variation regularized estimation problems. *IFAC Proceedings Volumes*, 45(16), 83-88.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models, 2nd Edition*, Springer, New York.

Yao, Y. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, 6(3), 181-89.

Ye, G, and Xie, X. (2011). Split bregman method for large scale fused lasso. *Computational Statistics & Data Analysis*, 55(4), 1552-69.

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49-67.

Zhang, B., Geng, J., and Lai, L. (2015). Multiple change-points estimation in linear regression models via sparse group lasso. *IEEE Trans. Signal Processing*, 63(9), 2209-24.

Zhu, Y. (2017). An augmented ADMM algorithm with application to the generalized lasso problem. *Journal of Computational and Graphical Statistics*, 26(1), 195-204.

(Chapter 4)

Bakin, S. (1999). Adaptive regression and model selection in data mining problems. Ph.D. thesis, Australian National Univ., Canberra.

Bertsimas, D., King, A., and Mazumder, R. (2016) Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2), 813-52.

Bien, J., Taylor, J., and Tibshirani, R. (2013) A lasso for hierarchical interactions. *Annals of statistics*, 41(3), 1111-41.

Bixby, R. E. (2012). A brief history of linear and mixed-integer programming computation. *Documenta Mathematica, Extra Volume: Optimization Stories*, 107-121.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1), 1-122.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4), 373-84.

Chen R., Chu, C., Yuan, S., and Wu, Y. (2016). Bayesian sparse group selection. *Journal of Computational and Graphical Statistics*, 25(3), 665-683.

Chen, S., Donoho, D. L. and Saunders, M. (1998). Atomic decomposition for basis pursuit. *SIAM Journal on Scientific Computing*, 20(1), 33-61.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1-38.

Horton, N.J. and Laird, N.M. (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, 8(1), 37-50.

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. (2007) Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2), 302-32.

Friedman, J., Hastie. T, and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432-41.

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

Furnival, G.M., and Wilson, R.W. (1974). Regressions by leaps and bounds. *Technometrics*, 16(4), 499-511.

Golub, G. and Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, 2(2), 205-24.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the Lasso. arXiv Preprint arXiv:1707.08692.

Hazimeh, H. and Mazumder, R. (2019). Fast best subset selection: coordinate descent and local combinatorial optimization algorithms. https://arxiv.org/abs/1803.01454

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, 27(4), 481-99.

Jaco, L, Obozinski, G., and Vert, J. (2009). Group lasso with overlap and graph lasso. *Proceedings of the 26th annual international conference on machine learning*, 433-440.

Kim, S., and Xing, E. (2012). Tree guided group lasso for multi-task regression with structured sparsity. *The Annals of Applied Statistics*, 6(3), 1095-117.

Liu, J. and Ye, J., (2010). Moreau-Yosida regularization for grouped tree structure learning. *In Advances in neural information processing systems*, 1459-67.

Mak, S. and Wu, J. (2019). cmenet: a new method for bi-level variable selection of conditional main effects. *Journal of American Statistician Association*, 114(526), 844-856.

Mazumder, R., Radchenko, P., and Dedieu, A. (2017). Subset selection with shrinkage: sparse linear modeling when the snr is low. arXiv Preprint arXiv:1708.03288.

Meier, L., Van De Geer, S., and Buhlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 53-71.

Natarajan, B.K. (1995). Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2), 227-34.

Nesterov, Y. (2012). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2), 341-362.

Obozinski, G., Jacob, L., and Vert, J. (2011). Group lasso with overlaps: the latent group lasso approach. *arXiv preprint*, arXiv:1110.0413.

Park, H., Niida, A., Miyano, S., and Imoto, S. (2015). Sparse overlapping group lasso for integrative multi-omics analysis. *Journal of Computational Biology*, 22(2), 73-84.

Rao, N., Cox, C., Nowak, R., and Rogers, T. (2013). Sparse overlapping sets lasso for multitask learning and its application to fmri analysis. *Advances in neural information processing systems*, 2202-2210.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R., (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2), 231-45.

Su, H. and Wu, J. (2017). CME Analysis: A new method for unraveling aliased effects in two-level fractional factorial experiments. *Journal of Quality Technology*, 49(1), 1-10.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-88.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1), 91-108.

Vazquez-Alcocer, A., Schoen, E., and Goos, P. (2018). A mixed integer optimization approach for model selection in screening experiments.

Weber, K., R. Eisman, S. Higgins, L. Morey, A. Patty, M. Tausek and Z.-B. Zeng (2001), An Analysis of Polygenes Affecting Wing Shape on Chromosome 2 in Drosophila Melanogaster. Genetics, 159, 1045-1057.

Wu, J. (2018). A fresh look at effect aliasing and interactions: some new wine in old bottles. *Ann Inst Stat Math*, 70, 249-68.

Wu, C. F. J., and Hamada, M. S. (2009), *Experiments: Planning, Analysis, and Optimization (2nd ed.)*, Hoboken, NJ: Wiley.

Xie, W. and Deng, X. (2018). The CCP selector: scalable algorithms for sparse ridge regression from chance-constrained programming. arXiv Preprint arXiv:1806.03756.

Xiong, S. (2014). Better Subset Regression. *Biometrika*, 101(1), 71-84.

Xiong, S., Dai, B., Huling, J., and Qian, P. (2016). Orthogonalizing EM: a design-based least squares algorithm. *Technometrics*, 58(3), 285-93.

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49-67.

Zhao, P., Rocha, G., and Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalites. *The Annals of Statistics*, 37, 3468-3497.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2), 894-942.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2). 301-20.

Appendices

Appendix A. Additive Heredity Model for the Analysis of Mixture-of-Mixtures Experiments

In the mixture-of-mixtures (MoM) experiments, the mixture components are called the major components and can be made up of sub-components. The sub-components within the major components are called the minor components. Assume that there are q major components, and let c_k be the proportion of the kth major component. Then,

$$\sum_{k=1}^{q} c_k = 1, 0 \leq c_k \leq 1, \quad k = 1, \ldots, q.$$

Moreover, each major component is composed of m_k minor components, whose proportions with respect to c_k are x_{kj}, such that,

$$\sum_{l=1}^{m_k} x_{kl} = 1, 0 \leq x_{kl} \leq 1, \quad l = 1, \ldots, m_k.$$

A.1 Lemma 1

Lemma 1. *Any major-minor model can be written in the form of additive models.*

Proof: In the major-minor model, we denote g_1 as a function to capture the relationship between the response y and the major components. The coefficients in g_1 are functions of minor components.

Without loss of generality, consider the second-order Scheffé model on the major components for g_1:

$$g_1(c_1, ..., c_q) = \sum_{k=1}^{q} \alpha_k c_k + \sum_{1 \leq k < j \leq q} \alpha_{kj} c_k c_j + \epsilon.$$

