2020

Approaches for Analyzing Multivariate Mixed Endpoints With High-Dimensional Covariates.

Yunusa Olufadi

APPROACHES FOR ANALYZING MULTIVARIATE MIXED ENDPOINTS WITH HIGH-DIMENSIONAL COVARIATES

by

Yunusa Olufadi

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Mathematical Sciences

The University of Memphis

May 2020

DEDICATION

My Mom, Alhaja Bilqees Olufadi

My Uncle, Dr. Rasaq Olufadi

My Dad, Late Alhaji Shuaib Laaro Olufadi

My Aunts, Late Alhaja Asiyah Titilayo and Risikat Babatunde

Grandparents, Late Alhaji AbdurRahman and Alhaja Rahmat Olufadi

ACKNOWLEDGMENTS

Completing a Ph.D. program is not a one man's job - it requires a tremendous amount of support from many individuals and institutions. While only my name appears on the cover of this dissertation work, a great many people, including my family and friends, well-wishers, colleagues, and various institutions, have contributed to accomplishing this colossal task. To this end, I take this opportunity to express my deepest gratitude and sincere appreciation to these incredible individuals and organizations.

First, I want to sincerely thank my advisor, Prof. E. Olusegun George, for his dedicated help, advice, constructive criticism, and positive appreciation, which led to the successful completion of this research work. I owe him lots of gratitude for his support and guidance. I will also like to especially appreciate the member of my thesis committee, Dr. Dale Bowman, Dr. Hongmei Zhang, and Dr. Fridtjof Thomas, for their brilliant comments, suggestions, and fruitful discussions.

I will be an ingrate if I fail to appreciate and acknowledge the financial support I received from the Tertiary Education Trust Fund (TETFUND), The Kwara State University, and The Department of Mathematical Sciences of The University of Memphis. This work may not have seen the light of the day without their support. My sincere appreciation goes to Prof. Alfred Olaiya Babatunde Soboyejo and his son (Prof. Wole Soboyejo) for their innumerable support, fatherly advice, and guidance since the time I started nursing the ambition of doing my Ph.D. program in the United States. I will forever remain indebted to you. I would also like to express my gratitude to my colleagues in the department John, Andrew, Kenneth, and Farnaz for making this journey a memorable one.

My mom (Alhaja Bilqees Olufadi) and siblings (Engr. Rasheed, Ibraheem, and Habeebat) have been the pillars that I rest on during my academic sojourn here in Memphis. I thank them immensely for their love and affection. I am so happy that God has answered your prayers! I do not have the right words to express my deepest gratitude to my wife and

children for their constant love, affection, and numerous support. May God bless you all. To the well-wishers, friends, and family too many to mention, I thank you immensely for your love and moral support.

ABSTRACT

Olufadi, Yunusa. PhD. The University of Memphis. May/2020. Approaches for Analyzing Multivariate Mixed Endpoints With High-Dimensional Covariates. Major Professor: E. Olusegun George, PhD

In clinical trials and observational studies, clinicians often observe measurements on multiple causes of clinical progression or synthesize information from various sources. These measurements are collected because a single outcome is usually inadequate to describe the disease complexities or because the primary outcomes of interest are abstract constructs (e.g., quality of life, disease conditions) that cannot be measured directly. It is usually necessary to collect multiple endpoints in order to fully understand the true associations that exist among several clinical outcomes and how they jointly affect the primary outcomes. In addition, such datasets are often useful for characterizing treatment effectiveness, evaluate the risk-factors, or investigate the impact of health policy initiatives. Examples of multivariate mixed outcomes data are ubiquitous in biomedical and bio-pharmaceutical studies, psychometric, behavioral research, and pre-clinical teratology and developmental toxicity studies, among others. The different data structures of endpoints present interesting statistical and computational challenges. For example, there would be several levels of correlations inherent in the outcomes data, especially when dealing with a clustered or longitudinal design. The common modeling strategy of analyzing each endpoint separately in a univariate manner usually leads to misleading findings because such an approach ignores the correlations and interactions among the outcomes. The introduction of high-dimensional covariates such as gene expressions and large dimensional clinical information further exacerbates the modeling and analysis (the $p \gg n$ problem), leading to a need for sophisticated variable selection strategy. While variable selection methods are well-developed for many statistical models, the procedure is underdeveloped for multivariate mixed endpoints. This dissertation is motivated by the statistical and computational challenges that arise from analyzing such data. The overarching goal of this dissertation is to develop statistical procedures for

jointly modeling, estimation, and efficient identification of significant predictors in the analysis of multivariate clustered/longitudinal mixed endpoints datasets that are characterized by high-dimensional covariates. Specifically, we develop a procedure to guide both the model estimation and the efficient extraction of potential active predictors. We demonstrate the advantages of our procedure in terms of variable selection, prediction, and computational scalability via extensive simulations study and apply the method to two real-life datasets. In addition to other properties, we find that the estimates identified by dynamic posterior exploration in our procedure stabilize rapidly and very early in their trajectories, especially in the implementation of the dynamic weighted LASSO.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In clinical trials (Oliveira & Teixeira-Pinto, 2015) and observational studies (Holmes et al., 1994), clinician often synthesize information from various sources or observed measurements on multiple causes of clinical progression for many reasons. Clinician/investigators collect these multivariate mixed measurements (e.g., discrete, ordinal, continuous, and nominal) because the primary outcomes of interest are an abstract construct (e.g., quality of life, Alzheimer's disease conditions, functional dependency, quality of care, cognitive level) that are hard to quantify, measured directly or expensive to measure, and a single outcome is not adequate to describe the disease complexities. In Alzheimer disease for example, many clinical and neuropathologic outcomes such as Functional Activities Questionnaire (FAQ), Clinical Dementia Rating (CDR), Montreal Cognitive Assessment (MoCA), Mini-Mental State Examination (MMSE), score derived from clinician judgment on neuropsychological tests (COG) etc., are collected on everyone under study to determine the disease condition. In psychiatric studies, several variables are measured as proxies of the underlying primary outcome of interest. For instance, in evaluating the effectiveness of a new anti-psychotic, clinician combine measurements such as symptoms of relapse, positive and negative syndrome scale (PNASS) score, quality of life and so on.

These multiple information are collected to understand the true associations that exists among these multiple endpoints and how they affect the primary outcome jointly. Other reasons might be to characterize the treatment effectiveness, evaluate the risk-factors, or investigate the impact of large policy initiatives. Examples of these multivariate mixed outcomes data are ubiquitous in biomedical and bio-pharmaceutical studies, psychometric, behavioral research, developmental toxicity study, and psychology among others. We present some of these examples in Table 1.

In addition to the multiple mixed responses, large number of variables that may help, say, predict disease, are also frequently collected. For example, in disease classification using microarray or proteomics data, tens of thousands of expressions of molecules or ions are potential predictors. Also, in genome-wide association studies between genotypes and phenotypes, hundreds of thousands of SNPs are potential covariates for phenotypes such as cholesterol levels or heights. The dimensionality of the data grows rapidly when interactions between predictors are also important. Such high-dimensional data are used to investigate questions such as which genes are potentially informative to predict the causes/pathway of disease.

The different data structures of endpoints that result from mixed outcomes design studies such as those described in Table 1 present interesting statistical and computational challenges. These includes:

1. Complex correlation structures. There are at least two levels of correlations inherent in such data set.

   (a) Statistical dependence between different endpoints on the same subject.

   (b) Correlation between repeated/longitudinal measurements.

   (c) Litter effects in teratology and developmental toxicity studies due to genetic similarity or shared maternal environment during gestation.

2. Further consideration are needed because the outcomes measurements on the same subject are of differing nature such as discrete, continuous, and nominal. We delay the discussion of some of the attempts that have been made in the past to handle these challenges to Section 1.2.

3. Further challenges are presented by the high dimensionality of covariates with relatively small sample sizes ($p \gg n$) leading to a need for dimension reduction for mean-

ingful estimating equations. Addressing the problem of high dimensionality is further complicated by multivariate mixed endpoints. While variable selection methods are well-developed for models such as linear regression (Efron et al., 2004; George & McCulloch, 1997; Park & Casella, 2008; Ročková & George, 2014), generalized linear models (Friedman et al., 2010), quantile regression (Alhamzawi & Ali, 2018; Alhamzawi & Yu, 2012; Li et al., 2010; Wu & Liu, 2009); graphical models (Deshpande et al., 2019; Gan et al., 2019), methodologies for variable selection to address multivariate mixed endpoints have received very rare statistical consideration in the literature. We devote Section 1.1 to the importance of the multivariate mixed endpoints high-dimensional data sets.

In Table 1, we present a tabulated summary of some of the studies that are relevant to the theme of this dissertation.

In each of the studies listed in Table 1, applications of the standard procedures are not adequate for analyzing such data set. The common modeling strategy of analyzing each endpoint separately ignores potential correlation among the outcomes which can lead to misleading conclusion. Joint modeling the multivariate mixed endpoints as opposed to the popular separate univariate analysis:

(a) provides a general framework to better describe the association among the outcomes. For instance, joint modeling can help elucidate the link between continuous progression of diseases through longitudinal outcomes such as biomarkers or more generally indicators of health, and the incidence of clinical events such as diagnosis, recurrence and death.

(b) offers the ability to answer fundamental multivariate question, for example, interest might be in assessing the impact of a policy change on the quality of care (the under-

Table 1. Some Examples of Multivariate Clustered and Longitudinal (Repeated-Measures) Mixed Endpoints Studies

	Studies	Study Goal	Discrete Endpoints	Continuous Endpoints	References
1	Anticonvulsant teratogenesis study	To assess the effect of in utero anticonvulsant exposure on a variety of birth outcomes	hypoplastic fingernails, tapered fingers, antiverted nostrils, hypoplastic toenails	birth weight, head diameter	Holmes et al. (1994)
2	Irwin's toxicity study	To determine treatment effects and association between some mixed outcomes	toe pinch, abnormal biting, restlessness, pinna reflex	temperature, pupil size grip strength, vocalization	Faes et al. (2008)
3	VHA performance monitoring study	To characterize trends in quality of individual service networks based on the mixed outcomes	visits, readmissions	days between visits, days between readmissions	Daniels and Normand (2006)
4	Restenosis study	to estimate restenosis for diabetic patients accounting for potential confounders	target lesion revascularization	proportion diameter stenosis	Teixeira-Pinto and Normand (2009)
5	Macular degeneration study	To investigate the relationship between the mixed endpoints	loss of at least three lines of vision	visual acuity difference	Teixeira-Pinto and Normand (2009)
6	Managed care & quality of care for schizophrenia	To compare care for patients who were & were not enrolled in managed care	atypical antipsychotic medication	self-reported interpersonal interaction	Dickey et al. (2003)
7	St Louis risk Research project	To determine effects of parental psychological disorders on children's development	number of adverse psychiatric symptoms found in a child	standardized reading score, standardized verbal comprehension score	Little and Schluchter (1985)
8	Harvard Six Cities Study	To determine the effect of maternal smoking on respiratory illness in children	AVF and FGS	UYH and HGF	Wang et al. (1994)
9	Restenosis after coronary stenting	To estimate the treatment effect and identify baseline risk factors predictive of outcome after the stenting procedure	target lesion revascularization, binary restenosis	Late lumen loss	Teixeira-Pinto and Mauri (2011)

lying outcome) rather than its impact on each outcome measured as a proxy of the quality of care.

(c) reduces the need for multiple testings that naturally leads to global tests that results in increased power and reduced false discovery rates.

(d) results in gains in efficiency as reported by Gueorguieva and Sanacora (2006), Teixeira-Pinto and Normand (2009) and others.

This theses also addresses the additional computational challenges that arise in the formu-

lation of methods in the context of mixed multivariate data. The procedure proposed here incorporate ways of handling the variable selection that result from high dimensionality of the data. The overarching goal of this dissertation is to address the following question: how do we jointly model, carry out estimation, and efficiently identify active potential predictors in a multivariate clustered/longitudinal mixed endpoints characterized by high-dimensional covariates?

1.1 Variable Selection

The collection of high-dimensional data has become increasingly common in diverse fields of sciences, engineering, and humanities, ranging from genomics and health sciences to economics, finance and machine learning. The development of appropriate statistical and computational methods to extract meaningful information from such data has become a universally important research preoccupation of statistician and computer scientists. While it is clear that not all the available variables can be included in the statistical model, it is not known *a priori* which variables should be included in the model. A crucial task is then to identify a sparse model that has better statistical interpretability and prediction accuracy. This task is usually accomplished through variable selection.

The essence of variable selection becomes even more apparent when the number of predictors is larger than the sample size (the so-called curse of dimensionality or NP-hard problem). The famous "Occam's razor" principle is quite apt when modeling high-dimensional variables. The inclusion of noise or irrelevant variables when modeling data makes it difficult to identify the true predictor variables that have an important influence on the response variable, leading to models that are not efficacious.

The earliest traditional procedures for variable selection include forward, backward or step-wise approaches. Among the drawbacks of these procedures are high computational cost for high-dimensional data and tendency to be stuck at local optimal points during stepwise

searching. These approaches can also induce biased estimates and non-valid p-values as well as nominal confidence levels. To overcome these issues, new advanced tools have been proposed, perhaps the most prominent frequentist of which is LASSO and its many variants (Tibshirani, 1994; Tibshirani et al., 2005; Zou, 2006). Another popular approach is through Bayes method (Buhlmann et al., 2010; Carvalho and Polson, 2010; Castillo et al., 2015; Castillo and van der Vaart, 2012; George and McCulloch, 1993, 1997; Huang et al., 2016; Mitchell and Beauchamp, 1988; Park and Casella, 2008; Polson and Scott, 2010; Zhang and Huang, 2008; Zhang et al., 2016a; Zhang et al., 2016b).

The literature on sparse estimation either from a frequentist perspective or Bayesian framework is vast. There are also some attempts to optimize variable selection and standardize it for any kind of data, however; there are no general variable selection methods for all statistical models. A comprehensive review of variable selection procedures is given by Bhadra et al. (2017) and van Erp et al. (2019). In the next section, we present a brief review of statistical modeling of mixed outcomes data.

1.2 Review of Existing Methods on Mixed Outcomes

The principal statistical challenge when analyzing mixed outcomes is the construction of the joint model. There have been several attempts in the past to address this issue, some have been non-Bayesian (e.g., Catalano and Ryan, 1992; Fitzmaurice and Laird, 1995; George et al., 2007; Geys et al., 1999; Gueorguieva and Agresti, 2001; Regan and Catalano, 1999) while others have adopted a Bayesian framework (e.g., Bowman and George, 2018; Das et al., 1999; Dunson, 2000; Dunson et al., 2003). We classify the existing attempts at modeling mixed outcomes into four groups.

1.2.1 Factoring Method

This method was first discussed by Tate (1954), later, Olkin and Tate (1961) introduced the general location model (GLOM). The main idea behind GLOM is to jointly factor the joint likelihood into a product of marginal distribution and conditional distribution, where the conditioning can be done either on the discrete or the continuous outcome. Fitzmaurice and Laird (1995) extended GLOM procedure to clustered data by proposing a model with logit link function for the marginal probability of binary response and a normal distribution for the continuous response given the binary response. Although Fitzmaurice and Laird (1995) has an attractive interpretational feature, the specification of the regression model conditional on the cluster-specific random effects is entirely unknown.

Cox and Wermuth (1992) compared different joint distribution models for analyzing data with continuous and discrete responses as a function of the covariates. Liu and Rubin (1998) extended the common covariance matrix to allow different, but proportional covariance matrices and replace the multivariate normal distribution specified for continuous variables by multivariate t distribution. George et al. (2007) made a novel contribution by assuming the exchangeability of joint bivariate outcomes within the litter. They factored the joint likelihood as the product of continuous response given the binary response and the marginal distribution of the binary responses to produce the parameter estimates. Other authors that used general location model include de Leon and Carrière (2007) and Fitzmaurice and Laird (1997).

A drawback of mixed outcome models based on factorization is that it may be challenging to implement it for quantitative risk assessment because there is no direct access to the marginal distributions (Geys et al., 2001). Also, factorization models do not easily extend to the setting of three or more outcomes and the correlation among the mixed responses itself cannot be directly estimated.

1.2.2 Latent Variable Method

The introduction of a continuous latent variable as an underlying mechanism for the generation of the discrete outcome is another method commonly used by Bayesian and non-Bayesian researchers when modeling mixed endpoints of different kinds. This approach was first introduced by Cox (1972) to model correlated binary data and extended to clustered mixed outcomes with missing data by Little and Schluchter (1985) and Little and Rubin (2002). Sammel et al. (1997) proposed a latent variable multivariate mixed effects model by assuming a latent variable, linearly linked to the observed covariates, for each subject under study. The distribution for each type of measurement given the latent variable is assumed to come from an exponential family model. These authors modeled the observed outcomes as functions of fixed covariates and subject-specific latent variable. A deficiency of this approach is that it is not robust to misspecification of the covariance because the mean parameters depend heavily on the covariance parameters. For example, if the outcomes are not correlated, the estimates of the covariance effects may be biased Sammel et al. (1999).

Catalano and Ryan (1992) noted that latent variable models provide a useful and intuitive way to motivate the distribution of discrete endpoints. In the bivariate case, a standard method assumes an unobservable normally-distributed random variable underlying the discrete outcomes, resulting in a probit-type model. One drawback of this method is that regression parameters for the binary response using the probit link function do not have an odds ratio interpretation. Alternatively, Geys et al. (2001) presented a model based on a Plackett-Dale approach, where a bivariate Plackett distribution is assumed. O'Malley et al. (2002) combined the general location model and the latent trait model for mixed outcomes. Several other studies such as Daniels and Normand (2006), Goldstein et al. (2009), Moustaki and Knott (2000) have employed the latent variable model when dealing with mixed endpoints data.

1.2.3 Pairwise Modeling Method

Many authors (Faes et al., 2008; Fieuws & Verbeke, 2006; Molenberghs & Verbeke, 2005) adopted the use of mixed models to directly specify the joint distribution of discrete and continuous outcomes instead of a latent variable or factorization approach. They did this by specifying the marginal distribution, conditioned on a correlated random effects. An advantage of the mixed model approach is that additional correlation structures in the data, such as the cluster effect or a longitudinal data structure, can be modeled within the same framework. However, it becomes difficult to implement this procedure as the number of outcomes increases. In fact, the higher the number of endpoints, the higher the dimension of the random-effects vector when modeling the correlation between the different outcomes via a random effect, and the more likely computational problems will arise during the estimation process.

A pairwise model-fitting procedure was presented by Fieuws and Verbeke (2006) to circumvent the computational complexities in the setting of many continuous outcomes, replacing the maximization of the full likelihood distribution by maximization of each pairwise density separately. The authors reported that pairwise estimation procedure achieves significant computational gains and yields unbiased estimates as well as valid standard errors. An extension of the pair-wise model is reported in Faes et al. (2008) who used a pseudo-likelihood approach to jointly model the mixed outcomes of differing types. However, in contrast to Fieuws and Verbeke (2006), they maximized the pseudo-likelihood function at once rather than maximizing all pair-wise likelihoods separately.

In general, Non-Bayesian methods are usually computationally expensive (see, for example, the procedure described in Fieuws and Verbeke, 2006; Gueorguieva and Agresti, 2001; Regan and Catalano, 1999). Except for Faes et al. (2008), Fieuws and Verbeke (2006), most of the frequentist methods described above can only model a single discrete outcome

and a single continuous outcome. The extension to multiple mixed endpoints of different kinds is computationally prohibitive. However, a Bayesian approach like the one developed in (Bowman & George, 2018; Das et al., 1999; Dunson, 2000) can be easily extended for applications to multiple outcomes of differing types, although, they are also not designed to handle high-dimensional covariates settings.

1.2.4 Bayesian Method

To our knowledge, the use of Bayesian procedure for jointly modeling mixed discrete and continuous outcomes was first introduced by Das et al. (1999). These authors proposed a set of latent variables to deal with binary responses in their data to construct a multivariate mixed response model and use Gibbs sampling to derive the joint conditional density of the outcomes. In a more generalized setting, Dunson (2000) described a Bayesian approach using a mixture of generalized linear models for the joint distribution of latent variables for the clustered mixed outcomes. Cluster and subject level latent variables were assigned multivariate Gaussian densities or linked to variables with simple exponential families. Dunson (2000) developed MCMC algorithms for estimation of the posterior distribution. Additionally, Dunson et al. (2003) proposed a Bayesian framework for jointly modeling cluster size and multiple categorical and continuous outcomes measured on each subunit. They used a continuation ratio probit model for the cluster size and underlying normal regression models for each of the subunit-specific responses and accommodated the dependency between cluster size and the different endpoints through a latent variable structure. This model facilitates posterior computation via a simple and computationally efficient Gibbs sampler. In another study, Weiss et al. (2011) assumed some exponential distributions for the different outcomes and then linearly linked the unknown mean functions with random effect variables to account for the correlated structure.

Recently, Bowman and George (2018) developed a Bayesian approach for a joint

regression model of mixed type by building on Bayesian procedure developed for developmental toxicity studies in George et al. (2007). Following Albert and Chib (1993), they assume Gaussian latent variables for binary/ordinal and continuous responses and employ Gibbs sampling algorithm to obtain the exact posterior distribution of the parameters and latent variable of the discrete outcomes. One advantage of Bayesian methods proposed by Bowman and George (2018) is that all responses can be modeled jointly without factoring the likelihood and all correlations between outcomes and litter-mates are easily accounted for.

Almost all of the existing procedures focused on bivariate outcomes with a single (clustered/longitudinal) binary outcome and a single continuous outcome (see, for example, Catalano and Ryan, 1992; Fitzmaurice and Laird, 1995; George et al., 2007). While the model proposed by Dunson (2000), Faes et al. (2008) accommodate multiple mixed endpoints, they are not flexible or designed to handle high-dimensional covariates even in the context of small data set. In particular, Faes et al. (2008) model based on the likelihood framework involves complex numerical optimization (usually non-trivial) and may be sensitive to the choice of starting value and can be heavily biased for small samples. Model proposed in this theses also differ from that of Dunson (2000), Faes et al. (2008) in that it is designed to simultaneously carry out parameter estimation and variable selection when dealing with mixed outcomes that are characterized by high-dimensional covariates.

1.3 Contributions of this Study

The rapid increase in the use of technological advancement in research has made high-dimensional data sets widely available in many fields such as health sciences, economics, engineering, humanities, business, and finance among others. Examples of such data includes functional and longitudinal data, genetic marker analysis data, tomography, DNA methylation, microarray and proteomics data, natural language processing, e-commerce and

marketing data, signal processing, functional magnetic resolution imaging (fMRI), high resolution images etc. In such data sets, the number of covariates is usually bigger than the number of samples. Typically, some of these high-dimensional data are collected or observed in addition to the primary endpoints. Statisticians have developed some procedures for handling such data, however, there is dearth of variable selection procedures in the context of multivariate mixed endpoints.

While there have been several proposals put forward by several authors to address the statistical challenges that emanate from the mixed outcome data, the computational difficulties introduced by the availability of high-dimensional data is rarely discussed in the literature for multivariate mixed outcomes. To the best of our knowledge, this is the first attempt at addressing both the statistical and computational issues related to such data sets. In this dissertation, we developed a novel procedure to efficiently estimate the parameters and extract potentially active predictors simultaneously.

For our procedure, we suggest the use of spike-and-slab prior (Mitchell & Beauchamp, 1988) on the regression coefficients, we referred to this procedure as EMMEVS (EM mixed endpoints variable selection). While the originality of EMMEVS relies on the MCMC procedure, EMMEVS is a deterministic alternative (inspired by Ročková and George, 2014) based on the EM algorithm and can be used to rapidly expose potential sparse high posterior probability submodels. We described the EMMEVS procedure in Chapter 4.

The computational speed of EMMEVS procedures allow for the exploration of many sub-models within a short period and both have the potential of lowering the computational burden of MCMC methods when estimating posterior distributions over subsets of potential predictors. Finally, the method developed here easily extends to clustered longitudinal (or repeated measures) binary/polychotomous and continuous outcomes as well as spatial outcomes data of different kinds.

To our knowledge, there is a lack of software/packages for the implementation of

the multivariate mixed endpoints procedures; we shall make freely available an R package "mme" (under development) to implement our methods at the following GitHub address: https://github.com/yone4real/mme. We also plan to include the implementation of other procedures such as general location model and some latent variable approach in the nearest future.

The remainder of the dissertation is organized as follows: Chapter 2 describes the model specification for high-dimensional multivariate mixed endpoints models for clustered design (Section 2.1) and longitudinal/repeated-measures design (Section 2.2). Chapter 3 presents the use of EMMEVS to answer our research question posed in Section 1.1. The details of the prior formulation and the EMMEVS procedure are described in Chapter 3. In Chapter 4, we present the analysis of two real-life data to demonstrate the application of our EMMEVS algorithm. In Chapter 5, we give a comprehensive accounts of the ongoing/future work. In particular, we give a partial discussion of two new methodologies that can be used to jointly modeled multivariate mixed endpoints and conduct variable selection simultaneously.

Chapter 2

High-dimensional Multivariate Mixed Endpoints Models

2.1 Introduction

The models proposed in this dissertation are fitted to clustered mixed endpoints, repeated measures mixed endpoints, longitudinal mixed endpoints, and clustered longitudinal (or repeated measures) mixed endpoints data. We start by providing the definitions of these designs.

2.1.1 Clustered Mixed Endpoints Data

By clustered mixed endpoints data, we mean data sets in which endpoints of different structures such as continuous, ordinal, binary/polychotomous are measured once on each subject or experimental unit, and the subjects are grouped or nested within the clusters of units. As an example, consider an experiment with G treatment groups in which m_i independent clusters are exposed to the ith $(i = 1, \cdots, G)$ treatment group of a test compound. In Figure 1, we illustrate the data structure of the clustered mixed endpoints characterized by high-dimensional covariates in the context of developmental toxicity experiment. The characteristics of the clustered mixed endpoints data described above naturally exhibits three levels (see, Figure 2). The subject (level 1 data) is nested within-cluster (level 2) which are in turn nested within the experimental group. The endpoints and other variables of interests are measured on the unit of analysis at level 1. Features obtained at the subject level help to understand the within-subject variability on the responses of interests; also, cluster characteristics help explain the within-cluster variability in the cluster-specific average responses, and treatment features help to understand the variations in treatment-specific mean responses. Three sources of variations to the responses arise in the experimental setting

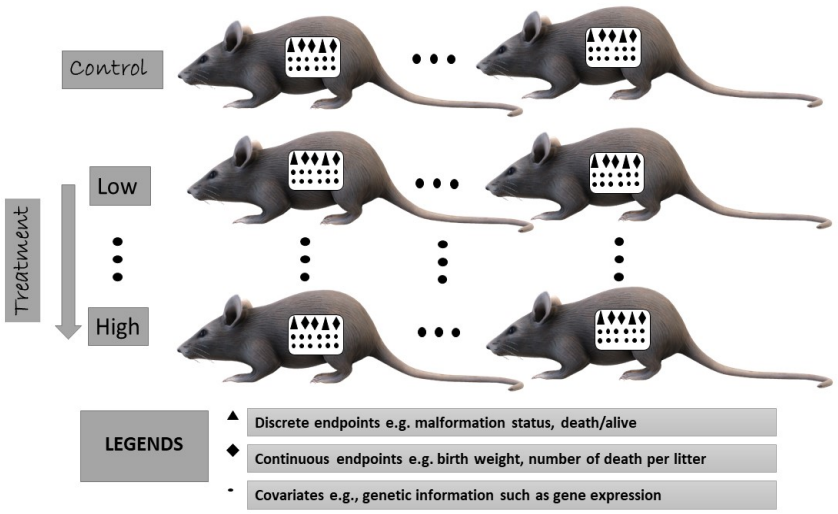

Figure 1. Data Structure for Multivariate Clustered Mixed Endpoints in a Developmental Toxicity Study

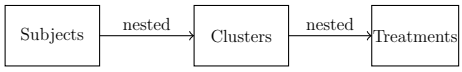

Figure 2. Nesting Structure for Clustered Mixed Endpoints.

above:

(a) σ^2_ϵ, variation in the responses that occurs across responses obtained from the kth subject (level 1) nested within the jth cluster (level 2) in the ith treatment group (level 3),

(b) $\sigma^2_{1\ell}$ and σ^2_{2f}, the variation in the average responses at the jth cluster (level 2) nested within the ith treatment group (level 3), and

(c) σ^2_a, variation in the mean responses across the G treatment groups (level 3).

2.1.2 Longitudinal (or Repeated Measures) Mixed Endpoints Data

We define longitudinal mixed endpoints data as data sets with repeated measurements on the same subject at the same treatment level. The repetition may be overtime and in

this case, measurements are labeled as longitudinal. An illustration of the data structure for the longitudinal mixed endpoints data is depicted in Figure 3.

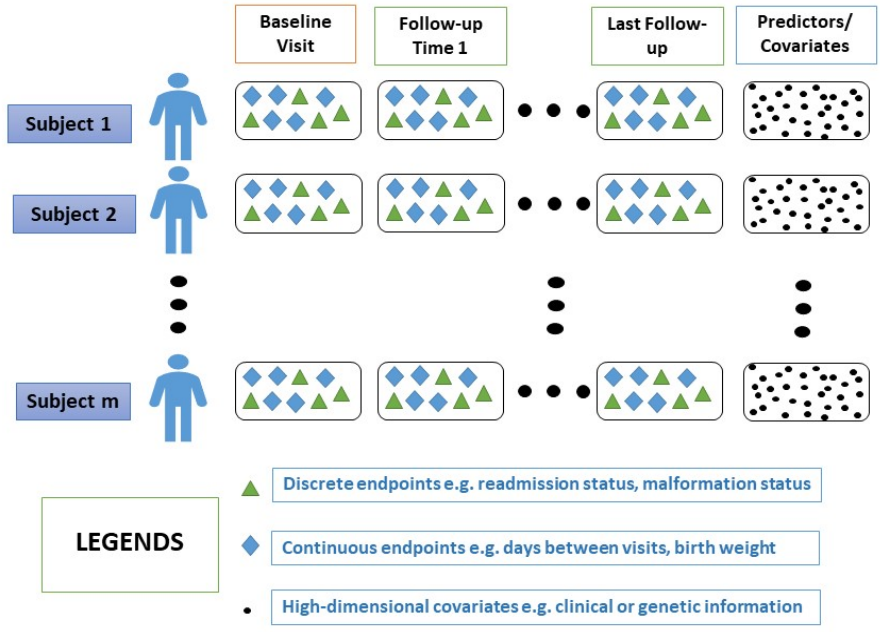

Figure 3. Data Structure for the Multivariate Longitudinal Mixed Endpoints Design

2.1.3 Clustered Longitudinal (or Repeated Measures) Mixed Endpoints Data

Here, we combined the features of both clustered and longitudinal (or repeated measures) data described in Sections 2.1.1 and 2.1.2 respectively. Consider a design in which at each of the t_{ij} time-points, multiple mixed endpoints of different nature are collected on the jth subject (or unit of analysis), $j = 1, \cdots, m,\ i = 1, \cdots, n_j$. For example, consider a dose-response animal teratology study in which mixed longitudinal measurements are made on each animal. This design is complex due to an extra layer of correlation introduced to accommodate the longitudinal structure in the data. Faes et al. (2004), Gueorguieva and

Sanacora (2006) and others addressed data of this type, however, not in the context of high-dimensional covariates.

2.2 Model Specification

The model for analyzing mixed longitudinal data accommodate more than two primary endpoints of differing types. In the next section, we focus on the model for clustered mixed endpoints and used that to lay foundation for mixed longitudinal/repeated-measures in Section 2.2.2

2.2.1 Model for Clustered Mixed Endpoints

Let $W_{\ell ijk}$ and X^*_{fijk} represent the ℓth continuous and fth discrete endpoints obtained from the kth subject of the jth cluster in the ith treatment group, $1 \leq i \leq G, 1 \leq j \leq m_i, 1 \leq k \leq n_{ij}$, where n_{ij} is the number of measurements in the jth cluster of the ith treatment group, m_i is the number of clusters in the ith treatment group, and G denote the number of treatment groups. Also, let the observed discrete outcomes, X^*_{fijk}, have s distinct values (i.e., s mutually exclusive ordered categories), say $x_1 < \cdots < x_s$. Corresponding to each observation X^*_{fijk} and following Albert and Chib (1993), a continuous latent variable X_{fijk} associated with the discrete endpoints, X^*_{fijk}, is introduced and defined by

$$X^*_{fijk} = \begin{cases} x_1 & \text{if } X_{fijk} \in (\zeta_0, \zeta_1) \\ \vdots & \vdots \\ x_O & \text{if } X_{fijk} \in [\zeta_O, \zeta_{O+1}) \\ \vdots & \vdots \\ x_s & \text{if } X_{fijk} \in [\zeta_{s-1}, \zeta_s) \end{cases} \tag{2.1}$$

with $\zeta_0 = -\infty$ and $\zeta_s = \infty$ and the remaining unknown cutpoints satisfying $\zeta_1 < \zeta_2 < \cdots < \zeta_{(s-1)}$. For binary data, it suffices to assume a single cutpoint in (2.1) is set to zero as adopted in Catalano and Ryan (1992).

The model equation for $W_{\ell ijk}$ and X_{fijk} is assumed to satisfy the linear mixed effects model

$$\begin{aligned} W_{\ell ijk} &= \alpha_{0\ell} + \alpha_{1\ell} d_i + \boldsymbol{z}_{i\ell}^{*'} \boldsymbol{\alpha}_\ell^* + a_{ijk} + b_{1_\ell ij} + \epsilon_{\ell ijk} \\ &= \boldsymbol{z}_{i\ell}^{'} \boldsymbol{\alpha}_\ell + a_{ijk} + b_{1_\ell ij} + \epsilon_{\ell ijk} \\ X_{fijk} &= \beta_{0f} + \beta_{2f} d_i + \boldsymbol{z}_{if}^{*'} \boldsymbol{\beta}_f^* + a_{ijk} + b_{2_f ij} + \epsilon_{fijk} \\ &= \boldsymbol{z}_{if}^{'} \boldsymbol{\beta}_f + a_{ijk} + b_{2_f ij} + \epsilon_{fijk} \end{aligned} \tag{2.2}$$

for $1 \leq f \leq h, 1 \leq \ell \leq c, 1 \leq i \leq g, 1 \leq j \leq m_i$ and $1 \leq k \leq n_{ij}$, where, $\alpha_{0\ell}$ and $\alpha_{1\ell}$ the intercept and slope of the ℓth outcome with β_{0f} and β_{2f} defined analogously for the fth discrete endpoint, $\boldsymbol{z}_{i\ell}$ and $\boldsymbol{z}_{if}$ represents the design vector for the fixed effects (modeled as a function of the treatment group and other covariates), $\boldsymbol{\alpha}_\ell^*$ and $\boldsymbol{\beta}_f^*$ are p-dimensional vector of unknown fixed regression coefficients, a_{ijk} the random effects for the association between continuous and discrete endpoints, $b_{1_\ell ij}$ and $b_{2_f ij}$ is the random effects that accommodates the clustering structure in the data, $\epsilon_{\ell ijk}$ and ϵ_{fijk} are the random (measurement) error.

The different terms in (2.2) reflects the eight different types of association between the endpoints, as illustrated in Figure 4. The

(a) cluster-mates on the same continuous outcomes is cov $\left(W_{\ell ijk}, W_{\ell ijk'}\right) = \sigma_{1_\ell}^2$,

(b) cluster-mates on the same latent variable have cov $\left(X_{fijk}, X_{fijk'}\right) = \sigma_{2_f}^2$,

(c) ℓth continuous outcomes and fth latent variable on the same subject have cov $(W_{\ell ijk}, X_{fijk})$ $= \sigma_a^2 + \sigma_{1_\ell 2_f}$,

(d) ℓth continuous outcomes and fth latent variables between cluster-mates have cov $\left(W_{\ell ijk}, X_{fijk'}\right)$ $= \sigma_{1_\ell 2_f}$,

(e) different continuous outcomes on the same subject have cov $\left(W_{\ell ijk}, W_{\ell' ijk}\right) = \sigma_a^2 + \sigma_{1_\ell 1_{\ell'}}$,

(f) two latent variables on the same fetus have cov $\left(X_{fijk}, X_{f' ijk}\right) = \sigma_a^2 + \sigma_{2_f 2_{f'}}$,

(g) different continuous outcomes on cluster-mates from different subject have cov $\left(W_{\ell ijk}, W_{\ell' ijk'}\right)$ $= \sigma_{1_\ell 1_{\ell'}}$,

(h) different latent variables on cluster-mates from different subject have cov $\left(X_{fijk}, X_{f' ijk'}\right) = \sigma_{2_f 2_{f'}}$,

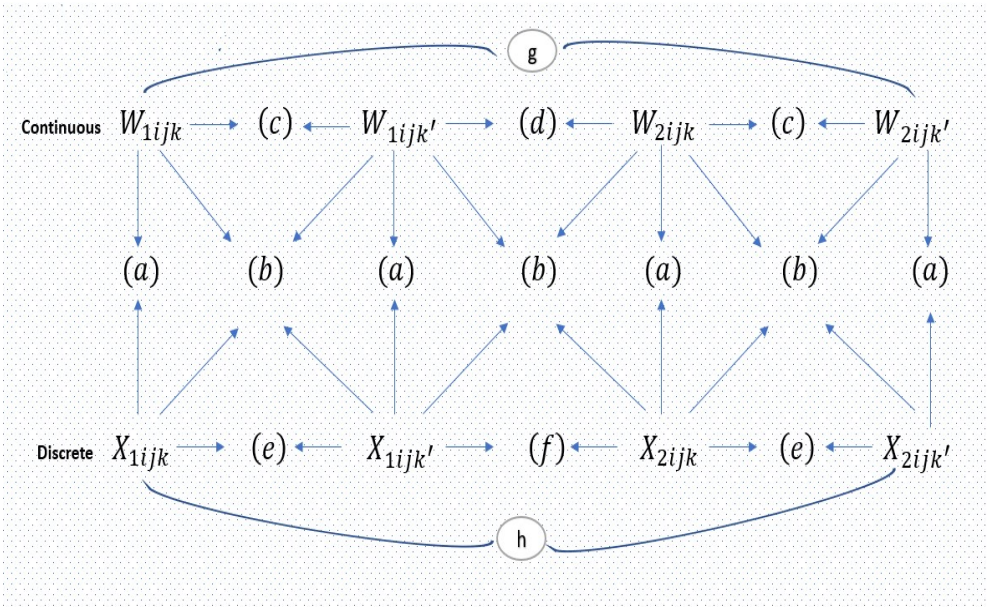

Figure 4. Association structure for two continuous and two latent variables for subject k and k' in the jth cluster of the ith treatment group.

Distributional Assumptions

In equation (2.2), we have four random variables (a_{ijk}, $b_{1_\ell ij}, b_{2_f ij}$, and $\epsilon_{\ell ijk}$), the condition for the variance components form of equation (2.2) is given as:

$$\begin{aligned}
\mathrm{E}\left(a_{ijk}\right) &= 0 \text{ and } \mathrm{var}\left(a_{ijk}\right) = \sigma_a^2,\ \forall\ i,j,k, \\
\mathrm{E}\left(b_{1_\ell ij}\right) &= \mathrm{E}\left(b_{2_f ij}\right) = 0 \text{ and } \mathrm{var}\left(b_{1_\ell ij}\right) = \sigma_{1_\ell}^2,\ \mathrm{var}\left(b_{2_f ij}\right) = \sigma_{2_f}^2,\ \forall\ \ell, f, i, j \\
\mathrm{E}\left(\epsilon_{\ell ijk}\right) &= \mathrm{E}\left(\epsilon_{fijk}\right) = 0,\ \text{and } \mathrm{var}\left(\epsilon_{\ell ijk}\right) = \mathrm{var}\left(\epsilon_{fijk}\right) = \sigma_\epsilon^2,\ \forall\ \ell, f, i, j, k \\
\mathrm{cov}\left(a_{ijk}, a_{i^{'}j^{'}k^{'}}\right) &= 0,\ \forall\ i \neq i^{'}, j \neq j^{'}, k \neq k^{'} \text{ except for } i = i^{'}, j = j^{'}, k = k^{'}, \\
\mathrm{cov}\left(b_{1_\ell ij}, b_{1_{\ell^{'}} ij}\right) &= \mathrm{cov}\left(b_{2_f ij}, b_{2_{f^{'}} ij}\right) = 0,\ \forall\ \ell \neq \ell^{'},\ f \neq f^{'} \text{ except for } \ell = \ell^{'}, f = f^{'}, \\
\mathrm{cov}\left(\epsilon_{\ell ijk}, \epsilon_{\ell^{'}i^{'}j^{'}k^{'}}\right) &= 0,\ \forall\ i \neq i^{'}, j \neq j^{'}, k \neq k^{'}, \ell \neq \ell^{'},\ \text{except for } i = i^{'}, j = j^{'}, k = k^{'}, \ell = \ell^{'}
\end{aligned} \tag{2.3}$$

We assumed a Gaussian distribution for the unobserved continuous latent variable underlying the discrete outcome, X^*_{fijk}, that is $X_{fijk} \sim N\left(\boldsymbol{z}_{if}^{'}\boldsymbol{\beta}_f + a_{ijk} + b_{2_f ij}, \sigma_\epsilon^2\right)$, leading to the the joint density of X^*_{fijk} and X_{fijk}, conditional on $\boldsymbol{\beta}_f$, a_{ijk}, $b_{2_f ij}$, and σ_ϵ^2, as

$$\begin{aligned}
f\left(x_{fijk}, x^*_{fijk} | \boldsymbol{\beta}_f, a_{ijk}, b_{2_f ij}, \sigma_\epsilon^2\right) \propto & \left[\sum_{q=1}^{s} \left\{ I\left(x^*_{fijk} = q-1\right) I\left(\zeta_{q-1} < X_{fijk} \leq \zeta_q\right) \right\}\right] \\
& \times \phi\left(x_{fijk}; \boldsymbol{z}_{if}^{'}\boldsymbol{\beta}_f + a_{ijk} + b_{2_f ij}, \sigma_\epsilon^2\right)
\end{aligned} \tag{2.4}$$

where $\phi\left(\cdot\right)$ is the probability density function of normal distribution and $I\left(.\right)$ is the indicator function. In particular, if we assumed X^*_{fijk} is binary, then, equation (2.4) becomes:

$$\propto \left[\left\{ I\left(x_{fijk} > 0, x^*_{fijk} = 1\right) + I\left(x_{fijk} \leq 0, x^*_{fijk} = 0\right) \right\}\right] \phi\left(x_{fijk}; \boldsymbol{z}_{if}^{'}\boldsymbol{\beta}_f + a_{ijk} + b_{2_f ij}, \sigma_\epsilon^2\right)$$

We make the following assumptions about the four random components in equations (2.2), $\epsilon_{\ell ijk}$, ϵ_{fijk} are iid $N(0, \sigma_\epsilon^2)$, $a_{ijk} \sim N(0, \sigma_a^2)$, and $\boldsymbol{b} = \begin{pmatrix} b_{1_1 ij} & \cdots & b_{1_c ij} & b_{2_1 ij} & \cdots & b_{2_h ij} \end{pmatrix}^{'} \sim$

$\mathrm{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$, where $q = c + h$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{1_1} & \sigma_{1_1 1_2} & \cdots & \sigma_{1_1 1_c} & \sigma_{1_1 2_1} & \sigma_{1_1 2_2} & \cdots & \sigma_{1_1 2_h} \\ & \sigma^2_{1_2} & \cdots & \sigma_{1_2 1_c} & \sigma_{1_2 2_1} & \sigma_{1_2 2_2} & \cdots & \sigma_{1_2 2_h} \\ & & \ddots & \sigma^2_{1_c} & \sigma_{1_c 2_1} & \sigma_{1_c 2_2} & \cdots & \sigma_{1_c 2_h} \\ & & & & \sigma^2_{2_1} & \sigma_{2_1 2_2} & \cdots & \sigma_{2_1 2_h} \\ & & & & & \sigma^2_{2_2} & \cdots & \sigma_{2_2 2_h} \\ & & & & & & \ddots & \sigma^2_{2_h} \end{bmatrix}$$

. We also assume that the random effects $\boldsymbol{b}$, a_{ijk}, as well as the random errors $\epsilon_{\ell ijk}$ and ϵ_{fijk} are all mutually independent.

Joint density of $\boldsymbol{W}$, $\boldsymbol{X}$, and $\boldsymbol{X}^*$

In view of equations (2.2), (2.4), and using the distributional assumptions given in section 2.2.1, the joint density of $\boldsymbol{W}$, $\boldsymbol{X}$, and $\boldsymbol{X}^*$ is given by

$$\begin{aligned} f(\cdot) &= f(W_{\ell ijk}|\boldsymbol{\alpha}_\ell, a_{ijk}, b_{1_\ell ij}, \sigma^2_\epsilon) \times f\left(X_{fijk}, X^*_{fijk}|\boldsymbol{\beta}_f, a_{ijk}, b_{2_f ij}, \sigma^2_\epsilon\right) \\ &= \prod_{\ell ijk} \left(\frac{1}{\sigma^2_\epsilon}\right)^{\frac{1}{2}} \exp\left(-\frac{\left(W_{\ell ijk} - \boldsymbol{z}'_{i\ell}\boldsymbol{\alpha}_\ell - a_{ijk} - b_{1_\ell ij}\right)^2}{2\sigma^2_\epsilon}\right) \\ &\quad \times \prod_{fijk} \left[\sum_{q=1}^{s} \left\{ I\left(X^*_{fijk} = q-1\right) I\left(\zeta_{q-1} < X_{fijk} \leq \zeta_q\right) \right\} \right. \\ &\quad \left. \times \left(\frac{1}{\sigma^2_\epsilon}\right)^{\frac{1}{2}} \exp\left(-\frac{\left(X_{fijk} - \boldsymbol{z}'_{if}\boldsymbol{\beta}_f - a_{ijk} - b_{2_f ij}\right)^2}{2\sigma^2_\epsilon}\right) \right] \end{aligned}$$

2.2.2 Model for Longitudinal (or Repeated Measures) Mixed Endpoints

Here, we define $W_{\ell ij}$ as the longitudinal measurements observed on the jth subject or unit of analysis at time i for the ℓth continuous endpoint. Similarly, we let X^*_{fij} be the

longitudinal measurements observed on the jth subject or unit of analysis at time i for the fth discrete endpoint. The observed discrete endpoints X^*_{fij} is assumed to be related to the unobserved latent variable X_{fij} by

$$X^*_{fij} = \begin{cases} 1 & \text{if } \zeta_0 < X_{fij} \leq \zeta_1 \\ l & \text{if } \zeta_{l-1} < X_{fij} \leq \zeta_l,\ l = 2, \cdots, L-1 \\ L & \text{if } \zeta_{L-1} < X_{fij} \leq \zeta_L \end{cases} \tag{2.5}$$

where $\zeta_0, \zeta_2, \cdots, \zeta_L$ are cutpoints whose coordinate satisfy $-\infty = \zeta_0 < \zeta_1 < \cdots < \zeta_{L-1} < \zeta_L = \infty$. Here, ζ_{L-1} and ζ_L are respectively defined as the lower and upper endpoints of the interval corresponding to observed outcome L. We also assume that $W_{\ell ij}$ and X_{fij} satisfy the linear mixed-effect model

$$\begin{aligned} W_{\ell ij} &= \alpha_{0\ell} + \alpha_{1\ell} t_{ij} + \boldsymbol{z}^{*'}_{i\ell} \boldsymbol{\alpha}^*_\ell + a_{ij} + b_{1_\ell i} + \epsilon_{\ell ij} \\ &= \boldsymbol{z}^{'}_{i\ell} \boldsymbol{\alpha}_\ell + a_{ij} + b_{1_\ell i} + \epsilon_{\ell ij} \\ X_{fij} &= \beta_{0f} + \beta_{2f} t_{ij} + \boldsymbol{z}^{*'}_{if} \boldsymbol{\beta}^*_f + a_{ij} + b_{2_f i} + \epsilon_{fij} \\ &= \boldsymbol{z}^{'}_{if} \boldsymbol{\beta}_f + a_{ij} + b_{2_f i} + \epsilon_{fij} \end{aligned} \tag{2.6}$$

for $1 \leq f \leq h, 1 \leq \ell \leq c, 1 \leq i \leq n_j, 1 \leq j \leq m$, where, $\alpha_{0\ell}$ and $\alpha_{1\ell}$ the intercept and slope of the ℓth outcome with β_{0f} and β_{2f} defined analogously for the fth discrete endpoint, $\boldsymbol{z}_{i\ell}$ and $\boldsymbol{z}_{if}$ represents the design matrix for the fixed effects (can be modeled as a function of the time/repeated-measures and other covariates), $\boldsymbol{\alpha}^*_\ell$ and $\boldsymbol{\beta}^*_f$ are p-dimensional vector of unknown fixed regression coefficients, a_{ij} the random effects for the association between the ℓth continuous and fth discrete endpoints observed on the jth subject at time i, $b_{1_\ell i}$ and $b_{2_f i}$ are the random effects that accounts for the correlation between measurements from the same subject, $\epsilon_{\ell ij}$ and ϵ_{fij} are the random (measurement) error, n_j is the number of times

the jth subject is observed, t_{ij} indexes the n_j longitudinal measurements made on the jth subject at time i.

Similar to the clustered mixed endpoint model, we assumed that $a_{ij} \sim \mathcal{N}(0, \sigma_a^2)$, $\boldsymbol{b} = \left(b_{1_\ell i}, b_{2_f i}\right)' \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma})$, and $\epsilon_{\ell ij}$ and ϵ_{fij} are independent pairs of $\mathcal{N}(0, \sigma_\epsilon^2)$ random errors. The matrix specification of equation (2.6) is the same as that given in (2.7) with slight notation in the definition.