For the coefficient α_k, we consider the second-order Scheffé model on the corresponding

minor components, that is,

$$\alpha_k(x_{k,1},...,x_{k,m_k}) = \sum_{l=1}^{m_k} \eta_l^{(k)} x_{kl} + \sum_{1\le l<l'\le m_k} \eta_{ll'}^{(k)} x_{kl}x_{kl'},$$
$$\alpha_{kj} = \alpha_k \alpha_j,$$

where $\eta_l^{(k)}$ and $\eta_{l,l'}^{(k)}$ are the coefficients of the minor components x_{kl} and $x_{kl}x_{kl'}$, respectively. Thus, the major-minor model is given by:

$$\begin{aligned}
g_1(c_1,...,c_q) &= \sum_{k=1}^{q}(\sum_{l=1}^{m_k}\eta_l^{(k)}x_{kl} + \sum_{1\le l<l'\le m_k}\eta_{ll'}^{(k)}x_{kl}x_{kl'})c_k \\
&+ \sum_{1\le k<j\le q}\big(\sum_{l=1}^{m_k}\eta_l^{(k)}x_{kl} + \sum_{1\le l<l'\le m_k}\eta_{ll'}^{(k)}x_{kl}x_{kl'}\big)\big(\sum_{l=1}^{m_j}\eta_l^{(j)}x_{jl} + \sum_{1\le l<l'\le m_j}\eta_{ll'}^{(j)}x_{jl}x_{jl'}\big)c_kc_j + \epsilon \\
&= \sum_{k=1}^{q}\sum_{l=1}^{m_k}\eta_l^{(k)}x_{kl}c_k + \sum_{k=1}^{q}\sum_{1\le l<l'\le m_k}\eta_{ll'}^{(k)}x_{kl}x_{kl'}c_k \\
&+ \sum_{1\le k<j\le q}\Bigg[(\sum_{l=1}^{m_k}\eta_l^{(k)}x_{kl})(\sum_{l=1}^{m_j}\eta_l^{(j)}x_{jl}) + (\sum_{l=1}^{m_k}\eta_l^{(k)}x_{kl})(\sum_{1\le l<l'\le m_j}\eta_{ll'}^{(j)}x_{jl}x_{jl'}) \\
&+ (\sum_{1\le l<l'\le m_k}\eta_{ll'}^{(k)}x_{kl}x_{kl'})(\sum_{l=1}^{m_j}\eta_l^{(j)}x_{jl}) + (\sum_{1\le l<l'\le m_k}\eta_{ll'}^{(k)}x_{kl}x_{kl'})(\sum_{1\le l<l'\le m_j}\eta_{ll'}^{(j)}x_{jl}x_{jl'})\Bigg]c_kc_j + \epsilon \\
&= \sum_{k=1}^{q}\sum_{j=1}^{m_k}\eta_l^{(k)}x_{kl}c_k + \sum_{k=1}^{q}\sum_{1\le l<l'\le m_k}\eta_{ll'}^{(k)}x_{kl}x_{kl'}c_k \\
&+ \sum_{1\le k<j\le q}\Big(\sum_{l=1}^{m_k}\sum_{l'=1}^{m_j}\eta_l^{(k)}\eta_{l'}^{(j)}x_{kl}x_{jl'} + \sum_{l=1}^{m_k}\sum_{1\le l*<l*'\le m_j}\eta_l^{(k)}\eta_{l*l*'}^{(j)}x_{kl}x_{jl*}x_{jl*'} \\
&+ \sum_{l=1}^{m_j}\sum_{1\le l*<l*'\le m_k}\eta_l^{(j)}\eta_{l*l*'}^{(k)}x_{jl}x_{kl*}x_{kl*'} + \sum_{1\le l<l'\le m_k}\sum_{1\le l*<l*'\le m_j}\eta_{ll'}^{(k)}\eta_{l*l*'}^{(j)}x_{kl}x_{kl'}x_{jl*}x_{jl*'}\Big)c_kc_j + \epsilon,
\end{aligned}$$

which is in the form of additive models. ■

A.2 Algorithm of Generating Maximin Distance Designs for Major Components

Algorithm 2 is to generate the maximin distance design for major components in MoM experiments.

Algorithm 2 The maximin distance design for major components

Input: lower and upper bounds on the major components
2: Partition the design space into small elements with predefined precision
Combine the Simplex-centroid design points and additional random sampled points from the partition pool as the initial design points
4: Calculate the minimum distance, d_{old}, of all pairs in the initial design
for 1:T **do**
6: Sample one point out of the initial design as the old point
Sample one point out of the partition pool to form the new design by replacing the old point in the last design
8: Calculate the minimum distance, d_{new}, of all pairs in the new design
if $d_{new} > d_{old}$ **then** update d_{old} with d_{new}
10: **else** do not form the new design with the new point in line 7
end if
12: **end for**
Output: the maximin distance design

A.3 More Details in Simulation

This section presents the details of simulation case (b), where there are three major components, c_1, c_2 and c_3, and the minor components nested under each major components are x_{11}, x_{12}, x_{13}, and x_{21}, x_{22}. Note that the major component c_3 has a single component.

There are five underlying models to be considered for generating the data:

$$\begin{aligned}
I: \ y =& 10c_1 + 30c_2 + 20c_3 + 18c_1c_2 + \epsilon, \\
II: \ y =& 15c_1x_{11} + 12.5c_1x_{12} + 15c_1x_{13} + 22.5c_2x_{21} + 20c_2x_{22} + \epsilon, \\
III: \ y =& 10c_1 + 30c_2 + 20c_3 + 15c_1^h x_{11} + 27.5c_2^h x_{21} + \epsilon, \text{ where } h = 0.5, \\
IV: \ y =& 10c_1 + 30c_2 + 20c_3 + 25c_2x_{11} + 22.5c_3x_{21} + \epsilon, \\
V: \ y =& 10c_1 + 30c_2 + 20c_3 + 7c_2c_3 + 13.75c_1^2x_{11}x_{12} + \epsilon,
\end{aligned}$$

where the major and minor components are independent of ϵ, $\epsilon \sim N(0, \sigma^2)$ and σ^2 is chosen such that the signal-to-noise (SN) ratio is three.

Same as case (a), the compared method includes the multiple-Scheffé model

$$y = (\alpha_1c_1 + \alpha_2c_2 + \alpha_3c_3) \times (\beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{13}x_{13}) \times (\beta_{21}x_{21} + \beta_{22}x_{22}) + \epsilon,$$

The major-only linear Scheffé model and the major-only quadratic Scheffé model are ex-

pressed respectively as

$$y = \gamma_1 c_1 + \gamma_2 c_2 + \gamma_3 c_3 + \epsilon,$$
$$y = \gamma_1 c_1 + \gamma_2 c_2 + \gamma_3 c_3 + \gamma_4 c_1 c_2 + \gamma_5 c_1 c_3 + \gamma_6 c_2 c_3 + \epsilon.$$

The 1st-order major-minor model and the 2nd-order major-minor model are expressed respectively as

$$y = (\gamma_1 + \gamma_2 x_{11} + \gamma_3 x_{12}) c_1 + (\gamma_4 + \gamma_5 x_{21}) c_2 + \gamma_6 c_3 + \epsilon$$
$$\begin{aligned} y = {} & \gamma_1 c_1 + \gamma_2 c_2 + \gamma_3 c_3 + \gamma_4 x_{11} c_1 + \gamma_5 x_{12} c_1 + \gamma_6 x_{21} c_2 + \gamma_7 c_1 c_2 + \gamma_8 c_1 c_3 + \gamma_9 c_2 c_3 \\ & + \gamma_{10} x_{11} c_1 c_2 + \gamma_{11} x_{12} c_1 c_2 + \gamma_{12} x_{21} c_1 c_2 + \gamma_{13} x_{11} c_1 c_3 + \gamma_{14} x_{12} c_1 c_3 + \gamma_{15} x_{21} c_2 c_3 \\ & + \gamma_{16} x_{11} x_{21} c_1 c_2 + \gamma_{17} x_{12} x_{21} c_1 c_2 + \epsilon. \end{aligned}$$

Same as case (a), the metrics to evaluate the model performance are R^2, AICc, MSE, MSCV, MSCVnorm and model size.

For both the unconstrained and constrained major components, we consider two different designs: the I-optimal design and the maximin distance design. For the minor components x_{11}, x_{12}, x_{13}, we choose the three-component simplex-centroid design assuming seven design points. For the minor components x_{21}, x_{22}, we choose three design points: the two end points and the middle point in the domain, i.e., (1,0), (0.5,0.5), and (0,1). Same as case (a), we applied the idea of crossed design to combine the designs for the major and minor components.