Matrix Specification of Regression Model (2.2)

In this section, we consider a general matrix specification of (2.2) for all subjects in the study. We do this by stacking formula (2.2) (given for individual k) into vectors and matrices. Let $n_1 = c \times G \times \sum_i m_i \times \sum_{ij} n_{ij}$ represent the total number of observations for the continuous responses and $n_2 = h \times G \times \sum_i m_i \times \sum_{ij} n_{ij}$ be similarly defined for the continuous latent variable. Define $n = n_1 + n_2$ as the total number of observations in the study, $q_1 = G \times \sum_i m_i \times \sum_{ij} n_{ij}$ and $q_2 = G \times \sum_i m_i$. A more compact notation for equation (2.2) is given by

$$\begin{aligned} \boldsymbol{W} &= \boldsymbol{Z}_1\boldsymbol{\alpha} + \boldsymbol{a} + \boldsymbol{b}_1 + \boldsymbol{\epsilon}_1 \\ \boldsymbol{X} &= \boldsymbol{Z}_2\boldsymbol{\beta} + \boldsymbol{a} + \boldsymbol{b}_2 + \boldsymbol{\epsilon}_2 \end{aligned} \tag{2.7}$$

where

- $\boldsymbol{W}' = \left(W_{1111}, \cdots, W_{cGm_Gn_{Gm_G}}\right)$ is an $n_1 \times 1$ vector of continuous responses,
- $\boldsymbol{X}' = \left(X_{1111}, \cdots, X_{hGm_Gn_{Gm_G}}\right)$ is an $n_2 \times 1$ vector of continuous latent variable,
- $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \cdots \boldsymbol{\alpha}_\ell)'$ is a $[(c \times (2+p)) \times 1]$-dimensional vector,
- $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots \boldsymbol{\beta}_f)'$ is a $[(h \times (2+p)) \times 1]$-dimensional vector,
- $\boldsymbol{Z}_1' = \left[(\mathbf{1}_{N_1} \otimes \boldsymbol{z}_{1\ell})', \cdots, (\mathbf{1}_{N_G} \otimes \boldsymbol{z}_{G\ell})' \right]$, is $(n_1 \times 2c)$ dimensional matrix,

- $\boldsymbol{Z}_2^{'} = \left[\left(\mathbf{1}_{N_1} \otimes \boldsymbol{z}_{1f}\right)^{'}, \cdots, \left(\mathbf{1}_{N_G} \otimes \boldsymbol{z}_{Gf}\right)^{'} \right]$, is $(n_2 \times 2h)$ dimensional matrix,
- $\boldsymbol{a}^{'} = \left(a_{111}, \cdots, a_{Gm_G n_{Gm_G}}\right)$, a $q_1 \times 1$ vector of random effects,
- $\boldsymbol{b}^{'} = \left(\boldsymbol{b}_1^{'}, \boldsymbol{b}_2^{'}\right)^{'}$ is a $q_2 \times 1$ vector of random effects associated with the cluster with $\boldsymbol{b}_1^{'} = \left(\mathbf{1}_{n_{11}} b_{111}, \cdots, \mathbf{1}_{n_{Gm_G}} b_{1n_{Gm_G}}\right)^{'}$, $\boldsymbol{b}_2^{'} = \left(\mathbf{1}_{n_{11}} b_{211}, \cdots, \mathbf{1}_{n_{Gm_G}} b_{2n_{Gm_G}}\right)^{'}$,
- $\boldsymbol{\epsilon}_1 = \left(\epsilon_{1111}, \cdots, \epsilon_{cGm_G n_{Gm_G}}\right)$ is an $n_1 \times 1$ vector of measurement errors for the continuous outcomes, and
- $\boldsymbol{\epsilon}_2 = \left(\epsilon_{1111}, \cdots, \epsilon_{hGm_G n_{Gm_G}}\right)$ is an $n_2 \times 1$ vector of measurement errors for the latent variables.

Chapter 3

The EMMEVS Procedure

3.1 Introduction

The MCMC stochastic search method is one of the most commonly used procedure for analyzing high-dimensional data and when the posterior is intractable (George & McCulloch, 1997; Li & Zhang, 2010; Mitchell & Beauchamp, 1988), however, using these methods have been shown to be slow and inefficient (Griffin & Brown, 2010; Ročková & George, 2014) especially when dealing with high-dimensional data. A rapid and efficient deterministic alternative expanding on the EM approach of Ročková and George (2014) is suggested to lower the computational burden of the MCMC procedure and simultaneously conduct variable selection and parameter estimation. Our proposal, referred to as EM for multivariate mixed endpoints variable selection (EMMEVS) can carry out parameter estimation and variable selection simultaneously. Also, EMMEVS can identify the posterior modes directly without the need for full stochastic search thereby leading to some computational time saving. Further, we showed that EMMEVS enables fast exploration of the posterior under a sequence of mixture priors.

The EMMEVS algorithm uses Laplace distribution on the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ rather than the commonly used Gaussian distribution, this strategy enables a simpler closed form update and leads to an adaptive LASSO-type objective function. With this property, the EMMEVS algorithm can benefit from some efficient algorithm such as glmnet of Friedman et al. (2010), dynamic weighted LASSO of Chang and Tsay (2010), and LARS algorithm of Efron et al. (2004) to estimate the parameters and conduct variable selection simultaneously. Thus, EMMEVS estimator can shrink most of the redundant variables to exactly zero or near zero.

The idea behind EMMEVS formulation is akin to (Figueiredo, 2003; Griffin & Brown,

2005, 2012; Kiiveri, 2003; Ročková & George, 2014), in which they combine the EM algorithm with Bayesian shrinkage estimation under sparsity priors. Although the originality of our proposal is anchored on stochastic search variable selection (SSVS) procedure developed by (George & McCulloch, 1993, 1997), EMMEVS is a flexible deterministic alternative to the SSVS based on EM algorithm (Dempster et al., 1977). In contrast to the SSVS where inference is drawn from the fully sampled posterior distribution using MCMC, EMMEVS estimates the posterior modes with the EM algorithm. Meanwhile, it is widely recognized that the EM algorithm is not guaranteed to converge to the global mode (this can result in biased estimates) and is sensitive to starting values, we suggest a deterministic annealing variant of the EMMEVS to help mitigate the potential problem of entrapment in local modes and thus, improves its performance. This is discussed in section 3.5.

3.2 The EMMEVS Algorithm

3.2.1 Setup and Hierarchical Prior Formulation

To conduct model selection through Bayesian method, two main ingredients are essential: (a) a prior to induce the posterior distribution over the subsets of potential covariates, and (b) a method to retrieve/identify the potential promising covariates from the posterior. In this section, we detailed the formulation of prior for our proposed method - EMMEVS. In what follows, we let $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, \sigma^2_\epsilon, \sigma^2_a, \boldsymbol{\Sigma})$ be the vector of parameters to be estimated. First, we consider an hierarchical prior formulation for the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

To identify active predictors in $\boldsymbol{Z}$ that are potentially associated with the continuous and discrete outcomes $\boldsymbol{W}$ and $\boldsymbol{X}$, two p-dimensional vector of binary latent variables $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)'$ and $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)'$ are introduced, for $(\gamma_r, \mu_r) \in \{0, 1\}$, $r = 1, \cdots, p$. The model selection proceeds by selecting the columns of $\boldsymbol{Z}$ for which $\gamma_r = 1$ or $\mu_r = 1$. Combined with suitable prior distributions over $\boldsymbol{\Theta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\mu}$, the induced posterior distribution $\tau(\boldsymbol{\gamma}|\boldsymbol{W})$

and $\tau(\boldsymbol{\mu}|\boldsymbol{X})$, then summarizes the post-data variable selection uncertainties. Next, we describe our prior specification on the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we placed the following spike-and-slab Laplace mixture of priors on each entries of $\alpha_{\ell r}$ and β_{fr}

$$\begin{aligned}
\tau(\alpha_{\ell r}|\gamma_r, \lambda_{\ell r}) &= \prod_{\ell=1}^{c}\prod_{r=1}^{p} \frac{1}{2\lambda_{\ell r}} e^{-\frac{|\alpha_{\ell r}|}{\lambda_{\ell r}}} \quad \text{with} \quad \lambda_{\ell r} = \lambda_0(1-\gamma_r) + \lambda_1\gamma_r \\
\tau(\beta_{fr}|\mu_r, \omega_{fr}) &= \prod_{f=1}^{h}\prod_{r=1}^{p} \frac{1}{2\omega_{fr}} e^{-\frac{|\beta_{fr}|}{\omega_{fr}}} \quad \text{with} \quad \omega_{fr} = \omega_0(1-\mu_r) + \omega_1\mu_r
\end{aligned} \tag{3.1}$$

where λ_0 and ω_0 are the variance of the spike distribution, and λ_1 and ω_1 are the variance of the slab distribution for both continuous and discrete endpoints respectively. Our prior specification in (3.1) differs from the Bayesian LASSO of Park and Casella (2008) in that the degree of shrinkage for the rth coefficient is controlled by the hyperparameters $\lambda_{\ell r}$ and ω_{fr} (for $r = 1, \cdots, p$), each a mixture of two different scales (λ_0, λ_1) and (ω_0, ω_1).

Several authors have recommend different values for the scale parameter of the spike distribution. In the traditional spike-and-slab prior, the spike component is set to be a mass at zero, which corresponds to our setting $\lambda_0 = 0$ and $\omega_0 = 0$. This approach of setting λ_0 and ω_0 to zero has been implemented in Brown et al. (2002), Panagiotelis and Smith (2008) and Hu et al. (2015). Here, we use a small continuous version of the spike-and-slab prior in which λ_0 and ω_0 is set to positive nonzero value but relatively small compared with λ_1 and ω_1, this strategy has also been used in George and McCulloch (1997) and Ročková and George (2014). We found this strategy to be very efficient in excluding unimportant nonzero effects (by inducing strong shrinkage on estimation) and to lead to a rapid EMMEVS procedure. An advantage of using a continuous spike-and-slab prior is that the continuous prior distribution on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ allows the use of efficient algorithm that do not require substituting the active dimension of the parameters (George & McCulloch, 1993, 1997).

As for λ_1 and ω_1, they are set to be relatively large, thereby serving as the "slab scale"

for modeling large coefficients and thus induce weak or no shrinkage on the estimation. In sum, by setting $\lambda_1 \gg \lambda_0$ and $\omega_1 \gg \omega_0$, the Laplace mixture of priors imposes a different strength of shrinkage for elements drawn from the slab parameters (λ_1, ω_1) and spike parameters (λ_0, ω_0). The advantage of this representation is that it allows us to shrink coefficients in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to zero if they are small in scale while not biasing the large coefficients.

The remaining components of the hierarchical prior specification is completed with a prior distribution on $\tau(\boldsymbol{\gamma})$ and $\tau(\boldsymbol{\mu})$ over the 2^p possible values of $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$. Here, we focus on an hierarchical prior specifications of the form

$$\tau(\boldsymbol{\gamma}) = E_{\tau(\boldsymbol{\theta}_1)}\tau(\boldsymbol{\gamma}|\boldsymbol{\theta}_1) \quad \text{and} \quad \tau(\boldsymbol{\mu}) = E_{\tau(\boldsymbol{\theta}_2)}\tau(\boldsymbol{\mu}|\boldsymbol{\theta}_2) \tag{3.2}$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ controls the sparsity and are defined as vectors of the proportion of non-zero regression coefficients for the continuous and discrete outcomes respectively. A default choice for $\tau(\gamma_r|\theta_{1\ell})$ and $\tau(\mu_r|\theta_{2f})$ is iid Bernoulli prior

$$\begin{aligned} \tau(\gamma_r|\theta_{1\ell}) &= \prod_{\ell=1}^{c} \theta_{1\ell}^{\sum_{r=1}^{p}\gamma_r}(1-\theta_{1\ell})^{1-\sum_{r=1}^{p}\gamma_r} \text{ with } 0 \leq \theta_{1\ell} \leq 1 \\ \tau(\mu_r|\theta_{2f}) &= \prod_{f=1}^{h} \theta_{2f}^{\sum_{r=1}^{p}\mu_r}(1-\theta_{2f})^{1-\sum_{r=1}^{p}\mu_r} \text{ with } 0 \leq \theta_{2f} \leq 1 \end{aligned} \tag{3.3}$$

however, in the presence structural information or networking group about the covariates, other structured priors such as logistic regression product prior (Stingo et al., 2010) and Markov random field prior (Li & Zhang, 2010) can be used. The probability parameters $\theta_{1\ell}$ and θ_{2f} can be view as the overall shrinkage parameters that equals the prior probabilities $Pr(\gamma_r = 1|\theta_{1\ell})$ and $Pr(\mu_r = 1|\theta_{2f})$. The prior expectation of the scales λ_r and ω_r are given as, $E(\lambda_r) = \lambda_0(1-\theta_{1\ell}) + \lambda_1\theta_{1\ell}$ and $E(\omega_r) = \omega_0(1-\theta_{2f}) + \omega_1\theta_{2f}$ with their values lying in the range $[\lambda_0, \lambda_1]$ and $[\omega_0, \omega_1]$ respectively; with this form, any marginal $\tau(\boldsymbol{\gamma})$ and $\tau(\boldsymbol{\mu})$ given in (3.2) will be exchangeable on the components of $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$. We consider

exchangeable priors $\tau(\theta_{1\ell}) \sim \text{beta}(l_1, o_1) \propto \theta_{1\ell}^{l_1-1}(1-\theta_{1\ell})^{o_1-1}$ and $\tau(\theta_{2f}) \sim \text{beta}(l_2, o_2) \propto \theta_{2f}^{l_2-1}(1-\theta_2)^{o_2-1}$ for $l_i, o_i > 0$ $(i = 1, 2)$ which results in beta-binomial priors $\tau(\boldsymbol{\gamma})$ and $\tau(\boldsymbol{\mu})$ that can favor parsimony (Scott & Berger, 2010).

One advantage of using the Laplace mixture of priors on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ over the Gaussian mixture formulation is that it leads to faster convergence (Armagan et al., 2013). Also, the regression coefficients cannot attain exact zeros in the normal mixture representation, and additional post-processing steps are required for variable selection, which can be sensitive to cut-off values. In the context of structured high-dimensional data, Chang et al. (2018) reported numerical inconveniences that may arise when $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ approaches zero under the Gaussian mixture formulation, noting that the conditional mean may explode to infinity.

The prior specifications for other parameters are $a_{ijk}|\sigma_a^2 \sim N(0, \sigma_a^2)$, $\boldsymbol{b}|\boldsymbol{\Sigma} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. As for the variance components, we assume the following non-informative prior distributions;

$$\sigma_\epsilon^2 \sim \mathcal{IG}(\nu_1, \kappa_1), \quad \sigma_a^2 \sim \mathcal{IG}(\nu_2, \kappa_2), \quad \boldsymbol{\Sigma} \sim \mathcal{IW}(\nu_0, \boldsymbol{\Sigma}_0^{-1}) \tag{3.4}$$

with $\nu_0 > (p-1)$ to ensure inversion of $\boldsymbol{\Sigma}$. Meanwhile, we assume $\nu_i = \kappa_i = 1000, i = 1, 2$ for the variance components of σ_ϵ^2 and σ_a^2 to ensure that they are non-informative. This is in contrast to the $\nu_i = \kappa_i = 1, i = 1, 2$ as used in (George & McCulloch, 1997; Ročková & George, 2014) because the form of the inverse gamma distribution we employed here is different from theirs. Similarly, the parameters ν_0 and $\boldsymbol{\Sigma}_0^{-1}$ can be chosen to make it flat and

non-influential. In sum, the hierarchical model formulation for our procedure is given as:

$$
\begin{aligned}
\boldsymbol{\alpha}|\boldsymbol{\gamma} &\sim \mathcal{DE}\left(\boldsymbol{\alpha}|\mathbf{0}, \boldsymbol{\lambda}\right) \\
\boldsymbol{\gamma}|\boldsymbol{\theta}_1 &\sim \text{Bern}\left(\boldsymbol{\gamma}|1, \boldsymbol{\theta}_1\right) \text{ with } \boldsymbol{\theta}_1 \sim \tau\left(\boldsymbol{\theta}_1\right) \\
\boldsymbol{\beta}_f|\boldsymbol{\mu} &\sim \mathcal{DE}\left(\boldsymbol{\beta}|\mathbf{0}, \boldsymbol{\omega}\right) \\
\boldsymbol{\mu}|\boldsymbol{\theta}_2 &\sim \text{Bern}\left(\boldsymbol{\mu}|1, \boldsymbol{\theta}_2\right) \text{ with } \boldsymbol{\theta}_2 \sim \tau\left(\boldsymbol{\theta}_2\right) \\
a_{ijk}|\sigma_a^2 &\sim \mathcal{N}\left(0, \sigma_a^2\right) \\
\sigma_a^2|\nu_2, \kappa_2 &\sim \mathcal{IG}(\nu_2, \kappa_2) \\
\sigma_\epsilon^2|\nu_1, \kappa_1 &\sim \mathcal{IG}(\nu_1, \kappa_1) \\
\boldsymbol{b}|\boldsymbol{\Sigma} &\sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right) \\
\boldsymbol{\Sigma}|\nu_0, \boldsymbol{\Sigma}_0 &\sim \mathcal{IW}(\nu_0, \boldsymbol{\Sigma}_0^{-1})
\end{aligned}
$$

3.3 A Closer Look at the EMMEVS Algorithm

We used EM algorithm for extracting information from the posterior distribution induced by the hierarchical prior formulation described in section 3.2.1. To implement the EM, we treated $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ as missing data and replace them by their conditional expectations given the observed data $\boldsymbol{Y} = (\boldsymbol{W}, \boldsymbol{X})$ and current parameter estimates denoted $\boldsymbol{\Theta}^{(m)}$ - this step is known as the expectation step (E-step). Thereafter, we maximize the expected complete-data log-posterior $log\left(\tau\left(\boldsymbol{\Theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\gamma}, \boldsymbol{\mu}|\boldsymbol{Y}\right)\right)$ with respect to $\boldsymbol{\Theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2$. Specifically, the EM algorithm indirectly maximizes $\tau\left(\boldsymbol{\Theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\boldsymbol{Y}\right)$ by iteratively maximizing the objec-

tive function $Q\left(\boldsymbol{\Theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{\Theta}_1^{(m)}, \boldsymbol{\theta}_1^{(m)}, \boldsymbol{\theta}_2^{(m)}\right)$

$$
\begin{aligned}
Q\left(\cdot\right)^{(m+1)} &= E_{\boldsymbol{\gamma}.\boldsymbol{\mu}|.}\left[log\left(\tau\left(\boldsymbol{\Theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\gamma}, \boldsymbol{\mu} | \boldsymbol{W}, \boldsymbol{X}\right)\right) | \boldsymbol{\Theta}_1^{(m)}, \boldsymbol{\theta}_1^{(m)}, \boldsymbol{\theta}_2^{(m)}, \boldsymbol{W}, \boldsymbol{X}\right] \\
&= \sum_{\boldsymbol{\gamma}} \log\left[\tau\left(\boldsymbol{\Theta}_{\setminus \boldsymbol{\beta}_f, \boldsymbol{\mu}}, \boldsymbol{\theta}_1, \boldsymbol{\gamma} | \boldsymbol{W}\right)\right] \tau\left(\boldsymbol{\gamma} | \boldsymbol{\Theta}_{\setminus \boldsymbol{\beta}_f}^{(m)}, \boldsymbol{\theta}_1^{(m)}\right) \\
&\quad \times \sum_{\boldsymbol{\mu}} \log\left[\tau\left(\boldsymbol{\Theta}_{\setminus \boldsymbol{\alpha}_\ell, \boldsymbol{\gamma}}, \boldsymbol{\theta}_2, \boldsymbol{\mu} | \boldsymbol{X}\right)\right] \tau\left(\boldsymbol{\mu} | \boldsymbol{\Theta}_{\setminus \boldsymbol{\alpha}_\ell}^{(m)}, \boldsymbol{\theta}_2^{(m)}\right)
\end{aligned} \tag{3.5}
$$

where $E_{\boldsymbol{\gamma}.\boldsymbol{\mu}|.}$ represent the conditional expectation $E_{\boldsymbol{\gamma}.\boldsymbol{\mu}} | \boldsymbol{\Theta}^{(m)}, \boldsymbol{\theta}_1^{(m)}, \boldsymbol{\theta}_2^{(m)}, \boldsymbol{W}, \boldsymbol{X}(\cdot)$ and $\boldsymbol{\Theta}_{\setminus M}$ denotes the parameter M is not included in $\boldsymbol{\Theta}$, for instance, $\boldsymbol{\Theta}_{\setminus \sigma_a^2} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{a}, \boldsymbol{b}, \sigma_\epsilon^2, \boldsymbol{\Sigma})$. Assuming a beta prior on $(\theta_{1\ell}, \theta_{2f})$ and using the spike-and-slab hierarchical formulation described earlier, the objective function in (3.5) is of the form

$$
Q\left(\cdot\right)^{(m+1)} = Q_1\left(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(m)}, \boldsymbol{\theta}_1^{(m)}, \boldsymbol{\theta}_2^{(m)}\right) + Q_2\left(\boldsymbol{\theta}_1^{(m)}, \boldsymbol{\theta}_2^{(m)} | \boldsymbol{\Theta}^{(m)}, \boldsymbol{\theta}_1^{(m)}, \boldsymbol{\theta}_2^{(m)}\right) + \text{Constant} \tag{3.6}
$$

where

$$
\begin{aligned}
Q_1\left(\cdot\right)^{(m+1)} \propto & -\frac{n_1 + n_2 + 2(\nu_1 + 1)}{2} \log \sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}\left[\sum_{\ell ijk} \vec{W}_{\ell ijk}^2 + \sum_{fijk} \vec{X}_{fijk}^2 + 2\kappa_1\right] \\
& - \sum_{\ell=1}^{c} \sum_{r=1}^{p} \frac{|\alpha_\ell|}{q_{\ell r}} - \sum_{f=1}^{h} \sum_{r=1}^{p} \frac{|\beta_f|}{q_{fr}} - \frac{q_2 + \nu_0 + q + 1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2\sigma_a^2}\left(\sum_{ijk} a_{ijk}^2 + 2\kappa_2\right) \\
& - \frac{q_1 + 2(\nu_2 + 1)}{2} \log \sigma_a^2 - \frac{1}{2} \boldsymbol{b}' \boldsymbol{\Sigma}^{-1} \boldsymbol{b} - \frac{1}{2} \text{tr}\left(\boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}^{-1}\right) \\
Q_2(\cdot)^{(m+1)} = & \sum_{\ell=1}^{c}\left[\sum_{r=1}^{p} \rho_{\ell r}^{(m)} \log\left(\frac{\theta_{1\ell}^{(m)}}{1 - \theta_{1\ell}^{(m)}}\right) + (l_1 - 1) \log \theta_{1\ell}^{(m)} + (p + o_1 - 1) \log\left(1 - \theta_{1\ell}^{(m)}\right)\right] \\
& + \sum_{f=1}^{h}\left[\sum_{r=1}^{p} \rho_{fr}^{(m)} \log\left(\frac{\theta_{2f}^{(m)}}{1 - \theta_{2f}^{(m)}}\right) + (l_2 - 1) \log \theta_{2f}^{(m)} + (p + o_2 - 1) \log\left(1 - \theta_{2f}^{(m)}\right)\right]
\end{aligned}
$$

where $\vec{W}_{\ell ijk} = W_{\ell ijk} - \boldsymbol{z}_{i\ell}^{'} \boldsymbol{\alpha}_\ell - a_{ijk} - b_{1_\ell ij}$ and $\vec{X}_{fijk} = X_{fijk} - \boldsymbol{z}_{if}^{'} \boldsymbol{\beta}_f - a_{ijk} - b_{2_f ij}$. It is obvious that $Q_2(\cdot)$ corresponds to sum of two beta-binomial priors, one each for binary latent

inclusion indicators $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$. If we assume a uniform prior for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, then we have

$$Q_2\left(\cdot\right)^{(m+1)} = \sum_{\ell=1}^{c}\left[\sum_{r=1}^{p}\rho_{\ell r}^{(m)}\log\left(\frac{\theta_{1\ell}^{(m)}}{1-\theta_{1\ell}^{(m)}}\right) + p\,\log\left(1-\theta_{1\ell}^{(m)}\right)\right] + \sum_{f=1}^{h}\left[\sum_{r=1}^{p}\rho_{fr}^{(m)}\log\left(\frac{\theta_{2f}^{(m)}}{1-\theta_{2f}^{(m)}}\right) + p\,\log\left(1-\theta_{2f}^{(m)}\right)\right]$$

Meanwhile, a closer look at the objective function $Q(\cdot)$ reveals that it has two appealing features that facilitates significant simplification of the EM steps.

1. The separable nature of the objective function (3.5) is due to the following hierarchical structure of the two binary latent variables $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$: $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1) \rightarrow (\boldsymbol{\gamma}, \boldsymbol{\mu}) \rightarrow \boldsymbol{\Theta} \rightarrow (\boldsymbol{W}, \boldsymbol{X})$ such that

$$E_{\boldsymbol{\gamma}\cdot\boldsymbol{\mu}|\cdot.} = E_{\boldsymbol{\gamma}\cdot\boldsymbol{\mu}}|\boldsymbol{\Theta}^{(m)}, \boldsymbol{\theta}_1^{(m)}, \boldsymbol{\theta}_2^{(m)}, \boldsymbol{W}, \boldsymbol{X}^*(\cdot) = E_{\boldsymbol{\gamma}\cdot\boldsymbol{\mu}}|\boldsymbol{\Theta}^{(m)}, \boldsymbol{\theta}_1^{(m)}, \boldsymbol{\theta}_2^{(m)}(\cdot) \tag{3.7}$$

   Thus, the posterior distribution of $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ given the observed data $\boldsymbol{Y}$ and current estimates $\left(\boldsymbol{\Theta}^{(m)}, \boldsymbol{\theta}_1^{(m)}, \boldsymbol{\theta}_2^{(m)}\right)$ depends on $\boldsymbol{Y}$ only through the current estimates $\left(\boldsymbol{\Theta}^{(m)}, \boldsymbol{\theta}_1^{(m)}, \boldsymbol{\theta}_2^{(m)}\right)$.

2. The separability of $Q(\cdot)$ into a pair of distinct functions $Q_1(\cdot)$ and $Q_2(\cdot)$ leads to an M-step that is obtained by maximizing each of these functions separately.