A.4 Unconstrained MoM Experiments

Tables S1 and S2 show the simulation results in terms of R^2, MSE, MSCV, MSCVnorm, AICc, and model size among different models in the three-component simplex-centroid design and in the maximin distance design for the unconstrained MoM experiments based on 50 simulation replications. The proposed AHM generally outperform the other models in prediction regarding MSCV and MSCVnorm, and in fitting regrading AICc in all simulation models but IV. For the simulation model I, which only contains the major components, the AHM has comparable prediction performance with the MajorQuad model. For the simulation models II and III, the AHM has competitive prediction and fitting performance compared with the 1st_MM and the 2nd_MM. For the simulation model IV, the AHM as well as the 1st_MM, 2nd_MM, has similar prediction and fitting performance, but worse than the multiple Scheffé model. For the simulation model V, the prediction performance of AHM is best and close to that of the true model.

In terms of model fitting, the measure R^2, AICc, and MSE values in the tables indicate that the AHM has good fitting performance when it has competitive prediction performance.

The model size of the AHM varies across different settings because of the variable selection performed via the nonnegative garrote method. We also observe that the model size of the AHM is often larger than that of 1st_MM but smaller than that of 2nd_MM.

Table S1: Performance comparisons of models under the unconstrained MoM experiment using I-optimal design for major components from 50 simulation replications (means and standard errors (in parenthesis)).

	Model	R^2	MSE	MSCV	MSCVnorm	AICc	Size
I	1st_MM	0.970 (0.002)	15.16 (1.03)	15.88 (1.07)	0.31 (0.03)	408.0 (10.2)	6.0 (0.0)
	2nd_MM	0.977 (0.001)	12.71 (0.52)	14.62 (0.71)	0.29 (0.02)	396.8 (6.1)	17.0 (0.0)
	AHM	0.976 (0.001)	12.49 (0.45)	13.33 (0.42)	**0.26** (0.02)	**383.5** (5.1)	9.1 (1.5)
	MajorLinear	0.970 (0.002)	15.09 (1.00)	15.39 (1.02)	0.30 (0.02)	403.9 (10.0)	3.0 (0.0)
	MajorQuad	0.975 (0.001)	12.81 (0.28)	13.35 (0.29)	**0.26** (0.02)	**383.5** (3.2)	6.0 (0.0)
	MultipleScheffe	0.972 (0.002)	15.37 (1.25)	17.97 (1.52)	0.35 (0.04)	426.0 (12.0)	18.0 (0.0)
	TrueModel	0.974 (0.001)	12.81 (0.20)	13.16 (0.21)	0.26 (0.02)	381.2 (2.3)	4.0 (0.0)
II	1st_MM	0.974 (0.001)	29.20 (1.01)	30.44 (1.05)	**0.26** (0.02)	**504.6** (5.3)	6.0 (0.0)
	2nd_MM	0.976 (0.001)	29.21 (1.32)	33.64 (1.86)	0.29 (0.03)	519.2 (6.9)	17.0 (0.0)
	AHM	0.975 (0.001)	28.74 (1.22)	30.49 (1.20)	**0.26** (0.02)	**505.0** (6.0)	8.3 (1.4)
	MajorLinear	0.963 (0.003)	40.45 (2.91)	41.41 (2.99)	0.36 (0.03)	548.8 (10.8)	3.0 (0.0)
	MajorQuad	0.963 (0.003)	40.76 (2.93)	42.59 (3.08)	0.37 (0.04)	553.3 (10.8)	6.0 (0.0)
	MultipleScheffe	0.976 (0.002)	29.14 (1.66)	33.85 (2.29)	0.29 (0.03)	520.2 (8.6)	18.0 (0.0)
	TrueModel	0.974 (0.001)	29.20 (1.01)	30.44 (1.05)	0.26 (0.02)	504.6 (5.3)	6.0 (0.0)
III	1st_MM	0.919 (0.005)	156.51 (8.51)	163.65 (8.92)	0.31 (0.02)	751.3 (7.9)	6.0 (0.0)
	2nd_MM	0.935 (0.004)	135.38 (5.49)	155.02 (7.21)	0.29 (0.03)	744.6 (6.1)	17.0 (0.0)
	AHM	0.932 (0.004)	133.22 (4.41)	141.56 (4.40)	**0.26** (0.02)	**730.9** (4.5)	8.6 (1.8)
	MajorLinear	0.812 (0.013)	353.27 (27.13)	360.66 (27.73)	0.67 (0.04)	867.3 (11.1)	3.0 (0.0)
	MajorQuad	0.821 (0.013)	345.10 (26.82)	359.94 (27.93)	0.67 (0.04)	867.3 (11.3)	6.0 (0.0)
	MultipleScheffe	0.925 (0.005)	156.80 (9.34)	182.86 (12.87)	0.34 (0.03)	767.6 (8.7)	18.0 (0.0)
	TrueModel	0.929 (0.003)	134.77 (2.22)	139.50 (2.35)	0.26 (0.02)	728.3 (2.5)	5.0 (0.0)
IV	1st_MM	0.843 (0.012)	240.36 (19.58)	251.37 (20.74)	0.65 (0.04)	814.1 (12.0)	6.0 (0.0)
	2nd_MM	0.864 (0.011)	226.50 (19.60)	248.01 (21.56)	0.64 (0.04)	819.9 (12.7)	17.0 (0.0)
	AHM	0.847 (0.012)	234.92 (18.53)	246.98 (19.44)	0.64 (0.04)	811.0 (11.5)	6.2 (1.3)
	MajorLinear	0.837 (0.012)	245.11 (19.93)	251.54 (20.52)	0.65 (0.04)	813.5 (12.0)	3.0 (0.0)
	MajorQuad	0.838 (0.012)	248.44 (20.10)	260.00 (21.12)	0.68 (0.04)	818.9 (11.9)	6.0 (0.0)
	MultipleScheffe	0.942 (0.004)	97.02 (4.02)	111.83 (5.98)	**0.29** (0.02)	**697.2** (6.2)	18.0 (0.0)
	TrueModel	0.936 (0.003)	97.27 (1.61)	100.73 (1.72)	0.26 (0.02)	680.4 (2.5)	5.0 (0.0)
V	1st_MM	0.970 (0.002)	15.33 (1.05)	16.23 (1.18)	0.36 (0.03)	409.6 (10.2)	6.0 (0.0)
	2nd_MM	0.973 (0.002)	15.08 (1.24)	17.51 (1.58)	0.39 (0.04)	421.6 (12.4)	17.0 (0.0)
	AHM	0.978 (0.001)	11.32 (0.45)	12.16 (0.45)	**0.27** (0.02)	**370.8** (5.5)	10.5 (1.1)
	MajorLinear	0.968 (0.002)	15.85 (1.09)	16.26 (1.13)	0.36 (0.03)	411.0 (10.4)	3.0 (0.0)
	MajorQuad	0.970 (0.002)	15.48 (1.13)	16.19 (1.19)	0.36 (0.03)	411.0 (10.9)	6.0 (0.0)
	MultipleScheffe	0.972 (0.002)	15.70 (1.20)	18.68 (1.67)	0.41 (0.04)	429.1 (11.5)	18.0 (0.0)
	TrueModel	0.977 (0.001)	11.42 (0.20)	11.86 (0.23)	0.26 (0.02)	365.5 (2.7)	5.0 (0.0)

Table S2: Performance comparisons of models under the unconstrained MoM experiment using the maximin distance design for major components from 50 simulation replications (means and standard errors (in parenthesis)).

	Model	R^2	MSE	MSCV	MSCVnorm	AICc	Size
I	1st_MM	0.973 (0.004)	13.56 (1.96)	14.12 (2.05)	0.30 (0.03)	444.4 (28.2)	6.0 (0.0)
	2nd_MM	0.978 (0.003)	11.85 (1.52)	13.37 (1.78)	0.28 (0.02)	436.4 (24.8)	17.0 (0.0)
	AHM	0.977 (0.003)	11.62 (1.45)	12.24 (1.53)	**0.26** (0.02)	**422.2** (24.4)	8.5 (1.8)
	MajorLinear	0.973 (0.004)	13.54 (1.92)	13.78 (1.95)	0.29 (0.02)	441.0 (27.6)	3.0 (0.0)
	MajorQuad	0.977 (0.003)	11.82 (1.46)	12.26 (1.51)	**0.26** (0.02)	**422.1** (23.9)	6.0 (0.0)
	MultipleScheffe	0.975 (0.004)	13.70 (2.02)	15.66 (2.32)	0.33 (0.03)	461.6 (28.5)	18.0 (0.0)
	TrueModel	0.976 (0.003)	11.82 (1.46)	12.10 (1.49)	0.26 (0.02)	419.7 (24.0)	4.0 (0.0)
II	1st_MM	0.975 (0.003)	26.98 (3.88)	28.02 (4.01)	**0.26** (0.02)	**560.0** (28.8)	6.0 (0.0)
	2nd_MM	0.977 (0.003)	26.82 (3.88)	30.28 (4.40)	0.28 (0.02)	573.1 (28.7)	17.0 (0.0)
	AHM	0.976 (0.003)	26.42 (3.77)	27.95 (3.96)	**0.26** (0.02)	**559.5** (28.2)	8.5 (1.4)
	MajorLinear	0.965 (0.004)	37.36 (5.56)	38.10 (5.67)	0.36 (0.03)	611.4 (28.1)	3.0 (0.0)
	MajorQuad	0.966 (0.004)	37.56 (5.58)	39.01 (5.80)	0.36 (0.03)	615.7 (28.0)	6.0 (0.0)
	MultipleScheffe	0.977 (0.003)	26.94 (4.15)	30.66 (4.76)	0.29 (0.03)	575.0 (30.7)	18.0 (0.0)
	TrueModel	0.975 (0.003)	26.98 (3.88)	28.02 (4.01)	0.26 (0.02)	560.0 (28.8)	6.0 (0.0)
III	1st_MM	0.924 (0.006)	146.76 (13.66)	152.57 (14.40)	0.30 (0.02)	845.9 (16.9)	6.0 (0.0)
	2nd_MM	0.939 (0.004)	127.48 (10.89)	143.58 (13.46)	0.29 (0.02)	836.5 (15.0)	17.0 (0.0)
	AHM	0.938 (0.005)	123.05 (10.36)	130.15 (10.68)	**0.26** (0.02)	**820.1** (14.7)	9.0 (1.7)
	MajorLinear	0.820 (0.013)	342.53 (36.48)	348.79 (37.19)	0.69 (0.05)	984.7 (18.9)	3.0 (0.0)
	MajorQuad	0.827 (0.012)	335.72 (35.09)	348.09 (36.24)	0.69 (0.05)	984.8 (18.5)	6.0 (0.0)
	MultipleScheffe	0.931 (0.006)	144.33 (14.54)	166.16 (17.85)	0.33 (0.03)	858.5 (18.6)	18.0 (0.0)
	TrueModel	0.935 (0.004)	124.96 (9.33)	128.80 (9.64)	0.26 (0.02)	818.0 (13.4)	5.0 (0.0)
IV	1st_MM	0.847 (0.015)	230.23 (27.33)	238.97 (28.37)	0.64 (0.05)	921.1 (21.0)	6.0 (0.0)
	2nd_MM	0.872 (0.015)	206.48 (26.68)	224.57 (28.71)	0.60 (0.06)	916.6 (23.3)	17.0 (0.0)
	AHM	0.867 (0.017)	202.32 (29.78)	214.45 (31.62)	0.58 (0.06)	901.2 (27.4)	8.2 (1.4)
	MajorLinear	0.838 (0.015)	238.72 (24.75)	243.95 (25.32)	0.66 (0.04)	924.2 (17.8)	3.0 (0.0)
	MajorQuad	0.839 (0.015)	241.48 (25.25)	251.34 (26.30)	0.68 (0.04)	929.4 (18.0)	6.0 (0.0)
	MultipleScheffe	0.943 (0.005)	92.76 (7.88)	104.97 (9.40)	**0.28** (0.03)	**784.5** (15.1)	18.0 (0.0)
	TrueModel	0.938 (0.005)	92.14 (6.96)	94.88 (7.12)	0.26 (0.02)	766.9 (13.4)	5.0 (0.0)
V	1st_MM	0.971 (0.004)	14.30 (1.89)	15.00 (2.02)	0.36 (0.03)	453.9 (24.7)	6.0 (0.0)
	2nd_MM	0.974 (0.004)	14.02 (1.91)	16.02 (2.29)	0.38 (0.04)	464.5 (25.4)	17.0 (0.0)
	AHM	0.979 (0.002)	10.59 (1.15)	11.26 (1.26)	**0.27** (0.02)	**409.3** (20.2)	10.3 (1.1)
	MajorLinear	0.970 (0.004)	14.66 (1.90)	14.98 (1.95)	0.35 (0.03)	454.7 (24.6)	3.0 (0.0)
	MajorQuad	0.971 (0.004)	14.27 (1.83)	14.85 (1.91)	0.35 (0.03)	453.6 (24.7)	6.0 (0.0)
	MultipleScheffe	0.973 (0.004)	14.54 (2.00)	16.78 (2.40)	0.40 (0.04)	472.1 (25.5)	18.0 (0.0)
	TrueModel	0.979 (0.002)	10.63 (1.10)	10.97 (1.16)	0.26 (0.02)	403.4 (19.4)	5.0 (0.0)

A.5 Constrained MoM Experiments

We consider simulations where lower and upper bounds are placed on the major components. Specifically, we assume that the major and minor components satisfy the following constraints:

$$c_1 + c_2 + c_3 = 1, \quad 0.2 \leq c_1 \leq 0.45,$$
$$0.4 \leq c_2 \leq 0.6, \quad 0.1 \leq c_3 \leq 0.25,$$
$$x_{11} + x_{12} + x_{13} = 1, \quad x_{21} + x_{22} = 1.$$

The comparison results from using the I-optimal design and the maximin distance design for the major components are reported in Table S3 and Table S4, respectively. From both tables we can learn that the AHM has competitive prediction performance among all models. We also note that the 1st_MM, the 2nd_MM, and the multiple Scheff*é* model have relatively good prediction performance in simulation models I, II, III, and IV. Note that V favors the AHM.

It is worth noting that in IV, the AHM has comparable prediction performance as the true model, which is different from the results for the unconstrained MoM experiments. This observation is likely due to the difference between the two scenarios in terms of the constraints of the design space. The more highly constrained design space, the less negative effects of model misspecification on the model's prediction performance. The AHM has similar R^2, AICc and MSE values as the 1st_MM and the 2nd_MM in all simulation models but V. The model size of the AHM is larger than that of the 1st_MM but smaller than that of the 2nd_MM.

Table S3: Performance comparisons of models under the constrained MoM experiment using I-optimal design for major components from 50 simulation replications (means and standard errors (in parenthesis)).