3.3.1 The E-Step

At the E-step, the following conditional expectations $E_{\gamma_r|\cdot}\left(\frac{1}{\lambda_{\ell r}}\right)$, $E_{\mu_r|\cdot}\left(\frac{1}{\omega_{fr}}\right)$, and $E_{\gamma_r|\cdot}\left(\lambda_{\ell r}\right)$, $E_{\mu_r|\cdot}\left(\omega_{fr}\right)$ for $Q_1(\cdot)$ and $Q_2(\cdot)$ respectively, is computed through the application

of the Bayes theorem to (3.7). Thus,

$$\begin{aligned}
E_{\gamma_r|\cdot}\left(\lambda_{\ell r}\right) = \rho_{\ell r}^{(m)} &= Pr\left(\gamma_r = 1|\theta_{1\ell}^{(m)}, \boldsymbol{\alpha}_{\ell}^{(m)}\right) = \frac{e_{1\ell r}}{e_{1\ell r} + e_{2\ell r}} \\
E_{\mu_r|\cdot}\left(\omega_{fr}\right) = \rho_{fr}^{(m)} &= Pr\left(\mu_r = 1|\theta_{2f}^{(m)}, \boldsymbol{\beta}_{f}^{(m)}\right) = \frac{g_{1fr}}{g_{1fr} + g_{2fr}} \\
E_{\gamma_r|\cdot}\left(\frac{1}{\lambda_{\ell r}}\right) = q_{\ell r}^{(m)} &= \frac{\rho_{\ell r}^{(m)}}{\lambda_1} + \frac{1 - \rho_{\ell r}^{(m)}}{\lambda_0} \\
E_{\mu_r|\cdot}\left(\frac{1}{\omega_{fr}}\right) = q_{fr}^{(m)} &= \frac{\rho_{fr}^{(m)}}{\omega_1} + \frac{1 - \rho_{fr}^{(m)}}{\omega_0}
\end{aligned} \tag{3.8}$$

where

$$e_{1\ell r} = \tau\left(\alpha_{\ell}^{(m)}|\boldsymbol{\Theta}_{\setminus\alpha_\ell}^{(m)}, \gamma_r = 1\right) Pr\left(\gamma_r = 1|\theta_{1\ell}^{(m)}\right),\ e_{2\ell r} = \tau\left(\alpha_{\ell}^{(m)}|\boldsymbol{\Theta}_{\setminus\alpha_\ell}^{(m)}, \gamma_r = 0\right) Pr\left(\gamma_r = 0|\theta_{1\ell}^{(m)}\right)$$
$$g_{1fr} = \tau\left(\beta_{f}^{(m)}|\boldsymbol{\Theta}_{\setminus\beta_f}^{(m)}, \mu_r = 1\right) Pr\left(\mu_r = 1|\theta_{2f}^{(m)}\right),\ g_{2fr} = \tau\left(\beta_{f}^{(m)}|\boldsymbol{\Theta}_{\setminus\beta_f}^{(m)}, \mu_r = 0\right) Pr\left(\mu_r = 0|\theta_{2f}^{(m)}\right)$$

Meanwhile, to facilitate the computation of $\rho_{\ell r}^{(m)}$ and $\rho_{fr}^{(m)}$, we take advantage of the conditional independence of the γ_r's and μ_r's given in (3.3), this results in $Pr\left(\gamma_r = 1|\theta_{1\ell}^{(m)}\right) = \theta_{1\ell}^{(m)}$, $Pr\left(\gamma_r = 0|\theta_{1\ell}^{(m)}\right) = 1 - \theta_{1\ell}^{(m)}$, $Pr\left(\mu_r = 1|\theta_{2f}^{(m)}\right) = \theta_{2f}^{(m)}$, and $Pr\left(\mu_r = 0|\theta_{2f}^{(m)}\right) = 1 - \theta_{2f}^{(m)}$.

3.3.2 The M-Step

The maximizer of the expected log-posterior with respect to $\boldsymbol{\Theta}$, $\boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$ is partly facilitated by the separability of the objective function $Q(.)$ given above and partly by the conjugacy of the prior formulation. In what follows, we describe the sequential optimization of the objective function given in equation (3.5).

(a) **M-step update for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$**: With $\boldsymbol{\alpha}^{(m+1)}$ and $\boldsymbol{\beta}^{(m+1)}$ fixed at $\boldsymbol{\alpha}_{\ell}^{(m)}$ and $\boldsymbol{\beta}_{f}^{(m)}$

respectively, the closed form solution of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is found to be

$$\boldsymbol{\theta}_1^{(m+1)} = \begin{bmatrix} \frac{\sum_{r=1}^{p} \rho_{1r}^{(m)} + l_1 - 1}{l_1 + o_1 + p - 2} \\ \vdots \\ \frac{\sum_{r=1}^{p} \rho_{cr}^{(m)} + l_1 - 1}{l_1 + o_1 + p - 2} \end{bmatrix} \text{ and } \boldsymbol{\theta}_2^{(m+1)} = \begin{bmatrix} \frac{\sum_{r=1}^{p} \rho_{1r}^{(m)} + l_2 - 1}{l_2 + o_2 + p - 2} \\ \vdots \\ \frac{\sum_{r=1}^{p} \rho_{hr}^{(m)} + l_2 - 1}{l_2 + o_2 + p - 2} \end{bmatrix} \tag{3.9}$$

(b) **M-step update for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$**: When we fixed $\boldsymbol{\Theta}_{\backslash\boldsymbol{\alpha}}^{(m+1)}$ at $\boldsymbol{\Theta}_{\backslash\boldsymbol{\alpha}}^{(m)}$ and $\boldsymbol{\Theta}_{\backslash\boldsymbol{\beta}}^{(m+1)}$ at $\boldsymbol{\Theta}_{\backslash\boldsymbol{\beta}}^{(m)}$, and let $\boldsymbol{G}_{1\ell} = \boldsymbol{W} - \boldsymbol{a} - \boldsymbol{b}_1$, $\boldsymbol{G}_{2f} = \boldsymbol{X} - \boldsymbol{a} - \boldsymbol{b}_2$, it is easy to recognize that the value of $\boldsymbol{\alpha}^{(m+1)}$ and $\boldsymbol{\beta}^{(m+1)}$ that maximizes $Q_1(.)$ as weighted LASSO problem, hence, the update for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are

$$\begin{aligned} \boldsymbol{\alpha}^{(m+1)} &= \begin{bmatrix} \arg\min_{\boldsymbol{\alpha}_1} \left(\frac{1}{2\sigma_\epsilon^2} \|\boldsymbol{G}_{11} - \boldsymbol{z}_{i\ell}' \boldsymbol{\alpha}_1\|_2^2 + \sum_{r=1}^{p} \frac{|\alpha_1|}{q_{1r}^{(m)}} \right) \\ \vdots \\ \arg\min_{\boldsymbol{\alpha}_c} \left(\frac{1}{2\sigma_\epsilon^2} \|\boldsymbol{G}_{1c} - \boldsymbol{z}_{i\ell}' \boldsymbol{\alpha}_c\|_2^2 + \sum_{r=1}^{p} \frac{|\alpha_c|}{q_{cr}^{(m)}} \right) \end{bmatrix} \\ \boldsymbol{\beta}^{(m+1)} &= \begin{bmatrix} \arg\min_{\boldsymbol{\beta}_1} \left(\frac{1}{2\sigma_\epsilon^2} \|\boldsymbol{G}_{21} - \boldsymbol{z}_{if}' \boldsymbol{\beta}_1\|_2^2 + \sum_{r=1}^{p} \frac{|\beta_1|}{q_{2r}^{(m)}} \right) \\ \vdots \\ \arg\min_{\boldsymbol{\beta}_h} \left(\frac{1}{2\sigma_\epsilon^2} \|\boldsymbol{G}_{2h} - \boldsymbol{z}_{if}' \boldsymbol{\beta}_h\|_2^2 + \sum_{r=1}^{p} \frac{|\beta_h|}{q_{hr}^{(m)}} \right) \end{bmatrix} \end{aligned} \tag{3.10}$$

where $\|\cdot\|^2$ denote the ℓ_2 norm. Many efficient algorithm are available to solve the maximization in (3.10). Some of these algorithms are the least angle regression of Efron et al. (2004), dynamic weighted lasso (DWL) algorithm developed in Chang and Tsay (2010), *glmnet* of Friedman et al. (2010), Wu and Lange (2008) procedure. We investigate the performance of our EMMEVS algorithm and study their behavior through *glmnet* and DWL.

(c) **M-step update for σ_ϵ^2, σ_a^2 and $\boldsymbol{\Sigma}$**: Fixing relevant parameters constant, we need to

solve the following equations in order to obtain the MLE solution of σ_ϵ^2, σ_a^2 and $\boldsymbol{\Sigma}$

$$\begin{aligned}
\sigma_\epsilon^{2(m+1)} &= \arg\min_{\sigma_\epsilon^2} \left(-\frac{c_1}{\sigma_\epsilon^{2(m)}} - c_2 \log \sigma_\epsilon^{2(m)} \right) \\
\sigma_a^{2(m+1)} &= \arg\min_{\sigma_a^2} \left(-\frac{c_4}{\sigma_a^{2(m)}} - c_3 \log \sigma_a^{2(m)} \right) \\
\boldsymbol{\Sigma}^{(m+1)} &= \arg\min_{\boldsymbol{\Sigma}} \left[-c_5 \log|\boldsymbol{\Sigma}^{(m)}| - \frac{1}{2} \boldsymbol{b}' \boldsymbol{\Sigma}^{-1(m)} \boldsymbol{b} - \frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}^{-1(m)} \right) \right]
\end{aligned}$$

where $c_1 = \frac{1}{2}\left[\sum_{\ell ijk} \vec{W}_{\ell ijk}^2 + \sum_{fijk} \vec{X}_{fijk}^2 + 2\kappa_1\right]$, $c_2 = \frac{n_1+n_2+2(\nu_1+1)}{2}$, $c_3 = \frac{q_1+2(\nu_2+1)}{2}$, $c_4 = \frac{\sum_{ijk} a_{ijk}^2 + 2\kappa_2}{2}$, $c_5 = \frac{q_2+\nu_0+q+1}{2}$. The MLE solution of σ_ϵ^2, σ_a^2, and $\boldsymbol{\Sigma}$ are given as

$$\sigma_\epsilon^{2(m+1)} = \frac{c_1}{c_2}, \quad \sigma_a^{2(m+1)} = \frac{c_4}{c_3}, \quad \boldsymbol{\Sigma}^{(m+1)} = \frac{1}{2c_5} \boldsymbol{I}_q \left(\boldsymbol{b}\boldsymbol{b}' + \boldsymbol{\Sigma}_0' \right)' \tag{3.11}$$

(d) **M-step update for a_{ijk} and $\boldsymbol{b}$**: Let $\tilde{W}_{\ell ijk} = W_{\ell ijk} - \boldsymbol{z}_{i\ell}' \boldsymbol{\alpha}_\ell - b_{1_\ell ij}$ and $\tilde{X}_f = X_{fijk} - \boldsymbol{z}_{if}' \boldsymbol{\beta}_f - b_{2_f ij}$. When we fixed $\boldsymbol{\Theta}_{\setminus a_{ijk}}^{(m+1)}$ at $\boldsymbol{\Theta}_{\setminus a_{ijk}}^{(m)}$, extract the part of $Q_1(\cdot)$ that is related to a_{ijk} (denote it as $Q_1(a_{ijk})$). Thereafter, take the partial derivative of $Q_1(a_{ijk})$ with respect to a_{ijk} (i.e., $\frac{\partial Q_1(a_{ijk})}{\partial a_{ijk}}$), set the result equal zero and simplifying, we have

$$a_{ijk}^{(m+1)} = \left(\frac{\sigma_a^{2(m)}}{\sigma_a^{2(m)}(c+h) + \sigma_\epsilon^{2(m)}} \right) \left(\sum_{\ell=1}^{c} \tilde{W}_{\ell ijk} + \sum_{f=1}^{h} \tilde{X}_{fijk} \right) \tag{3.12}$$

Similarly, focusing on the part of $Q_1(\cdot)$ that is related to $\boldsymbol{b} = \left(b_1', b_2'\right)'$ and treat others as constant, we have

$$\boldsymbol{b}^{(m+1)} = \left(\frac{n_{ij}}{\sigma_\epsilon^{2(m)}} \boldsymbol{I}_q + \boldsymbol{\Sigma}^{-1(m)} \right)^{-1} \left(\frac{1}{\sigma_\epsilon^{2(m)}} \right) \sum_{k=1}^{n_{ij}} \mu_{ijk} \tag{3.13}$$

where

$$\sum_{k=1}^{n_{ij}} \mu_{ijk} = \begin{bmatrix} \sum_{k=1}^{n_{ij}} \left(W_{1ijk} - \boldsymbol{z}_{i1}^{'}\boldsymbol{\alpha}_1 - a_{ijk}\right) \\ \vdots \\ \sum_{k=1}^{n_{ij}} \left(X_{hijk} - \boldsymbol{z}_{ih}^{'}\boldsymbol{\beta}_h - a_{ijk}\right) \end{bmatrix}$$

In sum, the goal of the EMMEVS algorithm is to locate the posterior mode of $\log \tau\left(\boldsymbol{\Theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{W}, \boldsymbol{X}\right)$. The idea is to of treat $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ as "missing data" and solve the complete-data log posterior, $\log \tau\left(\boldsymbol{\Theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\gamma}, \boldsymbol{\mu} | \boldsymbol{W}, \boldsymbol{X}\right)$ using EM approach. Thus, at the E-Step, we replace the unobservable $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ by their conditional expectations given $\boldsymbol{y}$ and $\boldsymbol{\Theta}^{(m)}, \boldsymbol{\theta}_1^{(m)}, \boldsymbol{\theta}_2^{(m)}$ and at the M-step, we maximize the expected complete-data log-posterior with respect to $\boldsymbol{\Theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2$. The EMMEVS algorithm proceed iteratively between the E and M steps, generates a sequence of parameter estimates which under regularity condition converges monotonically toward a local maximum of $\tau\left(\boldsymbol{\Theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \boldsymbol{y}\right)$. The EMMEVS algorithm is presented in Algorithm 1.

Algorithm 1 EMMEVS Algorithm

Initialize $\boldsymbol{\Theta}$, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ with fixed $\lambda_0,\ \omega_0,\ \lambda_1,\ \omega_1,\ l_1,\ l_2,\ o_1,\ o_2$.
E-step computation: Compute the E-step using equation (3.8)
Set $m = 0$.
while $m \leq M$ **do**
    update $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ using equations (3.9)
    update $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using equations (3.10)
    update σ_ϵ^2, σ_a^2, and $\boldsymbol{\Sigma}$ using (3.11)
    update $\boldsymbol{a}$ and $\boldsymbol{b}$ using equations (3.12) and (3.13)
  Set $m \leftarrow m + 1$.
end while
Stopping Criterion $\|\boldsymbol{\alpha}^{(m+1)} - \boldsymbol{\alpha}^{(m)}\| \leq 10^{-5}$ and $\|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}\| \leq 10^{-5}$

3.3.3 Ideas Behind EMMEVS Implementation

As noted earlier, EMMEVS employs MAP idea to select sparse model. This section lays out how EMMEVS algorithm performs variable selection. The variable selection is done in two specific appealing ways.

(1) *Thresholding Rule*: One of the procedures by which EMMEVS operates to carry out variable selection is by thresholding. Following from the hierarchical formulation in section 3.2.1, the idea here is that the modes for the parameters in the continuous endpoints $\tau\left(\boldsymbol{\Theta}_{\setminus\boldsymbol{\beta},\boldsymbol{\theta}_2},\boldsymbol{\theta}_1|\boldsymbol{W}\right)$ and the corresponding parameters for the discrete endpoints $\tau\left(\boldsymbol{\Theta}_{\setminus\boldsymbol{\alpha},\boldsymbol{\theta}_1},\boldsymbol{\theta}_2|\boldsymbol{X}\right)$ can be found deterministically. Thereafter, the associated modes of $\tau\left(\boldsymbol{\gamma}|\boldsymbol{W}\right)$ and $\tau\left(\boldsymbol{\mu}|\boldsymbol{X}\right)$ are obtained by thresholding rule. Specifically, once we obtain the posterior modes (MAP estimates) i.e., $\widehat{\boldsymbol{\Theta}},\widehat{\boldsymbol{\theta}}_1,\widehat{\boldsymbol{\theta}}_2$, we can find the most probable $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\mu}}$ given $\left(\widehat{\boldsymbol{\Theta}},\widehat{\boldsymbol{\theta}}_1,\widehat{\boldsymbol{\theta}}_2\right)$. Expanding on the univariate linear regression framework of Ročková and George (2014), the thresholding for our model occurs at the intersection of $\pm\alpha_\ell^*\left(\lambda_0,\lambda_1,\widehat{\theta}_{1\ell}\right)$ and $\pm\beta_f^*\left(\omega_0,\omega_1,\widehat{\theta}_{2f}\right)$ of the $Pr\left(\gamma_r=1|\widehat{\boldsymbol{\alpha}}_\ell,\widehat{\theta}_{1\ell}\right)$ and $Pr\left(\mu_r=1|\widehat{\boldsymbol{\beta}}_f,\widehat{\theta}_{2f}\right)$ weighted mixture of the spike-and-slab priors, namely

$$\alpha_\ell^*\left(\cdot\right)=\pm\sqrt{\left(2\lambda_0\log\left(d_\ell f_1\right)f_1^2\right)/\left(f_1^2-1\right)}\text{ and }\beta_f^*\left(\cdot\right)=\pm\sqrt{\left(2\omega_0\log\left(d_f f_2\right)f_2^2\right)/\left(f_2^2-1\right)}$$

where $f_1^2=\lambda_1/\lambda_0,\ d_\ell=\frac{1-\Pr\left(\gamma_r=1|\widehat{\theta}_{1\ell}\right)}{\Pr\left(\gamma_r=1|\widehat{\theta}_{1\ell}\right)},\ f_2^2=\omega_1/\omega_0,$ and $d_f=\frac{1-\Pr\left(\mu_r=1|\widehat{\theta}_{2f}\right)}{\Pr\left(\mu_r=1|\widehat{\theta}_{2f}\right)}$. Hence, the thresholding rule is

$$\widehat{\gamma}_r=\begin{cases}1 & \text{if } |\widehat{\alpha}_{\ell r}|\geq\alpha_\ell^*\left(\lambda_0,\lambda_1,\widehat{\theta}_1\right)\\ 0 & \text{otherwise}\end{cases};\quad \widehat{\mu}_r=\begin{cases}1 & \text{if } |\widehat{\beta}_{fr}|\geq\beta_f^*\left(\omega_0,\omega_1,\widehat{\theta}_2\right)\\ 0 & \text{otherwise}\end{cases}$$

(2) *Dynamic Posterior Exploration*: This is the second feature of the EMMEVS algorithm that makes it appealing. Here, rather than restricting attention to a single value

of (λ_0, ω_0), the computational speed of the EMMEVS algorithm makes it feasible to run the algorithm over a sequence of (λ_0, ω_0) to estimate the modes of a range of different posteriors. This is unlike the MCMC procedure developed in Chapter 3 which expends considerable computational effort sampling from a single posterior, the dynamic posterior exploration approach provides a snapshot of the several different posteriors. This feature of EMMEVS has practical significance that it helps the user to visualize the results and identify variables that should be included in the model.

The EMMEVS algorithm achieve this through regularization plot such as the LASSO regularization diagram of Hastie et al. (2009). The plot captures the evolution of the modal estimates as well as the model configurations and their posterior probabilities over a sequence of spike-and-slab Laplace mixture of priors with increasing $\lambda_0 > 0$ and $\omega_0 > 0$. As $\lambda_0 > 0$ and $\omega_0 > 0$ increases, the negligible coefficients are more and more absorbed in the spike part of the mixture, there by reducing the posterior multimodality and exposing sparse high-probability submodels for thresholding identification. Contrary to shrinkage estimators such as the LASSO or the Ridge, as λ_0 increases, the EMMEVS does not shrink the large coefficients to zero too much when negligible coefficients are getting closer to zero. In Section 3.5.1, we illustrate the implementation of the regularization plot using synthetic data.

3.4 Simulation Study

This section presents an artificial data to evaluate the performance of the procedures developed in this chapter. The data is generated as follows;

3.4.1 Data Generation

Clustered Mixed Endpoints Design

In this section, we describe a simulation study for illustrating the methods proposed in this dissertation and to evaluate their performances and behaviors. The simulated data are consistent with data that are commonly encountered in practice. Specifically, our simulated data include four dose levels and a control group. The dose levels are spaced equally between the control dose and the maximum dose. In particular, we set the dose levels to $0, 0.25, 0.50, 0.75, \text{and}, 1.0$. In addition to the five distinct dose levels, we assigned p-dimensional predictors, $\boldsymbol{z}_{ik}^{*}$, to each of the kth pup. The $\boldsymbol{z}_{ik}^{*}$ is generated from $\mathcal{N}_p(\mathbf{0}, \Theta)$ where $\Theta = \left(0.6^{|i-j|}\right)_{i,j=1}^{p}$. We assume that there are $c = 3$ continuous endpoints and $h = 2$ binary endpoints. Further, we assume the ℓth continuous outcomes are decreasing linearly with dose while the fth clustered binary outcomes are increasing linearly with dose and that the dose are constant over time (see Figure 5). To fit the joint model in (2.2), we set

$$\begin{aligned}
\alpha_{01} + \alpha_{11} d_i &= 1.5 - 2d_i \\
\alpha_{02} + \alpha_{12} d_i &= 1 - 0.5 d_i \\
\alpha_{03} + \alpha_{13} d_i &= 0.5 - 1.5 d_i \\
\beta_{01} + \beta_{11} d_i &= 5 + 3 d_i \\
\beta_{02} + \beta_{12} d_i &= 2 + 1.5 d_i
\end{aligned} \tag{3.14}$$

$\sigma_\epsilon^2 = 0.01$, $\sigma_a^2 = 0.1$, $a_{ijk} \sim \mathcal{N}(0, \sigma_a^2)$, $\epsilon_{\ell ijk} \sim \mathcal{N}(0, 0.02)$, $\epsilon_{fijk} \sim \mathcal{N}(0, 0.04)$. The dependence of the five endpoints is characterized by a multivariate normal distribution of $\boldsymbol{b} = (\boldsymbol{b}_{1_1 ij}, \boldsymbol{b}_{1_2 ij}, \boldsymbol{b}_{1_3 ij}, \boldsymbol{b}_{2_1 ij}, \boldsymbol{b}_{2_2 ij})' \sim \mathcal{N}(\mathbf{0}, \Sigma)$ with $\Sigma \sim \mathcal{IW}\left(\nu_0, \Sigma_0^{-1}\right)$. We consider a balanced design experiment consisting of equal number of dams per dose group and equal number of pups per dam. Thus, we fixed the number of pups in each dam to 10 and the number of

dams per dose group was set to 15. This corresponds to study sizes $(c+h)\times 5\times 15\times 10 = 3750$, $c \times 5 \times 15 \times 10 = 2250$ for the continuous endpoints and $h \times 5 \times 15 \times 10 = 1500$ for the discrete endpoints. Lastly, for the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we set their values to $\boldsymbol{\alpha} = \boldsymbol{\beta} = 1, 2, 3, 4$ and the remaining coefficients $(4 < r \leq p)$ were set to 0, such that only the first four coefficients are significant. The data is then generated according to (2.2).

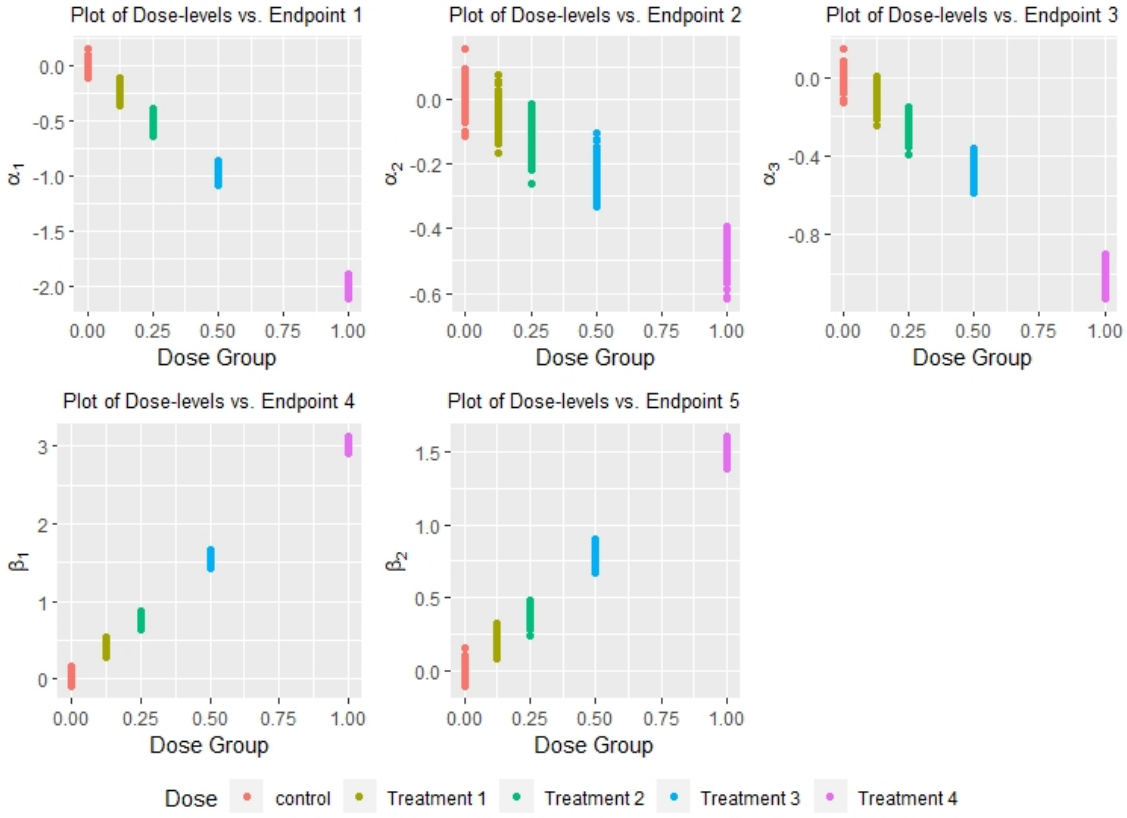

Figure 5. Simulated Data showing Trend

3.4.2 Simulation Results

To illustrate the performance of the EMMEVS algorithm, we applied the simulated data described in Section 3.4.1 using the spike-and-slab mixture priors given in (3.1) with fixed slab parameter at $\lambda_1 = \omega_1 = 100$ and a single value of λ_0 and ω_0 fix at 0.5. Further, we set $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} = \mathbf{1}_p$, $\sigma_\epsilon^{2(0)} = 0.01$, $\sigma_a^{2(0)} = 1$, $\boldsymbol{\Sigma} = \boldsymbol{I}_q$. The hyper-parameters ν_1, ν_2, κ_1,

and κ_2 are set to 1000 and $\nu_0 = p + 1$. Lastly, following the suggestion in Castillo and van der Vaart (2012), we set $a_1 = a_2 = 1$, $b_1 = b_2 = p$ to avoid putting an informative prior on $\theta_{1\ell}$ and θ_{2f} in order to obtain optimal posterior concentration rates.

The modal coefficient estimates for $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ are depicted in Figure 6 implemented for $p = 100$ using the DWL algorithm. The algorithm converges in less than 3 seconds with the number of iterations to convergence ranging between 3 and 7 iterations. It is obvious from Figure 6 that our algorithm is able to separate the significant coefficients from non-significant coefficients. Overall, we observed the powerful strength of the EMMEVS algorithm to constantly identify the true significant variables and drag to near zero (or exactly zero) those variables that are set to be unimportant.

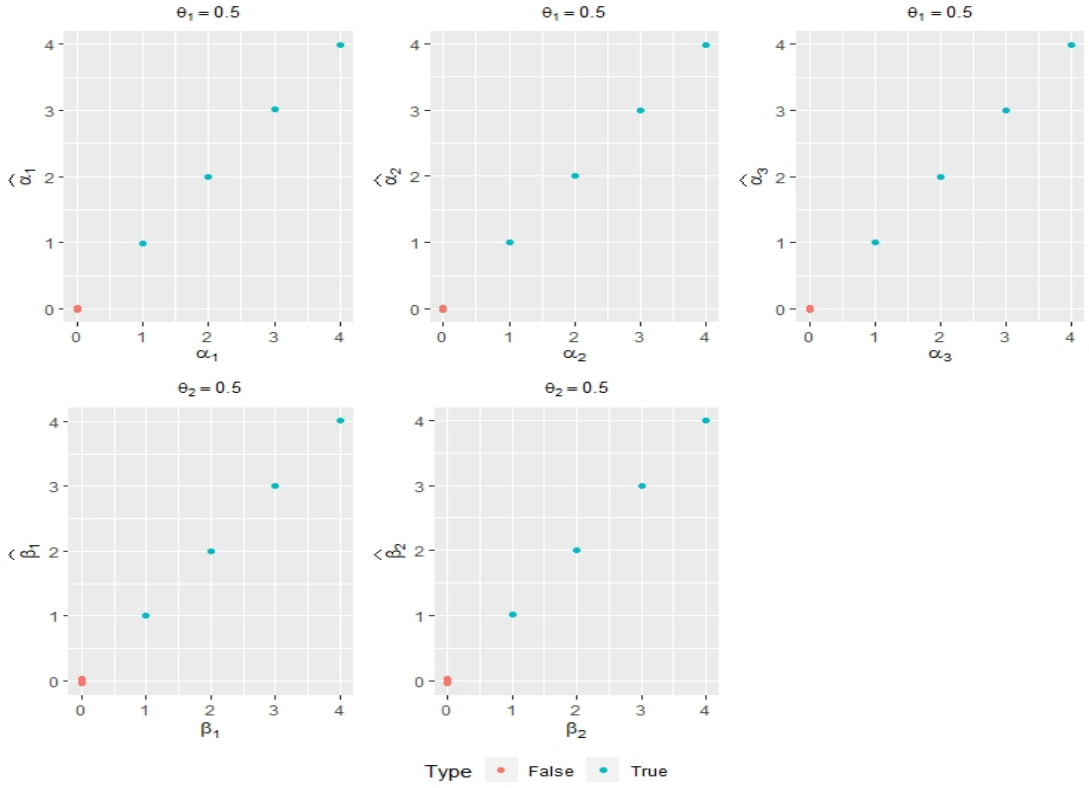

Figure 6. Plot of the true coefficients of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ against their MAP Estimates $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ for $p = 100$ using *glmnet* package in R.

3.5 Deterministic Annealing Variants of the EMMEVS Algorithm

We will start this section with a quotation from Wikipedia regarding the description of annealing as used in the steel industry.

> "Annealing is a heat treatment wherein a material is altered, causing changes in its properties such as strength and hardness. It is a process that produces conditions by heating to above the re-crystallization temperature and maintaining a suitable temperature, and then cooling." Wikipedia

The EM algorithm is one of the techniques used to find the ML estimates of parameters of interest. Although the method guarantees monotonical convergence towards at least a local maximum; there is no assurance that the estimates will converge to the global mode (because the estimates are prone to entrapment in local maximum mode), this is one of the potential drawbacks of the algorithm. Besides, the performance of the EM algorithm is dependent and sensitive to the starting values used to initialize the iterations in the algorithm. Several approaches have been suggested to reduce the dependence of the algorithm on the starting values. One such popular method recommended by McLachlan and Basford (2004) is to run the algorithm for various choice of the starting values or mitigate the entrapment to local modes. Another approach to mitigate this issue (an approach we use in this study) and which can further improve the chances of finding a global mode is the use of deterministic annealing variant of EM algorithm (DAEM) suggested by Ueda and Nakano (1998).

The deterministic annealing variant of EMMEVS referred to as DAEMMEVS, like the DAEM, uses the principle of entropy to redefine the objective function given in (3.5) with the aim of minimizing the so-called free-energy function at gradually cooler temperatures.

In our context, this is equivalent to maximizing the negative free-energy function

$$\begin{aligned} -\mathcal{F}_t\left(\boldsymbol{\Theta}, \theta_{1\ell}, \theta_{2f}\right) &= \underset{\text{internal energy}}{\mathcal{U}_t\left(\boldsymbol{\Theta}, \theta_{1\ell}, \theta_{2f}\right)} + \frac{1}{t} \underset{\text{entropy}}{\mathcal{S}_t\left(\boldsymbol{\Theta}, \theta_{1\ell}, \theta_{2f}\right)} \\ &= \sum_{\boldsymbol{\gamma}} \log\left[\tau\left(\boldsymbol{\Theta}_{\setminus \boldsymbol{\beta}_f, \boldsymbol{\mu}}, \theta_{1\ell}, \boldsymbol{\gamma} | \boldsymbol{W}\right)\right]^t \sum_{\boldsymbol{\mu}} \log\left[\tau\left(\boldsymbol{\Theta}_{\setminus \boldsymbol{\alpha}_\ell, \boldsymbol{\gamma}}, \theta_{2f}, \boldsymbol{\mu} | \boldsymbol{X}\right)\right]^t, \end{aligned} \tag{3.15}$$

for $0 < t \leq 1$, where

$$\begin{aligned} \mathcal{U}_t\left(\boldsymbol{\Theta}, \theta_{1\ell}, \theta_{2f}\right) = \sum_{\boldsymbol{\gamma}} \log\left[\tau\left(\boldsymbol{\Theta}_{\setminus \boldsymbol{\beta}_f, \boldsymbol{\mu}}, \theta_{1\ell}, \boldsymbol{\gamma} | \boldsymbol{W}\right)\right]^t \tau\left(\boldsymbol{\gamma} | \boldsymbol{\Theta}_{\setminus \boldsymbol{\beta}_f}^{(m)}, \theta_1^{(m)}\right)^t \\ \times \sum_{\boldsymbol{\mu}} \log\left[\tau\left(\boldsymbol{\Theta}_{\setminus \boldsymbol{\alpha}_\ell, \boldsymbol{\gamma}}, \theta_{2f}, \boldsymbol{\mu} | \boldsymbol{X}\right)\right] \tau\left(\boldsymbol{\mu} | \boldsymbol{\Theta}_{\setminus \boldsymbol{\alpha}_\ell}^{(m)}, \theta_{2f}^{(m)}\right)^t \end{aligned}$$

is the internal energy and $\mathcal{S}_t\left(\boldsymbol{\Theta}, \theta_{1\ell}, \theta_{2f}\right)$ is the entropy. In equation (3.15), $\frac{1}{t}$ acts as the temperature of the annealing process that regulates the degree of separation between the multiple modes of $\mathcal{F}_t\left(\boldsymbol{\Theta}, \theta_{1\ell}, \theta_{2f}\right)$. In practice, the annealing process starts with a high temperature (at t close to 0). At this high temperature, the landscape of $-\mathcal{F}_t\left(\boldsymbol{\Theta}, \theta_1, \theta_2\right)$ is smooth, this therefore prevents the algorithm from getting stuck in a local mode early in its iterations. And as the temperature cools down (at t close to 1), the effect of the inclusion posterior is strengthened. As a result, local modes begin to appear and the landscape of $-\mathcal{F}_t\left(\boldsymbol{\Theta}, \theta_{1\ell}, \theta_{2f}\right)$ progressively approaches the true, incomplete posterior. In fact, equation (3.15) embeds the actual log incomplete posterior as a special case when $t = 1$.

We formulate the deterministic annealing variant of EMMEVS by introducing an annealing loop within the algorithm. Doing this does not change the M-step; however, it changes the E-step which now requires the computation of the expected complete log posterior density with respect to a modified posterior distribution. Following Ročková and George (2014), the tempered probabilities of inclusion can be estimated by

$$\rho_{\ell r}^t = \frac{e_{1\ell r}^t}{e_{1\ell r}^t + e_{2\ell r}^t}, \text{ and } \rho_{fr}^t = \frac{g_{1fr}^t}{g_{1fr}^t + g_{2fr}^t} \tag{3.16}$$

where

$$e_{1\ell r} = \tau\left(\alpha_\ell^{(m)}|\boldsymbol{\Theta}_{\setminus\alpha_\ell}^{(m)}, \gamma_r = 1\right) Pr\left(\gamma_r = 1|\theta_{1\ell}^{(m)}\right),\ e_{2\ell r} = \tau\left(\alpha_\ell^{(m)}|\boldsymbol{\Theta}_{\setminus\alpha_\ell}^{(m)}, \gamma_r = 0\right) Pr\left(\gamma_r = 0|\theta_{1\ell}^{(m)}\right)$$

$$g_{1fr} = \tau\left(\beta_f^{(m)}|\boldsymbol{\Theta}_{\setminus\beta_f}^{(m)}, \mu_r = 1\right) Pr\left(\mu_r = 1|\theta_{2f}^{(m)}\right),\ g_{2fr} = \tau\left(\beta_f^{(m)}|\boldsymbol{\Theta}_{\setminus\beta_f}^{(m)}, \mu_r = 0\right) Pr\left(\mu_r = 0|\theta_{2f}^{(m)}\right)$$

The DAEMMEVS, is obtained by replacing $\rho_{\ell r}$ and ρ_{fr} given in (3.8) with $\rho_{\ell r}^t$ and ρ_{fr}^t respectively. The DAEMMEVS procedure is specifically described in Algorithm 3.

Algorithm 2 DAEMMEVS Algorithm

Initialize $\boldsymbol{\Theta}$, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ with fixed $\lambda_0,\ \omega_0,\ \lambda_1,\ \omega_1,\ l_1,\ l_2,\ o_1,\ o_2$

EM-step computation: Compute the EM step at the current temperature, t, until $\|\boldsymbol{\alpha}^{(m+1)} - \boldsymbol{\alpha}^{(m)}\| \leq 10^{-5}$ and $\|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}\| \leq 10^{-5}$

Set $m = 0$.

while $m \leq M$ **do**

(a) E-step: evaluate $\rho_{\ell r}^t$ and ρ_{fr}^t

(b) M-step: compute $\boldsymbol{\Theta}^{(m+1)}$, $\boldsymbol{\theta}_1^{(m+1)}$ and $\boldsymbol{\theta}_2^{(m+1)}$

(c) Set $m \leftarrow m + 1$.

end while

If $t < 1$, return to step 2, on each return, use the previous estimates to initiate the algorithm at the current temperature t. The algorithm stops at $t = 1$.

The strategy here is that at each cooling step, t, we find a global mode that is used to initiate the algorithm at the next temperature, thereby finding a new global mode. This strategy increases the probability of convergence to the true global mode if the new global mode is close to the previous one. While convergence at the global mode is still not guaranteed, the deterministic annealing approach removes the algorithm's dependence on the starting values and finds the global mode more often than the conventional EM algorithm (Ueda & Nakano, 1998).

3.5.1 Sensitivity Analysis

In this section, we investigate the sensitivity of the EMMEVS and its deterministic annealing variants at varying temperatures. We carry out the performance evaluation through the following: (1) regularization plots and (2) assessment of recovery support and estimation performance. For our illustration, we consider the same data set described in Section 3.4.1 to implement the tempered version of the EMMEVS algorithm and use the starting values for the parameters and hyper-parameters described in Section 3.5. We consider grid of λ_0 and ω_0 values, we ran these values from 0.001 to 0.05 equally divided into 20 values and we choose as the optimum model, the model that yield the largest log-posterior among the sequences of the scale parameters. The results of this is depicted in Figures 7, 8, 9, 10, and 11 at temperatures $t = 1, 5, 10, \text{and}, 20$ using DWL scheme.

Figures 7, 8, 9, 10, and 11 showed that both the EMMEVS and DAEMMEVS consistently identified the fixed covariates and shrunk redundant variables towards zero. For each of the Figure 7 to 11, we observed that the red dots (representing the irrelevant variables) are thresholded towards zero and that the trajectories of individual regression coefficients estimates appeared to stabilize relatively early in the path; this would mean that the parameter estimates had clearly segregated into groups of zero and non-zero values. Specifically, we observed that the DWL implementation of the EMMEVS and DAEMMEVS algorithm stabilizes very early in its trajectory and almost immediately identify the significant variables as early as the first computation compared to the ADL (adaptive LASSO implemented via the glmnet package of Friedman et al. (2010)) that consistently delay this till after about 10 computations. This is a consequence of the annealing process and the capability of the DWL to rapidly compute the solution by borrowing information from previous iterations when the regularization parameters $(\lambda_0, \lambda_1, \omega_0, \omega_1)$ change across the EM iterations.

The stabilization we observed in Figures 7, 8, 9, 10, and 11 allows us to focus and

report a single value of λ_0 and ω_0 out of the many that were computed without the need for cross-validation. The following analogy (reported in Deshpande et al. (2019)) of comparing the dynamic posterior exploration pre-stabilization to focusing a camera lens should bring home the points we are trying to make regarding the behavior of the regularization plots presented in Figures 7 to 11: "starting from a blurry image, turning the focus ring slowly brings an image into relief, with the salient features becoming increasingly prominent. In this way, the priors serve more as filters for the data likelihood than as encapsulations of any real subjective beliefs."

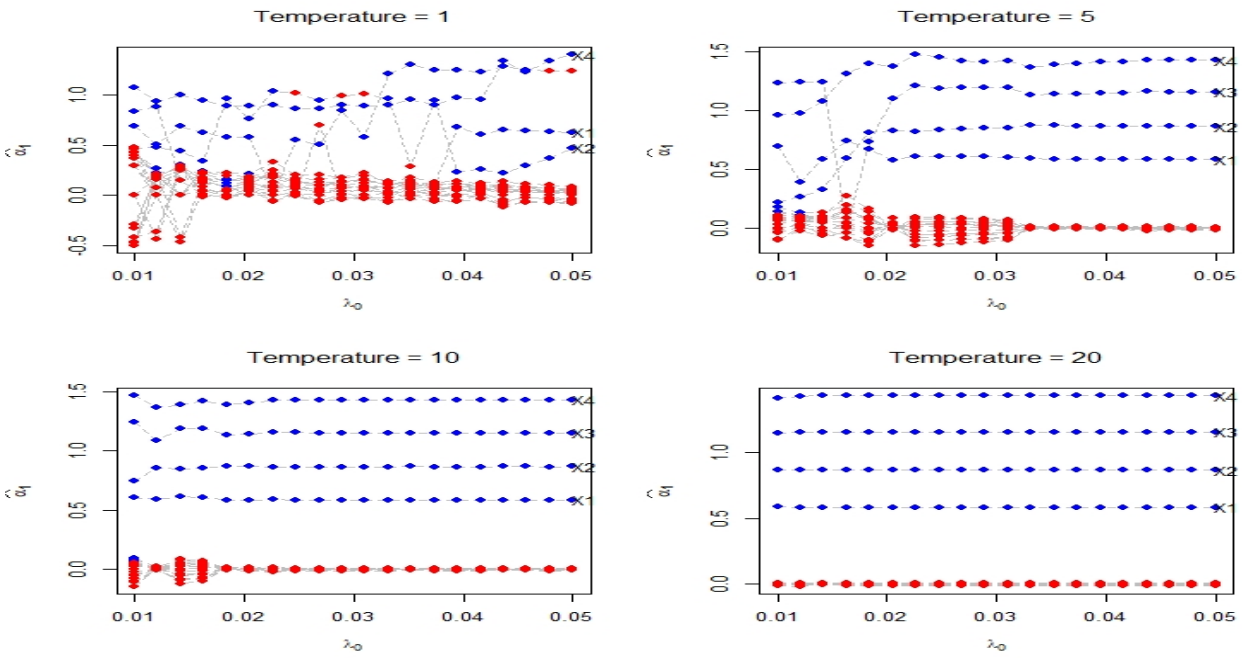

Figure 7. Plots for the trajectories of the regression coefficients $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ computed for varying choices λ_0 and ω_0. The blue dots corresponds to the variables that are in the slab component while the red dots are for variables in the spike component of the Laplace mixture.

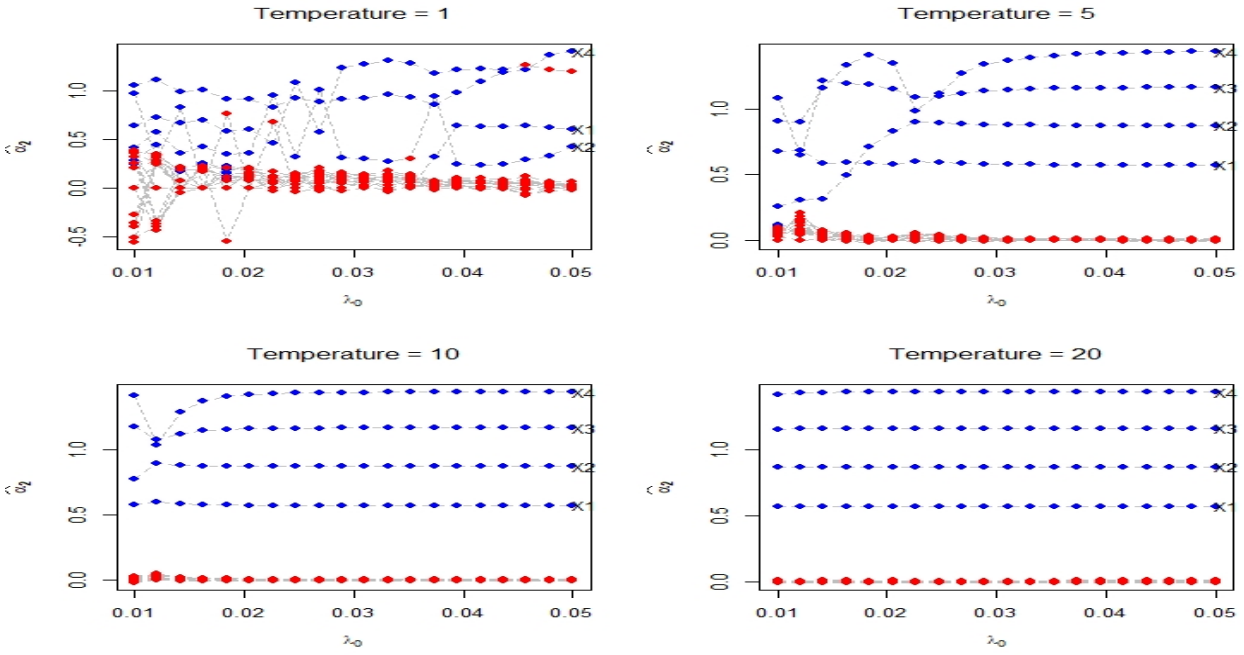

Figure 8. Plots for the trajectories of the regression coefficients $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ computed for varying choices λ_0 and ω_0. The blue dots corresponds to the variables that are in the slab component while the red dots are for variables in the spike component of the Laplace mixture.

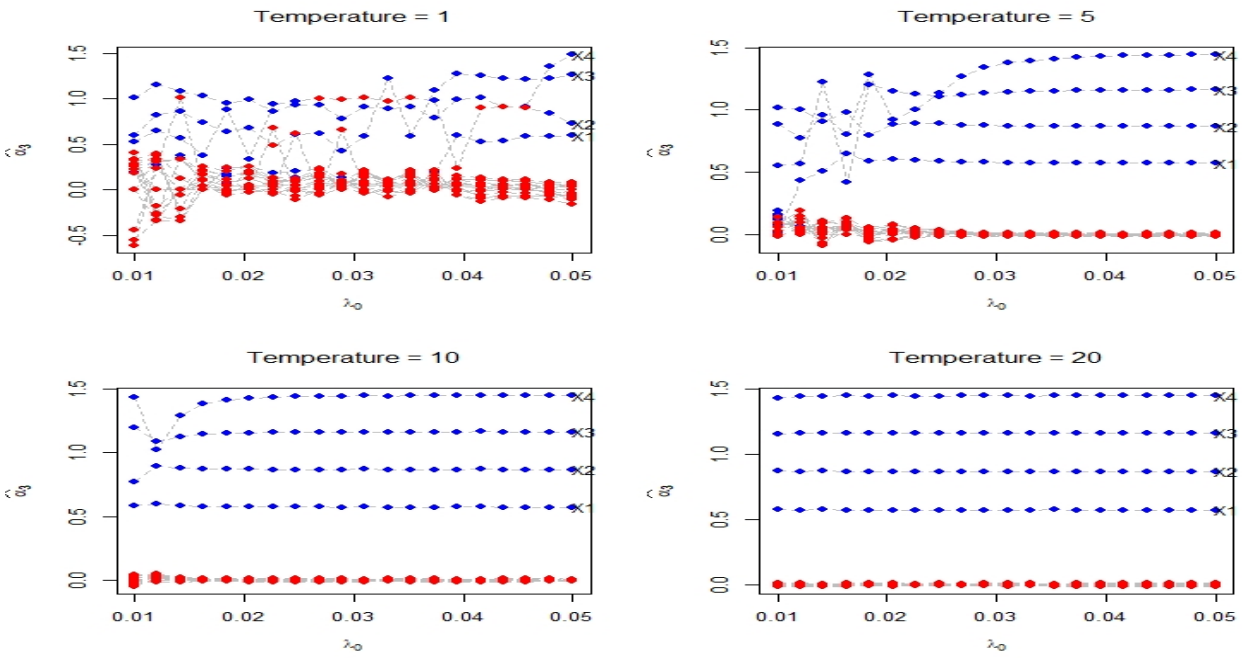

Figure 9. Plots for the trajectories of the regression coefficients $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ computed for varying choices λ_0 and ω_0. The blue dots corresponds to the variables that are in the slab component while the red dots are for variables in the spike component of the Laplace mixture.

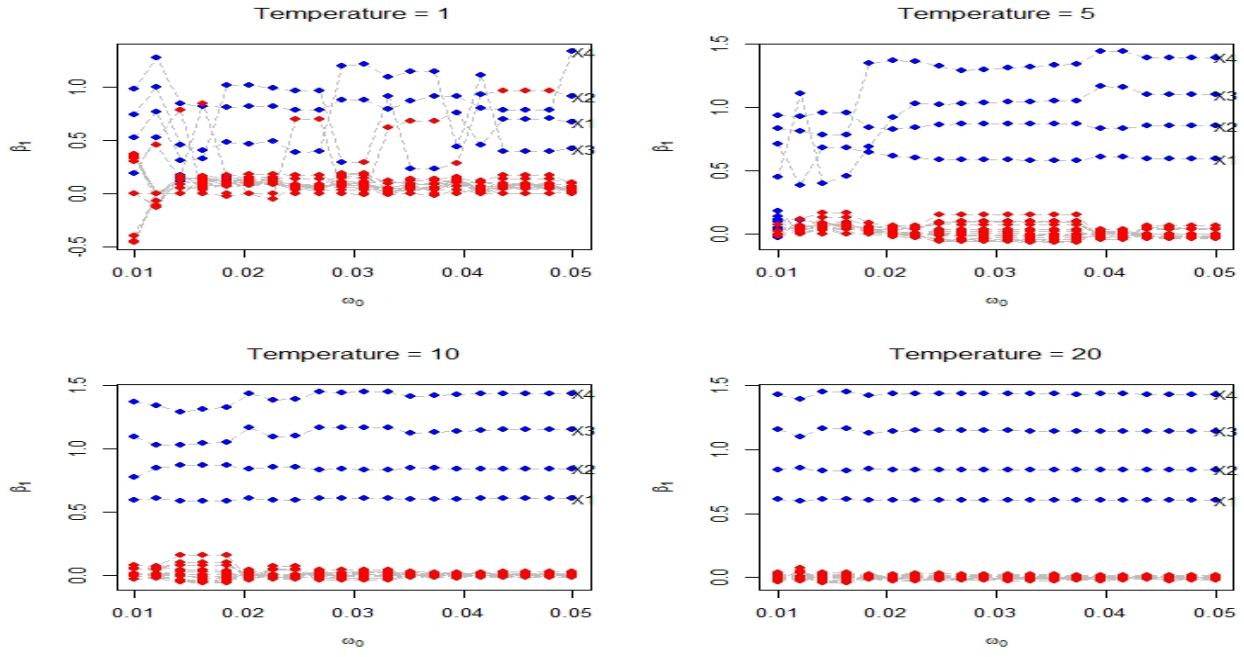

Figure 10. Plots for the trajectories of the regression coefficients $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ computed for varying choices λ_0 and ω_0. The blue dots corresponds to the variables that are in the slab component while the red dots are for variables in the spike component of the Laplace mixture.