	Model	R^2	MSE	MSCV	MSCVnorm	AICc	Size
I	1st_MM	0.999 (0.000)	0.48 (0.01)	0.50 (0.01)	0.27 (0.02)	-115.6 (3.5)	6.0 (0.0)
	2nd_MM	0.999 (0.000)	0.47 (0.01)	0.52 (0.02)	0.28 (0.02)	-104.9 (5.3)	17.0 (0.0)
	AHM	0.999 (0.000)	0.47 (0.01)	0.49 (0.01)	**0.26** (0.02)	-116.5 (4.1)	8.6 (1.7)
	MajorLinear	0.999 (0.000)	0.48 (0.01)	0.49 (0.01)	**0.26** (0.02)	**-119.2** (2.7)	3.0 (0.0)
	MajorQuad	0.999 (0.000)	0.47 (0.01)	0.49 (0.01)	**0.26** (0.02)	-118.4 (2.4)	6.0 (0.0)
	MultipleScheffe	0.999 (0.000)	0.48 (0.02)	0.54 (0.02)	0.29 (0.02)	-100.5 (6.2)	18.0 (0.0)
	TrueModel	0.999 (0.000)	0.47 (0.01)	0.48 (0.01)	0.26 (0.02)	-120.7 (1.9)	4.0 (0.0)
II	1st_MM	0.997 (0.000)	4.54 (0.08)	4.70 (0.09)	**0.26** (0.02)	**262.6** (3.1)	6.0 (0.0)
	2nd_MM	0.997 (0.000)	4.53 (0.14)	5.03 (0.17)	0.28 (0.02)	276.4 (5.1)	17.0 (0.0)
	AHM	0.997 (0.000)	4.49 (0.13)	4.73 (0.14)	**0.26** (0.02)	264.8 (5.0)	9.3 (1.1)
	MajorLinear	0.988 (0.001)	15.60 (1.09)	15.88 (1.11)	0.87 (0.03)	466.3 (12.0)	3.0 (0.0)
	MajorQuad	0.988 (0.001)	15.81 (1.11)	16.39 (1.16)	0.90 (0.03)	471.9 (12.1)	6.0 (0.0)
	MultipleScheffe	0.997 (0.000)	4.54 (0.16)	5.10 (0.22)	0.28 (0.02)	278.3 (6.0)	18.0 (0.0)
	TrueModel	0.997 (0.000)	4.54 (0.08)	4.70 (0.09)	0.26 (0.02)	262.6 (3.1)	6.0 (0.0)
III	1st_MM	0.962 (0.002)	101.23 (2.64)	105.11 (2.70)	**0.26** (0.02)	784.3 (4.4)	6.0 (0.0)
	2nd_MM	0.966 (0.002)	98.98 (3.37)	110.71 (4.16)	0.28 (0.02)	794.5 (5.8)	17.0 (0.0)
	AHM	0.964 (0.002)	98.13 (2.41)	102.58 (2.65)	**0.26** (0.02)	**782.2** (4.1)	8.6 (1.2)
	MajorLinear	0.850 (0.007)	397.23 (21.66)	404.41 (22.01)	1.01 (0.01)	1010.4 (9.3)	3.0 (0.0)
	MajorQuad	0.851 (0.007)	402.90 (21.84)	417.71 (22.58)	1.05 (0.01)	1016.1 (9.2)	6.0 (0.0)
	MultipleScheffe	0.966 (0.002)	98.97 (3.08)	112.07 (4.40)	0.28 (0.02)	796.0 (5.3)	18.0 (0.0)
	TrueModel	0.963 (0.001)	98.77 (1.40)	101.85 (1.39)	0.26 (0.02)	779.0 (2.4)	5.0 (0.0)
IV	1st_MM	0.968 (0.002)	43.55 (2.40)	45.39 (2.58)	0.36 (0.03)	642.3 (9.3)	6.0 (0.0)
	2nd_MM	0.977 (0.001)	32.80 (1.25)	36.66 (1.44)	**0.29** (0.02)	609.0 (6.4)	17.0 (0.0)
	AHM	0.975 (0.001)	34.97 (1.30)	36.66 (1.48)	**0.29** (0.02)	609.9 (6.6)	9.6 (1.4)
	MajorLinear	0.909 (0.005)	120.72 (8.15)	122.93 (8.31)	0.96 (0.02)	810.1 (11.3)	3.0 (0.0)
	MajorQuad	0.909 (0.005)	122.29 (8.31)	126.82 (8.64)	0.99 (0.02)	815.7 (11.4)	6.0 (0.0)
	MultipleScheffe	0.978 (0.001)	32.37 (1.24)	36.63 (1.60)	**0.29** (0.02)	**608.2** (6.5)	18.0 (0.0)
	TrueModel	0.976 (0.001)	32.37 (0.49)	33.37 (0.52)	0.26 (0.02)	591.6 (2.6)	5.0 (0.0)
V	1st_MM	0.998 (0.000)	1.08 (0.09)	1.13 (0.09)	0.37 (0.03)	21.6 (14.2)	6.0 (0.0)
	2nd_MM	0.998 (0.000)	1.10 (0.09)	1.24 (0.11)	0.40 (0.03)	38.7 (13.8)	17.0 (0.0)
	AHM	0.999 (0.000)	0.76 (0.02)	0.80 (0.02)	**0.26** (0.02)	**-32.0** (4.2)	9.9 (1.4)
	MajorLinear	0.998 (0.000)	1.12 (0.10)	1.14 (0.10)	0.37 (0.03)	24.1 (14.8)	3.0 (0.0)
	MajorQuad	0.998 (0.000)	1.13 (0.10)	1.17 (0.10)	0.38 (0.03)	28.2 (14.7)	6.0 (0.0)
	MultipleScheffe	0.998 (0.000)	1.11 (0.09)	1.25 (0.11)	0.41 (0.04)	40.6 (14.5)	18.0 (0.0)
	TrueModel	0.999 (0.000)	0.77 (0.01)	0.79 (0.01)	0.26 (0.02)	-37.3 (2.4)	5.0 (0.0)

Table S4: Performance comparisons of models under the constrained MoM experiment using the maximin distance design for major components from 50 simulation replications (means and standard errors (in parenthesis)).