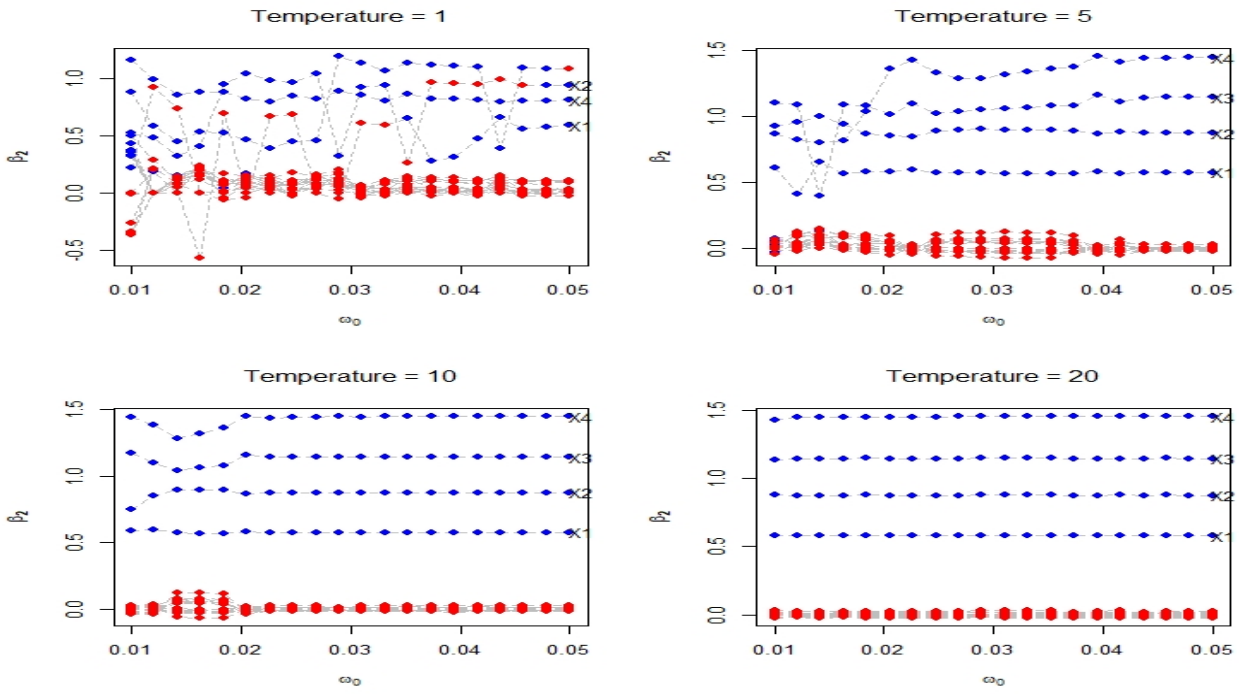

Figure 11. Plots for the trajectories of the regression coefficients $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ computed for varying choices λ_0 and ω_0. The blue dots corresponds to the variables that are in the slab component while the red dots are for variables in the spike component of the Laplace mixture.

To further explore the full potential of DAEMMEVS and the impact of starting values that are far away from the true coefficient vector, we consider two more randomly generated starting vectors $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim \mathcal{N}_p\left(\mathbf{0}, 3 \times \boldsymbol{I}_p\right)$ and $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim \mathcal{N}_p\left(\mathbf{0}, 5 \times \boldsymbol{I}_p\right)$ independently. For exposition, we set $\lambda_0 = \omega_0 = (0.1, 0.4, 0.7, 1)$ and applied both EMMEVS and DAEMMEVS at temperatures $t = 5, 10, \text{and}, 20$. To assess the recovery support and estimation performance, we tracked several quantities like numbers of iteration to convergence (ITER), number of times a true model is selected (FTM), Bias, mean squared error (MSE), execution time in seconds (TIME). In addition, we evaluate the sensitivity and specificity of our

procedure. Let

$$\begin{aligned}
\text{false positive rate (FPR)} &= \frac{FP}{FP+TN} \\
\text{false negative rate (FNR)} &= \frac{FN}{FN+TP} \\
\text{sensitivity (SEN)} &= \frac{TP}{TP+TN} \\
\text{specificity (SPE)} &= \frac{TN}{TN+FP} \\
\text{balanced accuracy rate (BAR)} &= \frac{SPE+SEN}{2} \\
\text{accuracy (ACC)} &= \frac{TP+TN}{TP+TN+FP+FN} \\
\text{Mathew's correlation coefficient (MCC)} &= \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}
\end{aligned}$$

where TP(TN) and FP(FN) are the total number of true positives (negatives) and false positives (negatives) identification made in the support recovery respectively. The MCC is a measure that shows the quality of the classification; its values ranges between -1 (indicates complete disagreement between the observed and predicted classification) and $+1$ (indicates complete agreement between the observed and predicted classification). The BAR is a comprehensive performance metric that combines both the sensitivity and the specificity of a classifier.

We collect the results of these analysis in Tables 2, 3, 4, and 5. We observe from Tables 2 to 5 that depending on the choice of the starting vectors $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ the bias and MSE of the EMMEVS algorithm converges to a different solution. In contrast, at higher temperatures, the DAEMMEVS converges to the same or approximate values of the bias and the MSE even for distant values. We also notice a downward decrease in the bias and MSE as the temperature increases. We also observed a similar trend when we look at other performance metrics (such as MCC, ACC, FTM, and BAR) to investigate the sensitivity of our procedure to identifying the true model. It is therefore evident from the Tables 2,

3, 4, and 5 that tempering act to reduce the posterior multimodality and gravitate smaller coefficient estimates towards zero.

Table 2. Simulation study for the average variable selection performance of EMMEVS algorithm when $p = 20$ using 100 repetitions: MSE (average mean squared error), BIAS (average bias), FTM (average number of true models detected), ACC (average accuracy), MCC (average Mathew's correlation coefficient), BAR (balance accuracy rate). The MSE has been re-scaled by a factor of 100 while the BIAS was re-scaled by a factor of 10000.

Temperature	Endpoints	$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} = \mathbf{1}_p$						$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim N_p(\mathbf{0}, I)$						$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim N_p(\mathbf{0}, 9 \times I)$						$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim N_p(\mathbf{0}, 25 \times I)$					
		BIAS	MSE	FTM	ACC	MCC	BAR	BIAS	MSE	FTM	ACC	MCC	BAR	BIAS	MSE	FTM	ACC	MCC	BAR	BIAS	MSE	FTM	ACC	MCC	BAR
1	$\boldsymbol{\alpha}_1$	-135.75	198.44	1.00	67.82	38.25	75.16	-162.48	187.77	2.00	63.41	34.30	73.72	-159.64	182.03	2.00	60.33	29.95	70.79	-152.21	199.72	1.00	65.33	36.38	74.35
	$\boldsymbol{\alpha}_2$	-141.66	197.36	3.00	66.60	36.87	74.39	-166.39	177.16	1.00	63.69	33.59	73.70	-156.84	182.38	0.00	59.18	28.46	70.44	-147.50	207.05	3.00	64.18	35.29	74.19
	$\boldsymbol{\alpha}_3$	-142.72	202.68	2.00	63.12	32.97	72.70	-165.68	185.67	0.00	61.97	31.88	73.00	-163.70	196.65	2.00	62.88	33.54	73.01	-158.37	188.11	2.00	63.33	33.65	72.92
	$\boldsymbol{\beta}_1$	-107.05	211.19	7.00	72.90	47.18	79.53	-168.72	191.48	2.00	61.42	31.61	72.67	-150.05	206.97	3.00	63.06	33.67	73.77	-136.12	214.29	3.00	62.84	33.61	72.54
	$\boldsymbol{\beta}_2$	-101.57	210.50	3.00	71.03	44.57	78.66	-159.39	187.87	1.00	62.05	32.47	72.60	-142.20	205.05	2.00	60.00	30.17	71.05	-125.85	184.30	1.00	60.70	31.25	71.76
2	$\boldsymbol{\alpha}_1$	-6.80	16.59	66.00	97.07	92.28	97.79	-10.36	46.61	61.00	96.02	89.94	97.32	-6.23	24.91	53.00	96.02	89.35	97.04	-6.19	21.51	64.00	97.27	92.48	97.73
	$\boldsymbol{\alpha}_2$	-8.21	19.37	92.00	99.52	98.39	99.13	-11.05	40.64	77.00	97.97	94.39	98.35	-8.79	30.43	81.00	98.43	95.73	98.73	-7.94	27.54	91.00	99.51	98.42	99.22
	$\boldsymbol{\alpha}_3$	-8.77	26.10	93.00	99.56	98.62	99.25	-9.80	43.24	57.00	96.20	90.15	97.53	-11.24	41.86	73.00	97.92	94.11	98.03	-8.06	33.44	86.00	99.26	97.67	99.06
	$\boldsymbol{\beta}_1$	-12.51	32.63	82.00	98.92	96.68	98.75	-10.13	40.22	59.00	95.70	89.17	97.31	-10.56	37.51	73.00	97.96	94.40	98.25	-9.27	33.88	77.00	98.43	95.39	98.26
	$\boldsymbol{\beta}_2$	-10.06	30.44	93.00	99.37	98.16	99.13	-14.44	54.88	71.00	97.40	93.24	98.19	-11.46	42.06	84.00	99.01	97.11	99.09	-10.95	37.06	82.00	98.72	96.26	98.63
3	$\boldsymbol{\alpha}_1$	-2.36	3.13	92.00	99.60	98.81	99.75	-1.89	4.75	98.00	99.90	99.70	99.94	-1.35	4.78	98.00	99.90	99.70	99.94	-2.99	4.11	96.00	99.70	99.20	99.81
	$\boldsymbol{\alpha}_2$	-2.73	1.26	100.00	100.00	100.00	100.00	-2.38	1.65	99.00	99.95	99.85	99.97	-2.74	4.67	100.00	100.00	100.00	100.00	-2.61	4.77	100.00	100.00	100.00	100.00
	$\boldsymbol{\alpha}_3$	-2.61	1.39	100.00	100.00	100.00	100.00	-2.54	1.62	100.00	100.00	100.00	100.00	-2.66	1.57	100.00	100.00	100.00	100.00	-3.11	1.58	100.00	100.00	100.00	100.00
	$\boldsymbol{\beta}_1$	-0.15	3.95	100.00	100.00	100.00	100.00	1.00	3.03	100.00	100.00	100.00	100.00	-6.01	5.22	97.00	99.80	99.44	99.88	-3.21	5.40	100.00	100.00	100.00	100.00
	$\boldsymbol{\beta}_2$	-2.38	3.78	99.00	99.95	99.85	99.97	-0.96	3.76	98.00	99.90	99.70	99.94	-2.81	5.19	98.00	99.90	99.70	99.94	-2.81	5.47	100.00	100.00	100.00	100.00
4	$\boldsymbol{\alpha}_1$	0.05	0.06	100.00	100.00	100.00	100.00	0.34	0.06	100.00	100.00	100.00	100.00	-0.08	0.07	100.00	100.00	100.00	100.00	-0.18	0.07	100.00	100.00	100.00	100.00
	$\boldsymbol{\alpha}_2$	-0.31	0.06	100.00	100.00	100.00	100.00	-0.45	0.06	100.00	100.00	100.00	100.00	-0.57	0.06	100.00	100.00	100.00	100.00	-0.25	0.06	100.00	100.00	100.00	100.00
	$\boldsymbol{\alpha}_3$	-0.12	0.06	100.00	100.00	100.00	100.00	0.07	0.06	100.00	100.00	100.00	100.00	-0.43	0.06	100.00	100.00	100.00	100.00	-0.22	0.06	100.00	100.00	100.00	100.00
	$\boldsymbol{\beta}_1$	-0.62	0.88	100.00	100.00	100.00	100.00	1.05	0.98	100.00	100.00	100.00	100.00	0.71	1.31	100.00	100.00	100.00	100.00	-0.56	1.32	100.00	100.00	100.00	100.00
	$\boldsymbol{\beta}_2$	-0.18	0.06	100.00	100.00	100.00	100.00	-0.40	0.06	100.00	100.00	100.00	100.00	0.19	0.06	100.00	100.00	100.00	100.00	-0.19	0.06	100.00	100.00	100.00	100.00

Table 3. Simulation study for the average variable selection performance of EMMEVS algorithm when $p = 50$ using 100 repetitions: MSE (average mean squared error), BIAS (average bias), FTM (average number of true models detected), ACC (average accuracy), MCC (average Mathew's correlation coefficient), BAR (balanced accuracy rate). The MSE has been re-scaled by a factor of 100 while the BIAS was re-scaled by a factor of 10000.

Temperature	Endpoints	$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} = \mathbf{1}_p$						$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim N_p(\mathbf{0}, I)$						$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim N_p(\mathbf{0}, 9 \times I)$						$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim N_p(\mathbf{0}, 25 \times I)$					
		BIAS	MSE	FTM	ACC	MCC	BAR	BIAS	MSE	FTM	ACC	MCC	BAR	BIAS	MSE	FTM	ACC	MCC	BAR	BIAS	MSE	FTM	ACC	MCC	BAR
1	$\boldsymbol{\alpha}_1$	-57.28	1219.78	4.00	79.54	45.09	81.35	-53.07	741.15	0.00	53.79	16.27	66.34	-52.62	830.07	0.00	59.41	21.47	69.73	-59.55	919.43	3.00	63.36	27.01	73.59
	$\boldsymbol{\alpha}_2$	-54.45	1205.11	2.00	69.41	31.75	76.42	-51.31	741.21	0.00	56.73	19.37	68.73	-53.56	853.24	0.00	58.52	21.19	70.73	-58.66	954.55	0.00	61.55	23.38	72.83
	$\boldsymbol{\alpha}_3$	-54.86	1238.30	0.00	68.67	30.03	75.90	-50.23	774.13	0.00	57.08	19.35	69.15	-50.63	877.72	0.00	60.24	22.42	71.44	-55.86	1005.88	0.00	63.67	25.07	72.61
	$\boldsymbol{\beta}_1$	-47.49	1936.16	2.00	80.27	44.32	83.79	-51.99	954.40	0.00	55.79	18.00	67.65	-54.68	1352.10	0.00	64.27	26.31	74.88	-57.05	1637.11	0.00	68.15	30.58	78.47
	$\boldsymbol{\beta}_2$	-46.59	2067.50	2.00	78.45	42.20	82.81	-50.05	953.34	0.00	57.85	20.03	68.31	-52.22	1357.61	0.00	66.64	28.26	75.60	-52.92	1670.32	1.00	69.57	31.25	78.10
2	$\boldsymbol{\alpha}_1$	-1.54	40.69	64.00	98.21	91.39	98.45	-0.98	57.84	59.00	98.30	91.63	98.73	-1.59	61.10	66.00	98.72	93.21	98.85	-2.10	53.15	59.00	98.11	90.47	98.28
	$\boldsymbol{\alpha}_2$	-0.70	46.54	94.00	99.86	99.04	99.47	-0.42	62.96	76.00	99.30	95.89	99.16	-2.87	67.41	92.00	99.76	98.58	99.64	-2.23	54.23	92.00	99.84	99.01	99.80
	$\boldsymbol{\alpha}_3$	-0.42	68.18	97.00	99.94	99.55	99.63	-0.96	64.41	43.00	96.60	84.58	97.01	-2.39	81.45	77.00	99.02	95.06	99.01	-1.71	71.52	86.00	99.64	97.83	99.46
	$\boldsymbol{\beta}_1$	-2.21	52.24	84.00	99.54	97.50	99.52	-2.76	72.56	63.00	98.68	93.03	98.83	-1.62	67.19	76.00	99.26	95.83	99.26	-2.47	70.68	76.00	99.30	96.11	99.39
	$\boldsymbol{\beta}_2$	-2.12	53.86	92.00	99.82	98.80	99.33	-1.59	69.55	79.00	99.50	97.01	99.39	-2.79	79.34	75.00	99.40	96.33	98.99	-3.53	77.64	93.00	99.86	99.07	99.58
3	$\boldsymbol{\alpha}_1$	-0.53	5.25	81.00	99.50	97.19	99.73	-0.34	10.19	87.00	99.68	98.17	99.83	-0.31	9.99	91.00	99.78	98.74	99.88	-0.71	10.83	91.00	99.80	98.83	99.89
	$\boldsymbol{\alpha}_2$	-0.29	1.44	100.00	100.00	100.00	100.00	-0.22	2.45	92.00	99.82	98.95	99.90	-0.98	8.77	99.00	99.98	99.88	99.99	-0.72	12.52	99.00	99.98	99.88	99.99
	$\boldsymbol{\alpha}_3$	-0.29	1.59	100.00	100.00	100.00	100.00	-0.03	2.40	100.00	100.00	100.00	100.00	-0.36	2.13	100.00	100.00	100.00	100.00	-0.66	2.42	100.00	100.00	100.00	100.00
	$\boldsymbol{\beta}_1$	0.26	8.32	96.00	99.92	99.52	99.96	-0.40	8.88	97.00	99.94	99.64	99.97	-0.15	12.60	93.00	99.84	99.07	99.91	-0.34	14.34	93.00	99.86	99.16	99.92
	$\boldsymbol{\beta}_2$	-0.30	7.72	96.00	99.92	99.52	99.96	-1.66	9.58	99.00	99.98	99.88	99.99	-0.85	11.74	97.00	99.94	99.64	99.97	0.00	12.86	95.00	99.90	99.40	99.95
4	$\boldsymbol{\alpha}_1$	0.08	0.17	100.00	100.00	100.00	100.00	-0.08	0.18	100.00	100.00	100.00	100.00	-0.05	0.17	100.00	100.00	100.00	100.00	-0.01	0.17	100.00	100.00	100.00	100.00
	$\boldsymbol{\alpha}_2$	0.10	0.16	100.00	100.00	100.00	100.00	-0.22	0.15	100.00	100.00	100.00	100.00	-0.12	0.16	100.00	100.00	100.00	100.00	-0.02	0.16	100.00	100.00	100.00	100.00
	$\boldsymbol{\alpha}_3$	0.12	0.15	100.00	100.00	100.00	100.00	-0.01	0.15	100.00	100.00	100.00	100.00	-0.09	0.15	100.00	100.00	100.00	100.00	0.02	0.16	100.00	100.00	100.00	100.00
	$\boldsymbol{\beta}_1$	-0.43	2.70	100.00	100.00	100.00	100.00	0.52	2.77	100.00	100.00	100.00	100.00	0.24	3.83	100.00	100.00	100.00	100.00	-0.81	3.55	100.00	100.00	100.00	100.00
	$\boldsymbol{\beta}_2$	-0.13	0.17	100.00	100.00	100.00	100.00	-0.02	0.18	100.00	100.00	100.00	100.00	-0.13	0.18	100.00	100.00	100.00	100.00	0.04	0.17	100.00	100.00	100.00	100.00

Table 4. Simulation study for the average variable selection performance of EMMEVS algorithm when $p = 100$ using 100 repetitions: MSE (average mean squared error), BIAS (average bias), FTM (average number of true models detected), ACC (average accuracy), MCC (average Mathew's correlation coefficient), BAR (balanced accuracy rate). The MSE has been re-scaled by a factor of 100 while the BIAS was re-scaled by a factor of 10000.

Temperature	Endpoints	$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} = \mathbf{1}_p$						$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim N_p(\mathbf{0}, I)$						$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim N_p(\mathbf{0}, 9 \times I)$						$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim N_p(\mathbf{0}, 25 \times I)$					
		BIAS	MSE	FTM	ACC	MCC	BAR	BIAS	MSE	FTM	ACC	MCC	BAR	BIAS	MSE	FTM	ACC	MCC	BAR	BIAS	MSE	FTM	ACC	MCC	BAR
1	$\boldsymbol{\alpha}_1$	-0.20	9772.04	1.00	87.76	42.10	86.08	-21.14	6252.80	0.00	68.79	18.80	72.36	-19.02	8609.77	0.00	76.96	24.67	77.22	-21.30	8968.26	0.00	80.16	29.37	80.80
	$\boldsymbol{\alpha}_2$	-0.18	10286.59	0.00	87.59	40.80	86.58	-19.28	6687.38	0.00	69.65	20.84	75.33	-16.81	9303.40	0.00	81.30	30.92	81.63	-20.91	9523.92	0.00	82.34	32.76	83.37
	$\boldsymbol{\alpha}_3$	-0.17	10374.77	0.00	87.99	40.85	86.31	-16.54	6623.88	0.00	70.24	22.43	77.44	-15.72	9551.39	0.00	83.68	35.11	84.31	-19.68	9771.44	0.00	84.59	35.86	84.43
	$\boldsymbol{\beta}_1$	-0.13	10565.65	0.00	88.61	44.50	88.67	-18.01	6915.19	0.00	70.44	21.37	76.34	-17.33	9096.29	0.00	81.13	31.44	83.23	-16.84	9516.13	0.00	83.94	35.98	85.88
	$\boldsymbol{\beta}_2$	-0.12	11355.56	0.00	90.56	47.25	88.61	-14.68	7784.59	0.00	73.96	23.95	77.93	-12.97	10203.50	1.00	85.68	37.97	85.59	-15.32	10929.46	0.00	87.40	41.71	87.33
2	$\boldsymbol{\alpha}_1$	0.00	587.90	44.00	98.21	84.85	95.71	-0.57	120.71	43.00	98.19	85.16	97.62	-0.88	383.95	37.00	97.59	81.62	96.59	-1.91	475.39	28.00	97.71	81.27	96.65
	$\boldsymbol{\alpha}_2$	0.00	730.23	79.00	99.73	96.29	97.22	-0.90	120.85	68.00	99.25	93.06	98.05	-0.81	504.39	67.00	99.24	92.33	97.45	-1.12	616.78	64.00	99.34	92.64	96.66
	$\boldsymbol{\alpha}_3$	0.00	977.78	62.00	99.17	92.17	96.21	-0.36	138.04	39.00	97.42	80.59	96.38	-1.55	889.73	37.00	97.90	82.75	96.03	-1.83	972.25	38.00	97.92	82.13	94.72
	$\boldsymbol{\beta}_1$	0.00	1655.84	72.00	99.59	95.24	97.87	-0.07	93.19	55.00	98.78	88.45	97.21	-0.36	649.68	41.00	98.66	86.57	96.42	-1.01	1525.09	42.00	98.80	88.13	97.10
	$\boldsymbol{\beta}_2$	0.01	2266.87	79.00	99.74	96.50	97.47	-1.23	121.82	62.00	99.38	92.93	97.76	-0.56	1517.35	36.00	98.60	85.16	95.44	-1.15	2637.24	33.00	98.56	84.78	95.77
3	$\boldsymbol{\alpha}_1$	0.00	10.03	47.00	98.95	90.17	99.45	-0.28	23.84	81.00	99.59	96.24	99.79	-0.01	24.80	80.00	99.52	96.04	99.75	-0.81	24.33	80.00	99.58	96.36	99.78
	$\boldsymbol{\alpha}_2$	0.00	1.92	98.00	99.98	99.77	99.99	-0.29	6.25	86.00	99.78	97.84	99.89	-0.22	17.04	94.00	99.89	98.94	99.94	-1.03	22.01	95.00	99.93	99.29	99.96
	$\boldsymbol{\alpha}_3$	0.00	2.56	98.00	99.94	99.49	99.97	-0.02	4.13	95.00	99.87	98.85	99.93	-0.15	8.01	100.00	100.00	100.00	100.00	-0.32	10.16	100.00	100.00	100.00	100.00
	$\boldsymbol{\beta}_1$	0.00	20.00	100.00	100.00	100.00	100.00	-0.88	25.83	87.00	99.83	98.25	99.91	0.12	25.62	77.00	99.66	96.56	99.82	0.47	24.67	69.00	99.57	95.58	99.78
	$\boldsymbol{\beta}_2$	0.00	21.56	87.00	99.78	97.84	99.89	-0.26	25.31	84.00	99.73	97.36	99.86	-0.38	22.09	79.00	99.64	96.43	99.81	0.19	18.62	73.00	99.60	95.93	99.79
4	$\boldsymbol{\alpha}_1$	0.00	0.38	100.00	100.00	100.00	100.00	0.01	0.39	100.00	100.00	100.00	100.00	-0.04	0.38	100.00	100.00	100.00	100.00	0.00	0.39	100.00	100.00	100.00	100.00
	$\boldsymbol{\alpha}_2$	0.00	0.34	100.00	100.00	100.00	100.00	-0.04	0.33	100.00	100.00	100.00	100.00	0.02	0.35	100.00	100.00	100.00	100.00	0.00	0.35	100.00	100.00	100.00	100.00
	$\boldsymbol{\alpha}_3$	0.00	0.34	100.00	100.00	100.00	100.00	0.02	0.33	100.00	100.00	100.00	100.00	0.03	0.36	100.00	100.00	100.00	100.00	-0.06	0.35	100.00	100.00	100.00	100.00
	$\boldsymbol{\beta}_1$	0.00	7.59	100.00	100.00	100.00	100.00	0.18	8.48	100.00	100.00	100.00	100.00	0.57	11.16	100.00	100.00	100.00	100.00	-0.08	11.07	100.00	100.00	100.00	100.00
	$\boldsymbol{\beta}_2$	0.00	0.37	100.00	100.00	100.00	100.00	0.00	0.38	100.00	100.00	100.00	100.00	-0.04	0.37	100.00	100.00	100.00	100.00	-0.18	0.37	100.00	100.00	100.00	100.00

Table 5. Simulation study for the average variable selection performance of EMMEVS algorithm when $p = 200$ using 100 repetitions: MSE (average mean squared error), BIAS (average bias), FTM (average number of true models detected), ACC (average accuracy), MCC (average Mathew's correlation coefficient), BAR (balanced accuracy rate). The MSE has been re-scaled by a factor of 100 while the BIAS was re-scaled by a factor of 10000.