	Model	R^2	MSE	MSCV	MSCVnorm	AICc	Size
I	1st_MM	0.999 (0.000)	0.44 (0.04)	0.45 (0.04)	0.26 (0.02)	-130.9 (15.1)	6.0 (0.0)
	2nd_MM	0.999 (0.000)	0.43 (0.04)	0.48 (0.05)	0.28 (0.02)	-118.9 (16.1)	17.0 (0.0)
	AHM	0.999 (0.000)	0.43 (0.04)	0.45 (0.04)	0.26 (0.02)	-131.9 (15.3)	8.2 (1.7)
	MajorLinear	0.999 (0.000)	0.44 (0.04)	0.45 (0.04)	0.26 (0.02)	-134.4 (15.0)	3.0 (0.0)
	MajorQuad	0.999 (0.000)	0.43 (0.04)	0.45 (0.04)	0.26 (0.02)	-133.2 (15.0)	6.0 (0.0)
	MultipleScheffe	0.999 (0.000)	0.44 (0.04)	0.50 (0.05)	0.29 (0.02)	-113.7 (16.9)	18.5 (0.5)
	TrueModel	0.999 (0.000)	0.43 (0.04)	0.44 (0.04)	0.25 (0.02)	-136.0 (15.0)	4.0 (0.0)
II	1st_MM	0.997 (0.000)	4.41 (0.15)	4.58 (0.15)	0.26 (0.02)	257.9 (5.7)	6.0 (0.0)
	2nd_MM	0.997 (0.000)	4.42 (0.22)	4.93 (0.24)	0.28 (0.02)	272.0 (8.4)	17.0 (0.0)
	AHM	0.997 (0.000)	4.38 (0.16)	4.62 (0.19)	0.26 (0.02)	260.8 (6.6)	9.5 (1.1)
	MajorLinear	0.988 (0.001)	15.11 (1.18)	15.39 (1.20)	0.86 (0.03)	460.9 (13.1)	3.0 (0.0)
	MajorQuad	0.988 (0.001)	15.31 (1.18)	15.88 (1.23)	0.89 (0.03)	466.5 (13.0)	6.0 (0.0)
	MultipleScheffe	0.997 (0.000)	4.41 (0.24)	5.01 (0.29)	0.28 (0.03)	274.1 (9.2)	18.5 (0.5)
	TrueModel	0.997 (0.000)	4.41 (0.15)	4.58 (0.15)	0.26 (0.02)	257.9 (5.7)	6.0 (0.0)
III	1st_MM	0.963 (0.002)	98.75 (2.76)	102.44 (2.92)	0.26 (0.02)	780.1 (4.7)	6.0 (0.0)
	2nd_MM	0.966 (0.002)	97.17 (3.48)	108.43 (4.03)	0.28 (0.02)	791.4 (6.1)	17.0 (0.0)
	AHM	0.964 (0.002)	96.77 (2.59)	101.17 (2.72)	0.26 (0.02)	780.3 (4.4)	9.1 (1.6)
	MajorLinear	0.849 (0.011)	394.27 (27.47)	401.45 (27.91)	1.02 (0.01)	1009.0 (11.6)	3.0 (0.0)
	MajorQuad	0.850 (0.011)	399.98 (28.02)	414.81 (29.02)	1.05 (0.01)	1014.7 (11.7)	6.0 (0.0)
	MultipleScheffe	0.966 (0.002)	97.19 (3.69)	110.13 (4.27)	0.28 (0.02)	793.6 (6.5)	18.5 (0.5)
	TrueModel	0.963 (0.001)	97.19 (1.79)	100.14 (1.87)	0.25 (0.02)	776.3 (3.1)	5.0 (0.0)
IV	1st_MM	0.968 (0.003)	42.00 (3.82)	43.72 (4.02)	0.35 (0.03)	635.8 (15.6)	6.0 (0.0)
	2nd_MM	0.978 (0.001)	32.01 (1.31)	35.76 (1.68)	0.28 (0.02)	604.8 (7.0)	17.0 (0.0)
	AHM	0.975 (0.001)	34.38 (2.02)	36.00 (2.06)	0.29 (0.02)	607.2 (10.0)	9.9 (1.1)
	MajorLinear	0.908 (0.006)	120.19 (8.91)	122.42 (9.06)	0.97 (0.02)	809.3 (12.6)	3.0 (0.0)
	MajorQuad	0.908 (0.006)	121.82 (9.09)	126.36 (9.42)	1.00 (0.02)	814.9 (12.7)	6.0 (0.0)
	MultipleScheffe	0.978 (0.001)	31.40 (1.23)	35.60 (1.85)	0.28 (0.02)	603.8 (6.9)	18.5 (0.5)
	TrueModel	0.976 (0.001)	31.50 (0.75)	32.48 (0.77)	0.26 (0.02)	587.0 (4.1)	5.0 (0.0)
V	1st_MM	0.998 (0.000)	1.02 (0.09)	1.06 (0.09)	0.38 (0.03)	10.9 (14.5)	6.0 (0.0)
	2nd_MM	0.998 (0.000)	1.04 (0.10)	1.16 (0.11)	0.41 (0.04)	28.0 (15.9)	17.0 (0.0)
	AHM	0.999 (0.000)	0.69 (0.05)	0.72 (0.05)	0.26 (0.02)	-48.8 (11.6)	10.2 (1.4)
	MajorLinear	0.998 (0.000)	1.07 (0.10)	1.09 (0.10)	0.39 (0.03)	16.3 (15.5)	3.0 (0.0)
	MajorQuad	0.998 (0.000)	1.08 (0.10)	1.12 (0.11)	0.40 (0.04)	20.7 (15.8)	6.0 (0.0)
	MultipleScheffe	0.998 (0.000)	1.04 (0.10)	1.19 (0.11)	0.42 (0.04)	30.9 (16.2)	18.5 (0.5)
	TrueModel	0.999 (0.000)	0.69 (0.04)	0.72 (0.05)	0.25 (0.02)	-54.2 (10.9)	5.0 (0.0)

Appendix B. Dynamic Variable Selection for Generalized Linear Models

B.1 Brief Review of Alternating Direction Method of Multipliers (ADMM)

Let us work through the standard ADMM approach for solving the optimization problem of the following form

$$\underset{\theta \in R^p}{\text{minimize}}\ f(\theta) + g(H\theta) \tag{B.1}$$

where both $f(\cdot)$ and $g(\cdot)$ are convex functions, and H is a matrix of size $m \times p$. To solve via ADMM, we introduce a copy of θ, called z, to decouple the two functions in Equatioin B.1. Then, ADMM solves an equivalent formulation of the problem

$$\underset{\theta \in R^p, z \in R^m}{\text{minimize}}\ f(\theta) + g(z) \text{ subject to } H\theta - z = 0.$$

ADMM alternatively updates the primal variables (θ, z) and associated scaled dual variable $u = \frac{\lambda'}{\rho}$:

$$\begin{aligned} \theta^{k+1} &= \underset{\theta}{\text{argmin}}\Big(f(\theta) + \frac{\rho}{2}||H\theta - z^k + u^k||_2^2\Big), \\ z^{k+1} &= \underset{\theta}{\text{argmin}}\left(g(z) + \frac{\rho}{2}||H\theta^{k+1} - z + u^k||_2^2\right), \\ u^{k+1} &= u^k + H\theta^{k+1} - z^{k+1}, \end{aligned} \tag{B.2}$$

where $\rho > 0$ is known as the positive augmented Lagrangian parameter and λ' is the Lagrange multiplier. The update of the primal variable θ^{k+1} is based on the current estimate of z^k and u^k. The update of the primal variable z^{k+1} is based on the current estimate of θ^{k+1} and u^k, while the update of the dual variable u^{k+1} is based on the current estimate of the primal variable θ^{k+1} and z^{k+1}.

B.2 Brief Review of Alternating Direction Method of Multipliers (ADMM)

Let us work through the standard ADMM approach for solving the optimization problem of the following form

$$\underset{\theta \in R^p}{\text{minimize}}\ f(\theta) + g(H\theta) \tag{B.3}$$

where both $f(\cdot)$ and $g(\cdot)$ are convex functions, and H is a matrix of size $m \times p$. To solve via ADMM, we introduce a copy of θ, called z, to decouple the two functions in Equatioin (A1). Then, ADMM solves an equivalent formulation of the problem

$$\underset{\theta \in R^p, z \in R^m}{\text{minimize}}\ f(\theta) + g(z) \text{ subject to } H\theta - z = 0.$$

ADMM alternatively updates the primal variables (θ, z) and associated scaled dual variable $u = \frac{\lambda'}{\rho}$:

$$\begin{aligned} \theta^{k+1} &= \underset{\theta}{\text{argmin}}\Big(f(\theta) + \frac{\rho}{2}||H\theta - z^k + u^k||_2^2 \Big), \\ z^{k+1} &= \underset{\theta}{\text{argmin}}\left(g(z) + \frac{\rho}{2}||H\theta^{k+1} - z + u^k||_2^2 \right), \\ u^{k+1} &= u^k + H\theta^{k+1} - z^{k+1}, \end{aligned} \tag{B.4}$$

where $\rho > 0$ is known as the positive augmented Lagrangian parameter and λ' is the Lagrange multiplier. The update of the primal variable θ^{k+1} is based on the current estimate of z^k and u^k. The update of the primal variable z^{k+1} is based on the current estimate of θ^{k+1} and u^k, while the update of the dual variable u^{k+1} is based on the current estimate of the primal variable θ^{k+1} and z^{k+1}.

B.3 The Algorithm for the Varying Coefficient Model with Smoothing Splines

Algorithm 2 summarizes the developed computational algorithm for parameter estimation of the benchmark method VCM with smoothing splines. The algorithm is implemented in the R package SeqADMM and is available in Bitbucket (`https://bitbucket.org/vtshen/rpackages/src/master/`).

Algorithm 2 The modified general local scoring algorithm

Input: $\boldsymbol{X}$ and $\boldsymbol{y}$
Initialize $s_0 = log\frac{\bar{y}}{1-\bar{y}}$, $\boldsymbol{s}_1^{(0)} = \boldsymbol{s}_2^{(0)} = \cdots = \boldsymbol{s}_p^{(0)} = \mathbf{0}$.
for iteration k **do**
$\boldsymbol{\eta}^{(k)} = \boldsymbol{s}_0 + \sum_{l=1}^{p} \boldsymbol{s}_l^{(k)}(x_l)$; $\boldsymbol{p}^{(k)} = e^{\boldsymbol{\eta}^{(k)}}/(1 + e^{\boldsymbol{\eta}^{(k)}})$
$W^{(k)} = diag(\boldsymbol{p}^{(k)}(1 - \boldsymbol{p}^{(k)}))$
$\boldsymbol{z}^{(k)} = \boldsymbol{\eta}^{(k)} + W^{-1,(k)}(\boldsymbol{y} - \boldsymbol{p}^{(k)})$
(The weighted backfitting algorithm)
for iteration t **do**
for iteration j over 1:p **do**
if j == 1 **then**
$\boldsymbol{r}_j^{(t)} = \boldsymbol{z}^{(k)} - \boldsymbol{s}_0 - \sum_{l=2}^{p} \boldsymbol{s}_l^{(t)}(\boldsymbol{x}_l)$
else if j == p **then**
$\boldsymbol{r}_j^{(t)} = \boldsymbol{z}^{(k)} - \boldsymbol{s}_0 - \sum_{l=1}^{p-1} \boldsymbol{s}_l^{(t)}(\boldsymbol{x}_l)$
else
$\boldsymbol{r}_j^{(t)} = \boldsymbol{z}^{(k)} - \boldsymbol{s}_0 - \sum_{l=1}^{j-1} \boldsymbol{s}_l^{(t)}(\boldsymbol{x}_l) - \sum_{l=j+1}^{p} \boldsymbol{s}_l^{(t-1)}(\boldsymbol{x}_l)$
end if
(Regress $W^{1/2,(k)}\boldsymbol{r}_j^{(t)}$ on $W^{1/2,(k)}\boldsymbol{s}_j(\boldsymbol{x}_j)$, where $\boldsymbol{s}_j(\boldsymbol{x}_j) = f_{smoothing\ spline}(time)x_j = diag(\boldsymbol{x}_j)H\theta$)
$\hat{\theta} = (H^T diag^T(\boldsymbol{x}_j) W diag(\boldsymbol{x}_j) H + \lambda\Omega)^{-1} H^T diag^T(\boldsymbol{x}_j) W \boldsymbol{r}_j$
$\boldsymbol{s}_j^{(t)}(\boldsymbol{x}_j) = diag(\boldsymbol{x}_j)H(H^T diag^T(\boldsymbol{x}_j) W diag(\boldsymbol{x}_j) H + \lambda\Omega)^{-1} H^T diag^T(x_j) W \boldsymbol{r}_j$
end for
if $\frac{\sum_{j=1}^{p} W_{j,j}^{(k)} (s_j^{(t)} - s_j^{(t-1)})^2}{\sum_{j=1}^{p} W_{j,j}^{(k)} + \sum_{j=1}^{p} W_{j,j}^{(k)} (s_j^{(t-1)})^2} \leq$ tolerance **then**
$\boldsymbol{s}_j^{(k)} = \boldsymbol{s}_j^{(t)}, j = 1, \cdots, p$; break.
end if
end for
Compute $\boldsymbol{\eta}^{(k)} = \boldsymbol{s}_0 + \sum_{l=1}^{p} \boldsymbol{s}_l^{(k)}(x_l)$; $\boldsymbol{p}_{\text{fitted}}^{(k)} = e^{\boldsymbol{\eta}^{(k)}}/(1 + e^{\boldsymbol{\eta}^{(k)}})$
if convergence in the deviance **then** break.
end if
end for

B.4 Overview of the Smoothing Spline

Consider an illustration example where we have the response variable $\boldsymbol{y}$ and the single predictor variable $\boldsymbol{x}$, with a choice of $\lambda \geq 0$, the smoothing spline estimator, $\hat{f}_\lambda$, is estimated by

$$f_\lambda = argmin \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt.$$

To directly estimate $\hat{f}_\lambda$ is difficult, but it is shown (Green and Silverman 1993) that (1) the solution to the above optimization problem is a natural cubic spline with knots at the unique values of x, and (2) the space of the such splines is n-dimensional.

Given the natural spline basis functions $\{h_1, ..., h_n\}$, we can reformulate the smoothing spline as $f(x) = \sum_{j=1}^{n} h_j(x)\theta_j$. With the matrix $H_{n\times n}$ having entries as $h_j(x_i)$, we note that solving the problem (1) is equivalent to solve the problem

$$argmin_{\theta \in R^n} ||y - H\theta||_2^2 + \lambda \theta^T \Omega \theta,$$

where the penalty matrix Ω has entries $\Omega_{jk} = \int h_j''(x)h_k''(x)dx$. This is a generalized Ridge regression problem, and its solution is known as

$$\hat{\theta} = (H^T H + \lambda\Omega)^{-1} H^T Y.$$

Note that in the case of weighted least squared regression, the optimization problem is then re-expressed as

$$argmin_{\theta \in R^n} ||W^{1/2}y - W^{1/2}H\theta||_2^2 + \lambda \theta^T \Omega \theta,$$

and the corresponding solution is

$$\hat{\theta} = (H^T W H + \lambda\Omega)^{-1} H^T W Y.$$

One choice of the basis is the natural cubic spline with knots at the unique values of x, and the space of such splines is n-dimensional. However, the associated penalty matrix Ω is in general not diagonal. There are other choices of spline basis, for example, the Demmler-Reinsch basis, which are orthonormal and therefore the penalty matrix is diagonal. In this work, we use the Demmler-Reinsch basis and generate an approximation to the Demmler-Reinsch orthonormal bases for smoothing spline, by using the function basis.gen in the R package gamsel (Chouldechova and Hastie 2015). Specifically, we choose a value $k = 10$ for the basis matrix of size $n \times k$.

B.4.1 Detailed Derivations in Algorithm 2

When the loss function has a weight matrix $W^{1/2}$,

$$\begin{aligned}
Loss &= ||W^{1/2}r_j - W^{1/2}diag(x_j)H\theta||_2^2 + \lambda\theta^T\Omega\theta \\
&= (W^{1/2}r_j - W^{1/2}diag(x_j)H\theta)^T(W^{1/2}r_j - W^{1/2}diag(x_j)H\theta) + \lambda\theta^T\Omega\theta \\
&= r_j^T W r_j - 2r_j^T W diag(x_j)H\theta + \theta^T H^T diag^T(x_j)W diag(x_j)H\theta + \lambda\theta^T\Omega\theta, \\
\frac{\partial Loss}{\partial\theta} &= 0 = -r_j^T W diag(x_j)H + H^T diag^T(x_j)W diag(x_j)H\theta + \lambda\Omega\theta, \\
\hat{\theta} &= (H^T diag^T(x_j)W diag(x_j)H + \lambda\Omega)^{-1}H^T diag^T(x_j)W r_j, \\
W^{1/2}diag(x_j)H\hat{\theta} &= W^{1/2}diag(x_j)H(H^T diag^T(x_j)W diag(x_j)H + \lambda\Omega)^{-1}H^T diag^T(x_j)W r_j.
\end{aligned}$$

Appendix C. Structured Variable Selection from an Experimental Thinking Perspective

C.1 Detailed Derivation of Solution to the Problem with the L_1-norm Penalty and the L_0-norm Constraint

Consider the following problem with L_1-norm and the L_0-norm

$$\underset{\boldsymbol{\beta}\in R^p}{\text{minimize}}\ ||\boldsymbol{y}-X\boldsymbol{\beta}||_2^2+\lambda||\boldsymbol{\beta}||_1 \text{ subject to } ||\boldsymbol{\beta}||_0\leq k, \tag{C.1}$$

where $||\boldsymbol{\beta}||_0=\sum_{i=1}^{p}1\{\boldsymbol{\beta}_i\neq 0\}$ is the L_0-norm of the coefficient $\boldsymbol{\beta}$ with $1(\cdot)$ denoting the indicator function. The L_0-norm term, different from the L_1-norm penalty term, does not depend on the magnitudes of coefficients. The turning parameter λ adjusts the effect of L_1-norm penalty, $||\boldsymbol{\beta}||_1$. The discrete tuning parameter k is the number of chosen predictors ranged between 0 and min{n, p}.