Temperature	Endpoints	$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} = \mathbf{1}_p$						$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim N_p(\mathbf{0}, I)$						$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim N_p(\mathbf{0}, 9 \times I)$						$\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} \sim N_p(\mathbf{0}, 25 \times I)$					
		BIAS	MSE	FTM	ACC	MCC	BAR	BIAS	MSE	FTM	ACC	MCC	BAR	BIAS	MSE	FTM	ACC	MCC	BAR	BIAS	MSE	FTM	ACC	MCC	BAR
1	$\boldsymbol{\alpha}_1$	-4.99	54569.13	3.00	93.41	44.77	87.21	-3.67	32077.69	0.00	73.31	15.18	74.26	-3.31	50261.30	0.00	86.06	25.66	82.11	0.24	51398.29	0.00	86.37	25.28	81.54
	$\boldsymbol{\alpha}_2$	-4.29	60201.17	2.00	92.35	40.12	87.53	-4.30	28684.30	0.00	68.04	13.36	72.80	-3.52	56145.78	0.00	85.35	25.69	83.47	1.10	58653.87	0.00	87.20	27.38	83.79
	$\boldsymbol{\alpha}_3$	-3.72	62260.86	1.00	91.59	37.83	87.99	-2.59	23920.66	0.00	63.18	11.60	70.68	-2.94	58007.38	0.00	85.63	26.11	83.85	1.64	60150.30	0.00	87.91	28.90	84.53
	$\boldsymbol{\beta}_1$	-4.53	57537.72	0.00	93.51	42.29	86.52	-1.17	36530.07	0.00	75.38	16.54	75.81	-0.77	49907.72	0.00	84.53	23.47	80.96	-0.21	51868.92	0.00	86.03	25.60	82.71
	$\boldsymbol{\beta}_2$	-3.00	59540.13	0.00	91.73	36.98	87.46	-1.71	31491.37	0.00	69.87	14.05	73.12	-2.46	51665.37	0.00	85.06	24.50	81.85	0.05	55094.51	0.00	86.48	26.76	83.18
2	$\boldsymbol{\alpha}_1$	0.81	28186.75	12.00	96.63	59.29	87.14	-0.15	3038.24	0.00	85.89	32.67	82.64	1.00	19546.60	2.00	94.68	46.56	85.29	1.52	22900.82	3.00	95.50	52.34	86.56
	$\boldsymbol{\alpha}_2$	1.14	35508.10	43.00	99.56	87.65	90.34	-0.17	4910.61	4.00	86.75	35.99	82.22	1.46	28849.13	17.00	98.60	71.85	88.02	2.23	29839.19	17.00	98.82	75.23	89.35
	$\boldsymbol{\alpha}_3$	1.06	40488.97	27.00	99.03	79.91	89.10	-0.54	5927.49	1.00	80.39	24.44	77.62	1.71	35718.24	12.00	97.74	66.16	89.30	2.95	35943.10	13.00	97.74	66.13	88.80
	$\boldsymbol{\beta}_1$	-0.21	38292.09	24.00	99.03	79.40	88.73	-0.60	2729.11	2.00	87.62	38.02	83.15	0.42	30101.75	4.00	97.67	60.37	86.68	-1.53	34637.61	9.00	98.10	66.67	88.13
	$\boldsymbol{\beta}_2$	0.24	40359.36	25.00	99.27	81.80	88.73	-0.15	5501.73	2.00	89.47	41.25	84.22	1.05	37024.72	9.00	98.21	66.72	88.67	-2.00	39645.65	13.00	98.38	69.38	88.52
3	$\boldsymbol{\alpha}_1$	0.24	58.79	64.00	98.77	86.35	99.37	0.12	71.13	80.00	99.18	91.80	99.58	0.01	64.36	63.00	98.49	85.80	99.23	-0.14	54.52	51.00	97.73	80.57	98.84
	$\boldsymbol{\alpha}_2$	0.00	3.14	100.00	100.00	100.00	100.00	-0.07	8.16	76.00	98.97	89.32	99.47	0.24	49.96	98.00	99.89	98.97	99.94	-0.08	63.29	99.00	99.93	99.45	99.96
	$\boldsymbol{\alpha}_3$	-0.03	4.96	95.00	99.64	97.37	99.82	-0.01	18.78	37.00	96.92	71.59	98.43	-0.18	94.48	96.00	99.81	98.19	99.90	-0.12	92.67	99.00	99.97	99.62	99.98
	$\boldsymbol{\beta}_1$	-0.04	48.59	64.00	99.50	91.97	99.74	-0.15	57.25	57.00	99.41	90.71	99.70	-0.17	58.21	63.00	99.49	91.97	99.74	0.29	239.85	66.00	99.64	93.85	99.82
	$\boldsymbol{\beta}_2$	-0.08	27.88	49.00	97.51	76.34	98.73	-0.10	41.51	55.00	97.84	79.46	98.90	-0.22	95.87	86.00	99.42	94.42	99.70	-0.22	1604.30	84.00	99.30	93.20	99.64
4	$\boldsymbol{\alpha}_1$	0.00	0.89	100.00	100.00	100.00	100.00	0.01	0.90	100.00	100.00	100.00	100.00	-0.03	0.90	100.00	100.00	100.00	100.00	0.00	0.92	100.00	100.00	100.00	100.00
	$\boldsymbol{\alpha}_2$	0.01	0.79	100.00	100.00	100.00	100.00	-0.02	0.81	100.00	100.00	100.00	100.00	0.02	0.83	100.00	100.00	100.00	100.00	0.00	0.82	100.00	100.00	100.00	100.00
	$\boldsymbol{\alpha}_3$	-0.02	0.79	100.00	100.00	100.00	100.00	0.01	0.82	100.00	100.00	100.00	100.00	0.00	0.81	100.00	100.00	100.00	100.00	-0.02	0.82	100.00	100.00	100.00	100.00
	$\boldsymbol{\beta}_1$	0.18	23.69	67.00	99.77	95.34	99.88	-0.01	42.82	71.00	99.80	95.93	99.90	-0.06	38.73	85.00	99.89	97.78	99.94	-0.18	33.72	85.00	99.92	98.20	99.96
	$\boldsymbol{\beta}_2$	-0.01	0.90	100.00	100.00	100.00	100.00	-0.01	0.89	100.00	100.00	100.00	100.00	-0.02	0.90	100.00	100.00	100.00	100.00	-0.05	0.87	100.00	100.00	100.00	100.00

Discussion and Chapter Summary

Motivated by the practical applications in clinical, medical, and behavioral studies as well as toxicology and psychometric among others, we propose a deterministic shrinkage procedure to select relevant covariates and conduct parameter estimation simultaneously. The method extends the univariate linear regression solution of Ročková and George (2014) to the complex multivariate (clustered or longitudinal) mixed outcomes with high-dimensional covariates.

We deployed our proposed EMMEVS algorithm and its deterministic annealing variant within a path-following scheme to identify the modes of several posterior distributions, corresponding to different choices of λ_0 and ω_0 from the spike distributions. In contrast to MCMC procedures which attempts to characterize a single posterior, our usage of the dynamic posterior exploration enables us to report the modes of several posterior distributions based on the grid of λ_0 and ω_0 from the spike distribution. We take advantage of some of the available algorithm (such as DWL of Chang and Tsay (2010) and glmnet of Friedman et al. (2010)) to obtain a computationally efficient scheme which is scalable to high-dimensional data.

We demonstrate the advantage of our procedure in terms of variable selection, prediction, and computational scalability via extensive simulation study. The results obtained from the simulation revealed that the modal estimates identified by our dynamic posterior exploration stabilized rapidly very early in their trajectories (especially with our implementation of DWL scheme and as the temperature increases); thus, allowing us to report a single estimate out of the many we computed without the need for cross-validation.

While there is no general guarantee that these trajectories will stabilize, figures like Figures 7, 8, 9, 10, and 11 provide a useful self-check: if one observes stabilization in the supports of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and in the log-posterior, one can safely report the final mode identified.

On the other hand, if the modal estimates have not stabilized, one can simply add larger values of λ_0 and ω_0 to the ladders and continue exploring. Although our focused in this dissertation was on the simplest setting where the $\boldsymbol{\gamma}$'s and $\boldsymbol{\mu}$'s are treated as exchangeable, it is not difficult to incorporate more thoughtful structured sparsity within our framework. For example, if the predictors displayed a known grouping structure, we could introduce several parameters, one for each group, with little additional computational overhead

Chapter 4

Real Data Analyses

4.1 The HELP Study Data Set

4.1.1 Background on the HELP study

Some patients with identified alcohol dependency, drug addiction, and substance abuse problems are fortunate to undergo detoxification programs by the federal/state governments or through their primary care physicians. However, despite an apparent need for medical services, many of these patients without primary medical care (PMC) do not receive adequate medical care (D'Aunno, 1997; Saitz et al., 1997). For this population of patients without PMC, there is a high chance of relapse due to other mental and emotional issues that go with substance abuse (cocaine, heroin) and alcohol dependency. To reduce relapse, patients are either linked to existing medical services or received treatment at the substance abuse treatment site. The practicality of linking patients with substance abuse problems to an off-site PMC has been implemented. Such implementation was carried out in the HELP (Health Evaluation and Linkage to Primary Care) study.

The HELP study was a clinical trial for adult inpatients recruited from a detoxification unit. Patients with no primary care physician were randomized to receive a multidisciplinary assessment and a brief motivational intervention or usual care, to link them to PMC. Eligible subjects were adults, who spoke Spanish or English, reported alcohol, heroin, or cocaine as their first or second drug of choice, and either resided in proximity to the primary care clinic to which they would be referred or were homeless. Patients with established primary care relationships they planned to continue, significant dementia, specific plans to leave the Boston area that would prevent research participation, failure to provide contact information

for tracking purposes, or pregnancy were excluded. Subjects were interviewed at baseline during their detoxification stay, and follow-up interviews were undertaken every six months for two years. A variety of continuous, count, discrete, and survival time predictors and outcomes were collected at each of these five occasions. The details of the randomized trials is described in Samet et al. (2003). The analyses carried out here are intended to illustrate our proposed procedures, a comprehensive study of the data is planned for future.

Meanwhile, since all the patients cannot be linked to the PMC, the physician have to decide which patient should be connected and who should not. To help the physician in making this decision, we are using the data from the HELP study to identify the risk factors for those that may be linked to the PMC using the novel methodology developed in this dissertation. In our analyses, we assembled five endpoints (two continuous and three discrete as presented in Table 6). There were 39 covariates including the time factor.

Table 6. Description of endpoints extracted from the HELP study data. Note that we used the cut-off point of to dichotomized the cesd scale. We computed the range for pcs and mcs from the baseline visit.

Variables	Description	Type	Range
pcs	SF-36 Physical Component Score	continuous	14.07 - 74.81
mcs	SF-36 Mental Component Score	continuous	6.76 - 62.18
cesd	Center for Epidemiological Studies Depression Scale	discrete	0-1
g1b	Experienced serious thoughts of suicide (last 30 days)	discrete	0-1
pcrec	Number of primary care visits last 6 months	discrete	0-2

4.1.2 Results - HELP Study

In this section, using the HELP study data described above, we apply the EMMEVS procedure to identify the risk factors to help physician decide on whether to link a patient to PMC or not after undergoing detoxification. In the analyses, we set $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} = \mathbf{1}_p$, $\sigma_\epsilon^{2(0)} = 0.1$, $\sigma_a^{2(0)} = 0.9$, $\boldsymbol{\Sigma} = \boldsymbol{I}_q$. The hyper-parameters ν_1, ν_2, κ_1, and κ_2 are set to 10000 and $\nu_0 = p + 1$. We also set $a_1 = a_2 = 1$, $b_1 = b_2 = p$ in order to obtain optimal posterior

concentration rates as suggested by Castillo and van der Vaart (2012). Further, the slab variance parameter λ_1 and ω_1 is fixed at 150 and for λ_0 (ω_0), we consider grid of 20 evenly spaced out values between 0.1 and 0.2 for each endpoints. We consider the sensitivity of the results to the tuning parameters $\lambda_0(\omega_0)$ and λ_1 (ω_1) and found them to be robust. Meanwhile, increasing the slab scale parameters (λ_1 and ω_1) only affect the number of iterations to convergence; in other words, the larger the λ_1 and ω_1, the more time it takes EMMEVS to converge. As with our simulation, we found the deterministic annealing variants of the EMMEVS to perform better and this is what we report here. Figure 12 presents the dynamic posterior exploration results and Table 7 displays the variables selected for each endpoints. The variables selected by each of the five endpoints differs, however.

Table 7. Variables selected using deterministic annealing version of EMMEVS at temperature = 20 for selected $\lambda_0(\omega_0)$ values along the regularization path leading to the selection of the predictors indicated with a bold font.

Variable Names	pcs	mcs	cesd	g1b	pcrec	Variable Description
a15a	0.42	0.27	0.94	**-6.89**	**20.03**	# nights in overnight shelter-last 6 months
a15b	-0.74	-0.10	**-7.48**	0.00	**16.56**	# nights on street-last 6 months
d1	-0.99	-0.75	**-2.15**	1.47	**-3.40**	# times ever hospitalized for medical problems
i1	-0.73	-0.15	**-0.33**	**0.22**	**-8.42**	Average # drinks in first 30 days before detoxification
i2	0.68	-0.01	-1.97	6.67	-3.08	Most drank any 1 day in first 30 before detoxification
age	**-1.85**	-0.10	**-3.18**	-0.62	-0.90	age at baseline (in years)
pss_fr	0.63	0.65	-0.20	0.12	-0.43	perceived social supports (friends)
daysdrink	-0.57	1.26	**-38.71**	**-80.71**	**-7.49**	Time (days) from baseline to first drink since leaving detoxification (6 months)
daysanysub	0.11	-0.50	**38.52**	**13.81**	**19.54**	Time (days) from baseline to first alcohol, heroin, or cocaine since leaving detoxification (6 months)
dayslink	-0.23	-0.31	**-31.48**	**-67.68**	**-14.84**	Time (days) to linkage to primary care within 12 months (by administrative record)
e2b	0.32	-0.13	**-2.16**	-0.19	**1.09**	# of times in past 6 months entered a detoxification program
drugrisk	-0.24	-0.20	0.12	**1.57**	**-1.44**	RAB-Drug Risk Total
sexrisk	0.00	0.12	-0.35	0.03	-0.97	RAB-Sex Risk Total
p1b	0.00	**1.64**	**-8.32**	**-5.63**	-0.89	Age first physical assaulted by person know
p2b	-0.63	-0.16	**-2.96**	**-8.02**	**-2.31**	Age first physical assaulted by stranger
p5b	1.14	**-2.16**	1.08	**3.30**	**-2.94**	Age first sex assaulted by person know
p6b	-0.38	**-1.92**	**-6.06**	**-2.21**	**-2.97**	Age first sex assaulted by stranger
c_au	-0.49	-0.93	0.04	0.01	-0.14	ASI-Composite Score for Alcohol Use
c_du	0.42	-0.47	-0.24	-0.08	-0.03	ASI-Composite Score for Drug Use
phys	**-1.92**	**-1.75**	0.30	0.11	0.01	InDUC-2L-Physical-RAW
inter	0.00	0.80	-0.22	-0.28	-0.66	InDUC-2L-Interpersonal-RAW
intra	0.38	-1.05	0.27	-0.08	0.00	InDUC-2L-Intrapersonal-RAW
impul	-0.06	-0.98	0.93	-1.01	-0.97	InDUC-2L-Impulse Control-RAW
sr	0.91	0.68	-0.25	-0.10	0.11	InDUC-2L-Social Responsibility-RAW
minage	-0.26	0.00	**-3.43**	**-1.75**	**1.61**	Age of first experience of physical or sexual abuse
female2	**-2.21**	**-5.36**	-0.41	-0.47	-0.43	Gender of respondent, baseline is male
substance2	**3.05**	**2.85**	-1.22	-0.21	0.14	primary substance of abuse cocaine, heroin is baseline
substance3	**2.82**	**2.56**	-0.76	**-0.61**	**-0.30**	primary substance of abuse alcohol, heroin is baseline
racegrp2	-0.86	-1.41	-1.49	-0.25	**-1.79**	hispanic, baseline is others
racegrp3	1.25	0.58	-1.16	-1.52	**-1.18**	white, baseline is others
racegrp4	0.34	1.26	-1.49	**-0.78**	**-1.13**	black, baseline is others
homeless2	0.82	-1.21	-0.50	**0.47**	0.37	Homeless-shelter/street past 6 months
satreat2	0.14	-0.13	-0.04	-0.06	0.63	Any substance abuse treatment this time point
drinkstatus2	-0.85	**1.71**	0.75	-0.05	-0.27	Drank alcohol since leaving detoxification? (6 month)
anysubstatus2	1.00	**-3.14**	-0.58	-0.37	-0.56	Used alcohol, heroin, or cocaine since leaving detoxification? (6 months)
abuseage2	0.49	**3.97**	0.17	0.00	-0.76	Age of onset of physical or sexual abuse is between 13 and 17 years, baseline is <13 years
abuseage3	0.64	**3.95**	-0.09	-0.58	0.00	Age of onset of physical or sexual abuse is more than 17 years, baseline is <13 years
hs_grad2	**2.05**	1.25	0.24	-0.26	-0.37	High School graduate, baseline is no

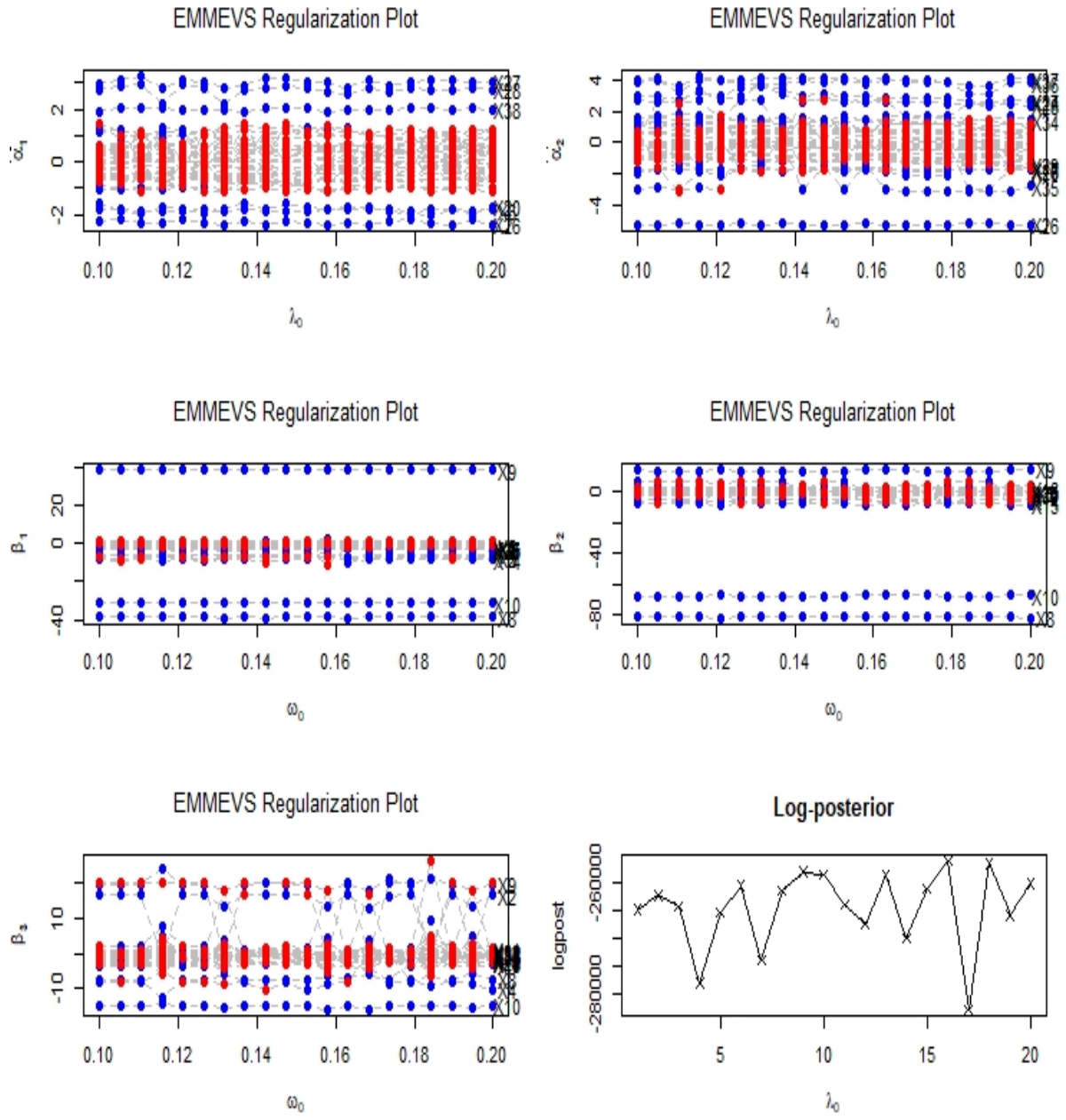

Figure 12. Plots of estimated regression coefficients for the trajectories of $\widehat{\boldsymbol{\alpha}}_1$, $\widehat{\boldsymbol{\alpha}}_2$, $\widehat{\boldsymbol{\beta}}_1$, $\widehat{\boldsymbol{\beta}}_2$, and $\widehat{\boldsymbol{\beta}}_3$ for varying choices λ_0 and ω_0. The estimates for variables with conditional posterior inclusion probability $P\left(\gamma_r = 1|\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}_1\right)$ and $P\left(\mu_r = 1|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}_2\right)$ above (below) 0.5 depicted in blue (red). The last log posterior plot in the third row is used for submodel evaluation, i.e, the selection of the optimum $\lambda_0(\omega_0)$ defined as the model with maximum λ_0, in this case was found to be 16 at $\lambda_0 = 0.1789474$

4.2 The National Alzheimer's Coordinating Center's Uniform Data Set

4.2.1 About the Data

The National Alzheimer's Coordinating Center is responsible for developing and maintaining a database of participant information collected from the 34 past and present Alzheimer's Disease Centers (ADC). The NIA organized the Alzheimer's Disease Centers Clinical Task Force and defined a standardized Uniform Data Set (UDS) (Beekly et al., 2007). The ADC provides researchers a standard set of assessment procedures to characterize mild Alzheimer's disease (AD) and mild cognitive impairment (MCI) in comparison with non-demented controls. NACC provided the patient data set used in this study. The data includes patients' demographics, health history, physical information, and four primary measurements of AD condition presented in Table 8. Our analysis is based on a subset of this data set concerning the status of AD for patients that attended the clinic for at least four times. The total number of subjects that met this requirement is 16070.

To focus on the identification of the risk factors of AD, patients who were marked with probable AD and had more than three visits were chosen from the original data set. As a result, 16,070 patients with likely AD between August 2005, and December 2019, as a subset of the NACC data set, were used in this study. The number of visits varies with patients in the subgroup, with an average of 4.98 visits for a patient while the maximal number reaches as high as 14. In total, there are 78 features/predictors (including time interval) in the created data subset. Detailed descriptions of other variables are available in Table 9. There are different data types in this data subset, such as continuous variables, ordinal variables, and nominal variables (including dummy variables). Four primary measurements mainly characterized the condition of AD for each patient:

We note that our proposed model is neutral to etiologic diagnoses, which implies that

Table 8. Description of endpoints extracted from the NACC uniform data set.

Endpoints	Endpoint Description	Type	Range
CDRGLOB (Morris, 1993)	Global Staging CDR	Ordinal	0.0 = No impairment 0.5 = Questionable impairment 1.0 = Mild impairment 2.0 = Moderate impairment 3.0 = Severe impairment
FAQ (Pfeffer et al., 1982)	Functional Activities Questionnaire	Discrete	0-30
MMSE and MoCA (Folstein et al., 1975; Hobson, 2015)	Mini-Mental State Examination	Discrete	0-30
CDRSUM (Morris, 1993)	Standard CDR sum of boxes	Discrete	0 - 18

other metrics defining AD progression stages can be used to replace the primary endpoints defined above.

4.2.2 Research Question

Clinicians and researchers have conducted numerous studies to identify the risk factors of AD. Most of these studies are carried out using a single endpoint (Gomar et al., 2011; Helzner et al., 2009; Lindsay et al., 2002; Ravaglia et al., 2006). However, due to the complex nature of the AD, AD condition is usually measured by multiple neuropathologic and clinical measurements. Thus, failing to leverage the association between these mixed endpoints when trying to identify the risk factors of AD may cast doubt on the results reported. In the present study, we assembled four of the most commonly used measurements for diagnosing the AD in order to address this shortcoming. Why is it essential to identify the risk factors for AD? Identifying the risk factors can help

1. retard disease progression during presymptomatic phases of AD/MCI, when it is more likely that pathologic changes can be arrested or reversed, and

2. demonstrate intervention efficacy, for instance, during clinical trials, it is essential to

recruit only subjects that are highly likely to have the disease (e.g., AD) because little or no prevention can be detected among the subjects that are not likely to have the disease. Selecting participants to maximize treatment benefits (such as study enrichment, for example, recruiting those with high risk for developing AD) is critical for a reduction in the cost of following false-positive subjects longitudinally and avoid trial failure due to the benefits among the experimental group.