With the augmented complete response $\boldsymbol{y}_c$ and the augmented complete orthogonalized regression matrix X_c, the problem C.1 is reformulated as:

$$\underset{\boldsymbol{\beta}\in R^p,\ \boldsymbol{y}^*\in R^{m-n}}{\text{minimize}}\ ||\boldsymbol{y}-X\boldsymbol{\beta}||_2^2+||\boldsymbol{y}^*-X^{\dagger}\boldsymbol{\beta}||_2^2+\lambda||\boldsymbol{\beta}||_1 \text{ subject to } ||\boldsymbol{\beta}||_0\leq k, \tag{C.2}$$

We apply the EM algorithm to solve the problem C.2. The complete log-likelihood function, similar to 4.1, can be written as

$$l_c(\boldsymbol{\beta};X_c,\boldsymbol{y}_c)=-m\log(\sqrt{2\pi}\sigma)-\frac{\boldsymbol{y}^T\boldsymbol{y}+\boldsymbol{y}^{*,T}\boldsymbol{y}^*-2(\boldsymbol{y}^TX+\boldsymbol{y}^{*,T}X^{\dagger})\boldsymbol{\beta}+\boldsymbol{\beta}^TX_c^TX_c\boldsymbol{\beta}+\lambda||\boldsymbol{\beta}||_1}{2\sigma^2},$$

which is subject to $||\boldsymbol{\beta}||_0\leq k$.

In the **E-step**, we replace the missing values in response, $\boldsymbol{y}^*$, with the conditional expectation

given the observed data, $\mathrm{E}(\boldsymbol{y}^*|X^\dagger;\ \boldsymbol{\beta})$. We have

$$Q(\boldsymbol{\beta}; X_c, \boldsymbol{y}_c, \boldsymbol{\beta}^{(t)}) = E[l_c(\boldsymbol{\beta}; X_c, \boldsymbol{y}_c, \boldsymbol{\beta}^{(t)})] = -m\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\Big[\boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^{T,(t)}X^{\dagger,T}X^\dagger\boldsymbol{\beta}^{(t)} + \sigma^2 - 2(\boldsymbol{y}^TX + \boldsymbol{\beta}^{T,(t)}X^{\dagger,T}X^\dagger)\boldsymbol{\beta} + \boldsymbol{\beta}^TD\boldsymbol{\beta} + \lambda||\boldsymbol{\beta}||_1\Big],$$

where $\boldsymbol{\beta}^{(t)}$ is computed from the last iteration and the conditional expectations are given by

$$\begin{aligned} E(\boldsymbol{y}^*|X^\dagger, \boldsymbol{\beta}) =& X^\dagger\boldsymbol{\beta}, \\ E(\boldsymbol{y}^{*,T}\boldsymbol{y}^*|X^\dagger, \boldsymbol{\beta}) =& \boldsymbol{\beta}^TX^{\dagger,T}X^\dagger\boldsymbol{\beta} + \sigma^2. \end{aligned}$$

In the following **M-step**, we maximize $E[l_c(\boldsymbol{\beta}; X_c, \boldsymbol{y}_c, \boldsymbol{\beta}^{(t)})]$ over $\boldsymbol{\beta}$. That is

$$\begin{aligned} &\underset{\boldsymbol{\beta}}{\text{maximize}} -\Big[\boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^{T,(t)}X^{\dagger,T}X^\dagger\boldsymbol{\beta}^{(t)} + \sigma^2 - 2(\boldsymbol{y}^TX + \boldsymbol{\beta}^{T,(t)}X^{\dagger,T}X^\dagger)\boldsymbol{\beta} + \boldsymbol{\beta}^TD\boldsymbol{\beta} + \lambda||\boldsymbol{\beta}||_1\Big], \\ &\text{subject to } ||\boldsymbol{\beta}||_0 \le k, \end{aligned}$$

which can be further reduced to

$$\begin{aligned} &\underset{\boldsymbol{\beta}}{\text{maximize}} -\boldsymbol{\beta}^TD\boldsymbol{\beta} + 2(X^T\boldsymbol{y} + X^{\dagger,T}X^\dagger\boldsymbol{\beta}^{(t)})^T\boldsymbol{\beta} - \lambda||\boldsymbol{\beta}||_1,\ \text{subject to } ||\boldsymbol{\beta}||_0 \le k, \\ \Leftrightarrow\ &\underset{\boldsymbol{\beta}}{\text{maximize}} \sum_{j=1}^{p}\big[-d_j\beta_j^2 + 2\mu_j^{(t)}\beta_j - \lambda|\beta_j|\big],\ \text{subject to } ||\boldsymbol{\beta}||_0 \le k, \end{aligned} \tag{C.3}$$

where $\mu_j^{(t)}$ is the j-th component in $\boldsymbol{\mu}^{(t)} = X_c^T\boldsymbol{y}_c = X^T\boldsymbol{y} + X^{\dagger,T}X^\dagger\boldsymbol{\beta}^{(t)}$.

In the problem C.3, the objective function term is decomposable in the dimension of $\boldsymbol{\beta}$. We obtain the maximum objective function value in the dimension j of $\boldsymbol{\beta}$, $\max(OF_j)$, by setting the first derivative to 0.

$$\frac{\partial OF_j}{\partial\beta_j} = 2d_j\beta_j - 2\mu_j^{(t)} + \lambda\frac{\partial|\beta_j|_1}{\partial\beta_j} = 0,$$

When $\beta_j > 0$, we have

$$\begin{aligned} 0 =& 2d_j\beta_j - 2\mu_j^{(t)} + \lambda, \\ \hat{\beta}_j =& \frac{\mu_j^{(t)}}{d_j} - \frac{\lambda}{2d_j}, \end{aligned}$$

where $\mu_j^{(t)} > \frac{\lambda}{2} > 0$.

When $\beta_j < 0$, we have

$$\begin{aligned} 0 =& 2d_j\beta_j - 2\mu_j^{(t)} - \lambda, \\ \hat{\beta}_j =& \frac{\mu_j^{(t)}}{d_j} + \frac{\lambda}{2d_j}. \end{aligned}$$

where $\mu_j^{(t)} < -\frac{\lambda}{2} < 0$.

When $\beta_j = 0$,

$$\begin{aligned} 0 =& -2\mu_j^{(t)} - \lambda[-1, 1], \\ \lambda >& |2\mu_j^{(t)}|. \end{aligned}$$

In summary, we have the j-th dimension of the solution $\boldsymbol{\beta}$, $\hat{\beta}_j$, as

$$\hat{\beta}_j = \begin{cases} \frac{\mu_j^{(t)}}{d_j} - \frac{\lambda}{2d_j} sign(\mu_j^{(t)}), & 0 < \lambda < |2\mu_j^{(t)}|, \\ 0, & \lambda > |2\mu_j^{(t)}| > 0. \end{cases}$$

We then have

$$\max(\mathrm{OF}_j) = \begin{cases} \frac{(2\mu_j^{(t)} - sign(\mu_j^{(t)})\lambda)^2}{4d_j}, & 0 < \lambda < |2\mu_j^{(t)}|, \\ 0, & \lambda > |2\mu_j^{(t)}| > 0. \end{cases}$$

Note that $\hat{\beta}_j$ and $\mu_j^{(t)}$ have the same sign.

We introduce the binary variable z_j, j-th component in $\boldsymbol{z}$, as $1\{\beta_j^{(t+1)} \neq 0\}$, then $||\boldsymbol{\beta}||_0 = \sum_{i=1}^{p} z_i$. In addition, the $\max(OF_j)$ can be viewed as $z_j \frac{(2\mu_j^{(t)} - sign(\mu_j^{(t)})\lambda)^2}{4d_j}$. Then the problem C.3 is equivalent to

$$\begin{aligned} \underset{\boldsymbol{z} \in R^p}{\text{maximize}} & \sum_{j=1}^{p} z_j \frac{(2\mu_j^{(t)} - sign(\mu_j^{(t)})\lambda)^2}{4d_j}, \\ s.t. & \sum_{j=1}^{p} z_j \leq k, z_j \in \{0, 1\}, j = 1, ..., p, \end{aligned} \tag{C.4}$$