The question is, how do we identify patients that are likely to develop the AD (or other potential outcomes of interests) accurately when multivariate mixed endpoints characterized by high-dimensional predictors?

4.2.3 Results - UDS Data Analyses

Our analysis in this section was based on the NACC UDS data set described above. We used this data to illustrate the power of EMMEVS and answer the research question posed in Section 4.2.2. In the analyses, we set $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)} = \mathbf{1}_p$, $\sigma_\epsilon^{2(0)} = 3$, $\sigma_a^{2(0)} = 0.1$, $\boldsymbol{\Sigma} = \boldsymbol{I}_q$. The hyper-parameters ν_1, ν_2, κ_1, and κ_2 are set to 10000 and $\nu_0 = p + 1$. We also set $a_1 = a_2 = 1$, $b_1 = b_2 = p$ in order to obtain optimal posterior concentration rates as suggested by Castillo and van der Vaart (2012). Further, the slab variance parameter λ_1 and ω_1 is fixed at 100 and for $\lambda_0(\omega_0)$, we consider grid of 20 evenly spaced out values between 0.01 and 0.015 for each endpoints. We consider the sensitivity of the results to the tuning parameters $\lambda_0(\omega_0)$ and $\lambda_1(\omega_1)$ and found them to be robust. Meanwhile, increasing the slab scale parameters ($\lambda_1$andω_1) only affect the number of iterations to convergence; in other words, the larger the $\lambda_1$andω_1, the more time it takes EMMEVS to converge. As with our simulation, we found the deterministic annealing variants of the EMMEVS to perform better and this is what we report here. Figure 13 presents the dynamic posterior exploration results and Table 9 displays the variables selected for each endpoints.

It is pertinent to remark here that the analysis of the UDS data presented here is for illustrative purposes only, we plan to do a more substantive and comprehensive analysis of this data in the nearest future.

Table 9. Variables selected using deterministic annealing version of EMMEVS at temperature = 20 for selected $\lambda_0(\omega_0)$ values along the regularization path leading to the selection of the predictors indicated with bold font.

Variable Names	CDRGLOB	FAQ	MMSE	CDRSUM	Variable Description
SEX2	**-0.38**	**0.83**	**-0.98**	**0.07**	Subject's sex
NACCAPOE2	**0.72**	**-0.21**	**1.84**	**4.33**	APOE genotype e3,e4
NACCAPOE3	**-0.04**	**0.31**	**0.12**	**-2.37**	APOE genotype e3,e2
NACCAPOE4	1.31	**-0.73**	**3.05**	**-3.23**	APOE genotype e4,e4
NACCAPOE5	**0.58**	**-0.23**	**0.75**	**2.70**	APOE genotype e4,e2
NACCAPOE6	0.54	**0.31**	1.08	0.01	APOE genotype e2,e2
NACCAGE	**0.01**	0.05	**0.38**	**0.36**	Subject's age at visit
NACCAGEB	0.00	0.03	**-0.36**	**1.35**	Subject's age at initial visit
EDUC	-0.01	**0.32**	-0.07	**-1.78**	Years of education
NACCBMI	0.02	0.13	0.04	**-0.64**	Body mass index (BMI)
BPDIAS	0.01	0.05	0.01	**-1.67**	Subject blood pressure (sitting), systolic
BPSYS	0.00	0.02	0.00	**-1.73**	Subject blood pressure (sitting), diastolic
HRATE	0.01	0.04	0.03	**-0.54**	Subject resting heart rate (pulse)
SMOKYRS	0.00	0.00	0.00	**1.81**	Total years smoked cigarettes
QUITSMOK	0.00	0.04	-0.01	**-1.60**	Age at which subject quit smoking
NACCETPR	-0.09	0.09	**-0.22**	**1.20**	Primary etiologic diagnosis e.g., MCI, impaired, not MCI

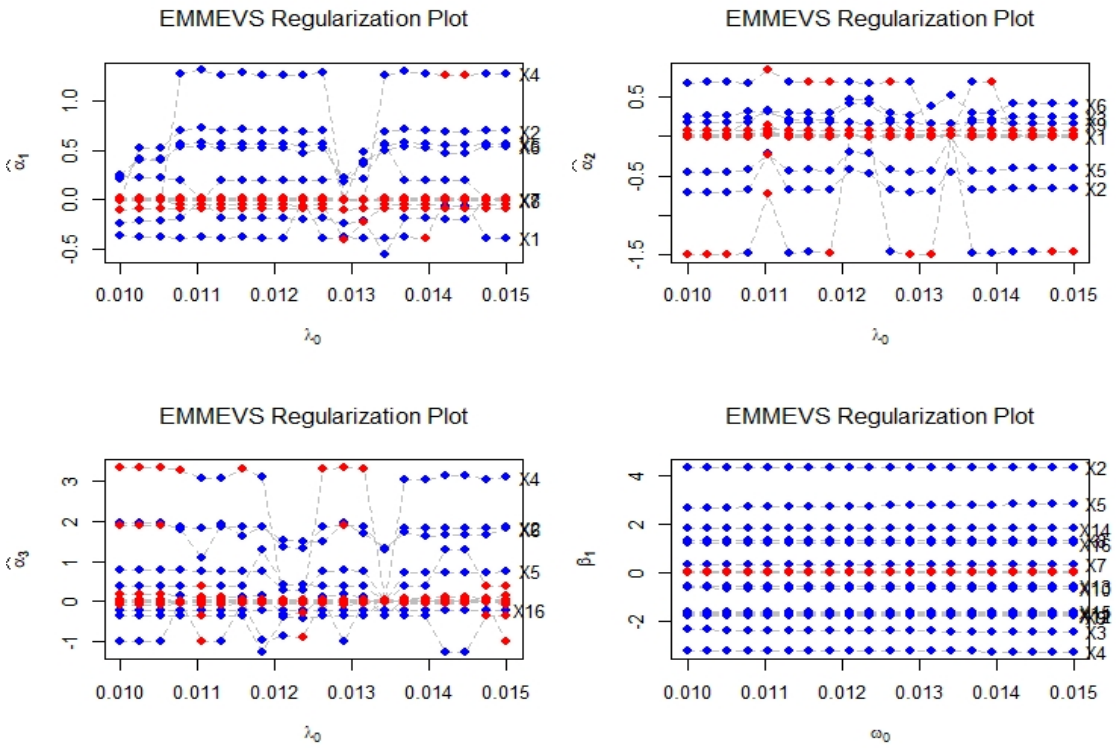

Figure 13. Plots of estimated regression coefficients for the trajectories of $\widehat{\boldsymbol{\alpha}}_1$, $\widehat{\boldsymbol{\alpha}}_2$, $\widehat{\boldsymbol{\alpha}}_3$, and $\widehat{\boldsymbol{\beta}}_1$ for varying choices λ_0 and ω_0. The estimates for variables with conditional posterior inclusion probability $P\left(\gamma_r = 1|\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}_1\right)$ and $P\left(\mu_r = 1|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}_2\right)$ above (below) 0.5 depicted in blue (red).

Chapter 5

Conclusion and Further Research

In this chapter, we present the details of the ongoing work and further research. In particular, we give a comprehensive account of two key methodologies that can be used to model and analyze multivariate mixed endpoints with high-dimensional covariates.

5.1 Use of Scale Mixture of Uniform Distribution

To address the statistical and computational challenges described in Chapter 1, we are working on another novel procedure that makes use of a data augmentation strategy motivated by a scale mixture representation of the uniform distribution. We consider a fully Bayesian solution to answer the research question posed in Section 1 by assigning independent priors on the regression parameters. Due to the Bayesian interpretation of the frequentist LASSO, the Bayesian regularization framework has enjoyed increased applicability in the literature (see, for example, Armagan et al., 2013; Bhadra et al., 2017; Figueiredo, 2003; Hans, 2009; Park and Casella, 2008. Park and Casella (2008) introduced the Gibbs sampling procedure using a conditional Laplace prior with a particular focus on addressing the multimodality issues. Other methods based on Laplace priors include the Bayesian LASSO via reversible-jump MCMC (Chen et al., 2011) and Bayesian LASSO regression of Hans (2009). In our ongoing study, we suggest a new hierarchical formulation of the Bayesian LASSO utilizing the SMU representation of the Laplace distribution in the context of multivariate mixed endpoints data. We are motivated to consider a regularized Bayesian procedure because compared to the classical counterpart, the Bayesian framework:

1. Is endowed with several model summaries and parameter uncertainties (e.g., mean, standard errors), which follows naturally from the posterior distributions.

2. Have estimates with an intuitive interpretation. For example, a 95% Bayesian credibility interval can simply be interpreted as the interval in which the actual value lies with 95% probability.

3. Is flexible and computationally efficient, leading to scalable MCMC algorithms with good convergence and mixing properties. The cost of the flexibility of MCMC, however, is that it requires more computation time compared to standard optimization procedures.

4. Can estimate the penalty parameter(s) simultaneously with the model parameters in a single step.

5. Can handle multimodal optimization problems well. Indeed, this is one of the most persuasive arguments for pursuing a fully Bayesian approach, as summarizing a multimodal surface with a single frequentist point estimate can be vastly misleading (Polson et al., 2014). This multimodal issue is particularly worrisome in our case, where we are dealing with multivariate mixed endpoints with high-dimensional covariates.

Andrews and Mallows (1974), Feller (1971) introduced the use of scale mixtures in the statistics literature. These authors used it to sample symmetric distribution with normal components and become what is known as the scale mixture of normal (SMN) distribution. Walker and Gutierrez-Pena (1999), Walker et al. (1997) proposed a new class of scale mixture of distribution known as the scale mixtures of uniform (SMU) distribution.

The SMU density representation is similar to the SMN representation but with the normal distribution replaced by a uniform distribution whose support is determined by the mixing parameter. Since the appearance of SMU, several authors Choy et al., 2009; Choy and Chan, 2008; Mallick and Yi, 2014; Qin et al., 1998a, 1998; Qin et al., 2003 have applied the SMU distribution in their research. Walker et al. (1997) used SMU distribution in normal regression models in the non-Bayesian framework. Qin et al. (1998) provided Gibbs sampler

by using SMU in variance regression models and also to derive Gibbs sampler for auto-correlated heteroscedastic regression models (Qin et al., 1998a). Choy et al. (2009) used it in a stochastic volatility model with a two-stage scale mixture representation of the student-t distribution. The use of SMU in multivariate mixed endpoints data settings has received less attention. Our goal is to propose a Gibbs sampler that utilizes the SMU distribution for the Laplace density.

Typically, a popular choice of the prior distribution for the regression coefficient is a normal distribution with zero mean and unknown variance. The use of the Gaussian prior distribution leads to a ridge estimator and has been reported to perform poorly if there are large differences in the size of the fixed coefficients (Griffin & Brown, 2010). According to Tibshirani (1996), Laplace prior is a generalization of the ridge prior and leads to the LASSO estimator, we, therefore, elected to place the following independent Laplace distribution as a prior on the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

$$\tau(\alpha_\ell|\lambda) = \prod_{\ell=1}^{c} \frac{1}{2\lambda} \exp\left\{-\frac{|\alpha_\ell|}{\lambda}\right\} \text{ and } \tau(\beta_f|\omega) = \prod_{f=1}^{h} \frac{1}{2\omega} \exp\left\{-\frac{|\beta_f|}{\omega}\right\} \tag{5.1}$$

A noteworthy feature of the Laplace densities given in (5.1) is that it can be reformulated as an SMU distribution; this formulation which makes the model inference tractable is presented in the proposition

Proposition 1. The Laplace density $\tau(x) = \frac{1}{2\Xi} \exp\left(-\frac{|x|}{\Xi}\right)$ can be written as a scale mixture of uniform distribution, the mixing density being a gamma distribution. That is

$$\frac{1}{2c} \exp\left(-\frac{|x|}{\Xi}\right) = \int_{-u\Xi<x<u\Xi} \frac{1}{2u} \text{Gamma}\left(2, \frac{1}{\Xi}\right) du \tag{5.2}$$

As with many penalized regression problems, we add a penalty term to the minimization of the sum of squared residuals, resulting in the following regularization problem (with

the goal of shrinking small coefficients towards zero while leaving large coefficients large.)

$$\text{minimize}\{RSS + penalty\} \tag{5.3}$$

However, rather than minimizing (5.3), we addressed the problem from a Bayesian perspective. We solve the problem by constructing a Markov chain whose stationary distribution is the joint posterior for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and the minimizer of (5.3) as its global mode.

5.2 Use of the Spike-and-Slab LASSO

In chapter 3, we used the spike-and-slab mixtures of Laplace priors to address the statistical and computational challenges inherent in the multiple mixed endpoints characterized with high-dimensional variables. The procedure derived therein considered the problem of estimation and variable selection from a non-penalized perspective. In this chapter, we present a new method that combine the advantage of the Bayesian procedure with the readily available computational algorithm in the frequentist paradigm. Specifically, we suggest a penalized procedure that borrows strength from the spike-and-slab Laplace densities and the frequentist LASSO which induces sparsity through penalty functions. This chapter draws inspiration from Ročková and George (2018) spike-and-slab LASSO for sparse normal means estimation in a univariate linear regression setting.

The Spike-and-Slab LASSO (SSL) was introduced by Ročková and George (2018), SSL places a mixture of two Laplace densities on each regression coefficients, δ_j, as follows:

$$\tau(\boldsymbol{\delta}|\boldsymbol{\vartheta}) = \prod_{j=1}^{p} \left[\vartheta_j \tau\left(\delta_j|\eta_1\right) + \left(1-\vartheta_j\right)\tau\left(\delta_j|\eta_0\right)\right] \tag{5.4}$$

where $0 < \theta < 1$ is defined as the mixing proportion and $\tau\left(\delta|\eta\right) = \frac{\eta}{2}\exp\{-\eta|\delta|\}$ denotes a univariate Laplace distribution with mean 0 and variance $\frac{2}{\eta^2}$. Typically, it is assumed that

$\eta_0 \gg \eta 1 > 0$, this allows the spike distributions $\tau(\delta_j|\eta_0)$ to be concentrated around zero and the slab distribution, $\tau(\delta_j|\eta_1)$ to be relatively diffuse (Ročková & George, 2018).

The SSL model has wide-applicability outside of univariate linear regression: it has been adapted to address wide-ranging statistical and computational problems such as generalized linear models problems (Tang et al., 2018; Tang et al., 2017a), covariance matrix estimation(Deshpande et al., 2019; Gan et al., 2019), causal inference (Antonelli et al., 2019), group LASSO and generalized additive models, (Bai et al., 2019) factor analysis (Ročková & George, 2016), and Cox proportional hazard models (Tang et al., 2017b). Meanwhile, the use of SSL in multivariate mixed endpoints model with high-dimensional covariates is sparse, and to the best of our knowledge, this is the first study to apply SSL in that setting.

We extend the use of the SSL to analyze data with multiple mixed endpoints and large number of covariates. We referred to our formulation as the spike-and-slab LASSO with mixed endpoints (hereafter, SSLME). Under the SSLME prior, the global posterior mode is exactly sparse, thereby allowing the mode to automatically separate the active from the non-active regression coefficients. Our development here is based on the non-separable (fully Bayes) and self-adaptive penalty that allow us to automatically adapt to ensemble information about sparsity. The continuous nature of our prior is critical in facilitating efficient coordinate ascent algorithm for the maximum *a posteriori* (MAP) estimation that allow us to bypass the use of MCMC such as the Gibbs sampling procedure described in Section 5.1.

5.3 Further Research

In Chapter 3, we assumed a Gaussian distribution for the unobserved continuous latent variable underlying the discrete outcomes to jointly modeled the discrete and continuous responses. However, there are other potential alternatives to the Gaussian distribution. To this end, we intend to investigate the effects of assuming other distributions like t-latent

distributions (de Leon & Wu, 2011; Liu, 2005; Tan et al., 1999), latent logistic regression (Nikoloulopoulos & Karlis, 2008), and the Gaussian copula method (de Leon & Wu, 2011) on our model. It will be interesting to investigate the behavior and sensitivity of our model under each of these assumptions.

Another extension of our proposed methodology which we intend to research in the future is when the structural information about the predictors is available such as biological information in a genetic context. to achieve the modeling here, a more flexible priors, such as the logistic regression product prior (Stingo et al., 2010) and the Markov random field prior (Li & Zhang, 2010; Stingo & Vannucci, 2011) can be employed to transmit the biological information.

One crucial reality with any modeling and data analysis is the practicality of having to deal with missing data – usually, this is a common situation. For example, in the two real-data sets we analyzed in Chapter 4, there are so many missing observations. Therefore, part of our future plan is to continue research on variable selection in clustered/longitudinal multivariate mixed endpoints characterized by high dimensional covariates to a more realistic case of missing responses in both discrete and continuous responses. The analysis of mixed endpoints data with missing values have been investigated in the past in the context of bivariate data. For example, Little and Schluchter (1985) and Fitzmaurice and Laird (1997) modeled such data using the general location model of Olkin and Tate (1961) and assume the missing values are missing at random to justify their ignoring the missing data.

Another potential future research focus of mine is to extend the model selection research to the informative visit processes in a longitudinal design.

5.4 Conclusion

Statisticians have developed several procedures for the estimation and variable selection in high-dimensional dynamic regression models for linear regression (Fan & Li, 2001;

Tibshirani, 1996; Tibshirani et al., 2005; Zou & Hastie, 2005), logistic/multinomial data (Holmes & Held, 2006; Tutz & Pössnecker, 2012), quantile regression (Alhamzawi & Yu, 2012), count/zero-inflated models (Algamal, 2019; Buu et al., 2011; Wang et al., 2015), censored survival data (Faraggi & Simon, 1998; Johnson, 2009; Zhanfeng et al., 2010), ordinal regression (Aljabri & Alhamzawi, 2019; Feng et al., 2017; Fu & Archer, 2020; Sukthuayat & Chaimongkol, 2018), and Tobit regression (Huang et al., 2020; Liu et al., 2013) among others. However, despite the presence of the mixed endpoints data of differing types for over two decades and more importantly during this era of big data analysis, it is surprising that we do not have a method to carry out both parameter estimation and variable selection. This dissertation fills the gap.

In the dissertation, we present the first-ever known variable selection procedure for multivariate mixed endpoints that are characterized by high-dimensional covariates. We referred to this method as EMMEVS - a rapid deterministic method based on EM algorithm. It is a deterministic alternative to MCMC methods that has the potential to lower the computational burden commonly encountered with the use of MCMC. The computational speed of the EMMEVS algorithm allows for the exploration of many sub-models within a short period.

We demonstrate the advantage of our procedure in terms of variable selection, prediction, and computational scalability via extensive simulation study and apply the method to two real-life data. The results obtained from the simulation and analysis of the two real data revealed that the modal estimates identified by our dynamic posterior exploration stabilized rapidly very early in its trajectories (especially with our implementation of the dynamic weighted LASSO scheme and as the temperature increases); thus, allowing us to report a single estimate out of the many we computed without the need for cross-validation.

Bibliography

Albert, J., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Stat. Assoc.*, *88*, 669.

Algamal, Z. (2019). Variable selection in count data regression model based on firefly algorithm. *Statistics, Optimization & Information Computing*, *7*(2), 520–529.

Alhamzawi, R., & Ali, H. T. M. (2018). The bayesian adaptive lasso regression. *Mathematical Biosciences*, *303*, 75–82.

Alhamzawi, R., & Yu, K. (2012). Variable selection in quantile regression via gibbs sampling. *Journal of Applied Statistics*, *39*(4), 799–813.

Aljabri, D. H. Q., & Alhamzawi, R. (2019). Bayesian bridge regression for ordinal models with a practical application. *Journal of Physics: Conference Series*, *1294*, 032030.

Andrews, D. F., & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(1), 99–102.

Antonelli, J., Parmigiani, G., & Dominici, F. (2019). High-dimensional confounding adjustment using continuous spike and slab priors. *Bayesian Anal.*, *14*(3), 805–828.

Armagan, A., Dunson, D. B., & Lee, J. (2013). Generalized double pareto shrinkage. *Stat Sin.*, *23*(1), 119–143.

Bai, R., Moran, G. E., Antonelli, J., Chen, Y., & Boland, M. R. (2019). Spike-and-slab group lassos for grouped regression and sparse generalized additive models.

Beekly, D. L., Ramos, E. M., Lee, W. W., Deitrich, W. D., Jacka, M. E., Wu, J., Hubbard, J. L., Koepsell, T. D., Morris, J. C., & Kukull, W. A. (2007). The nia alzheimer's disease centers the national alzheimer's coordinating center (nacc) database: The uniform data set. *Alzheimer Disease & Associated Disorders*, *21*(3), 249–258.

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. T. (2017). Lasso meets horseshoe : A survey.

Bowman, D., & George, E. O. (2018). A bayesian analysis of clustered discrete and continuous outcomes. *Journal of Applied Statistics*, *45*(3), 438–449.

Brown, P. J., Vannucci, M., & Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *64*(3), 519–536.

Buhlmann, P., Kalisch, M., & Mathuis, M. H. (2010). Variable selection in high-dimensional linear models: Partially faithful distributions and the pc-simple algorithm. *Biometrika*, *97*(2), 261–278.

Buu, A., Johnson, N. J., Li, R., & Tan, X. (2011). New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statistics in medicine*, *30*(18), 2326–2340.

Carvalho, C., & Polson, N. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*, 465.

Castillo, I., Schmidt-Hieber, J., & van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, *43*, 1986.

Castillo, I., & van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, *40*, 2069.

Catalano, P. J., & Ryan, L. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Amer. Stat. Assoc.*, *87*, 651.

Chang, C., Kundu, S., & Long, Q. (2018). Scalable bayesian variable selection for structured high-dimensional data. *Biometrics*, *74*(4), 1372–1382.

Chang, C., & Tsay, R. S. (2010). Estimation of covariance matrix via the sparse cholesky factor with lasso. *Journal of Statistical Planning and Inference*, *140*(12), 3858–3873.

Choy, B., Wan, Y., & Chan, C. (2009). Bayesian student-t stochastic volatility models via scale mixtures. *Advances in Econometrics*, *23*.

Choy, B., & Chan, J. (2008). Scale mixtures distributions in statistical modelling. *Australian & New Zealand Journal of Statistics*, *50*(2), 135–146.

Cox, D. R. (1972). The analysis of multivariate binary data. *Appl. Stat.*, *21*, 113.

Cox, D. R., & Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika*, *79*(3), 441–461.

Daniels, M. J., & Normand, S. L. (2006). Longitudinal profiling of health care units based on continuous and discrete patient outcomes. *Biostatistics*, *7*(1), 1–15.

Das, K., Rao, J., & Biswas, A. (1999). Bayesian analysis of clustered regression models for mixed discrete and continuous outcomes – an application to depressive disorder data. *Calcutta Statistical Association Bulletin*, *49*(3-4), 255–268.

D'Aunno, T. (1997). Linking substance abuse treatment and primary health care. (J. A. Egerton, D. M. Fox, & A. I. Leshner, Eds.). In J. A. Egerton, D. M. Fox, & A. I. Leshner (Eds.), *Treating drug abusers effectively*. London: Basil Blackwell.

de Leon, A. R., & Wu, B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, *30*(2), 175–185.

de Leon, A. R., & Carrière, K. C. (2007). General mixed-data model: Extension of general location and grouped continuous models. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, *35*(4), 533–548.

Dempster, A. P., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, *39*, 1.

Deshpande, S. K., Ročková, V., & George, E. I. (2019). Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics*, *28*(4), 921–931.

Dickey, B., Normand, S.-L. T., Hermann, R. C., Eisen, S. V., Cortés, D. E., Cleary, P. D., & Ware, N. (2003). Guideline Recommendations for Treatment of Schizophrenia: The Impact of Managed Care. *Archives of General Psychiatry*, *60*(4), 340–348.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *J. R. Stat. Soc. B*, *62*, 355.

Dunson, D. B., Chen, Z., & Harry, J. (2003). A bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics*, *59*(3), 521–530.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, *32*(2), 407–499.

Faes, C., Aerts, M., Molenberghs, G., Geys, H., Teuns, G., & Bijnens, L. (2008). A high-dimensional joint model for longitudinal outcomes of different nature. *Statistics in Medicine*, *27*(22), 4408–4427.

Faes, C., Geys, H., Aerts, M., Molenberghs, G., & Catalano, P. J. (2004). Modeling combined continuous and ordinal outcomes in a clustered setting. *Journal of Agricultural, Biological, and Environmental Statistics*, *9*(4), 515–530.

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

Faraggi, D., & Simon, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics*, *54 4*, 1475–85.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* New York, John Wiley & Sons Inc.

Feng, X.-N., Wu, H.-T., & Song, X.-Y. (2017). Bayesian adaptive lasso for ordinal regression with latent variables. *Sociological Methods & Research*, *46*(4), 926–953.

Fieuws, S., & Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, *62*(2), 424–431.

Figueiredo, M. A. (2003). Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*, 1150.

Fitzmaurice, G. M., & Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, *90*(431), 845–852.

Fitzmaurice, G. M., & Laird, N. M. (1997). Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*, *53*(1), 110–122.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.*, *33*(1), 1–22.

Fu, H., & Archer, K. J. (2020). High-dimensional variable selection for ordinal outcomes with error control. *Briefings in Bioinformatics*.

Gan, L., Narisetty, N. N., & Liang, F. (2019). Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association*, *114*(527), 1218–1231.

George, E. O., Armstrong, D., Catalano, P. J., & Srivastava, K. (2007). Regression models for analyzing clustered binary and continuous outcomes under an assumption of exchangeability. *J. Stat. Plan. Inference*, *137*, 3462.

George, E. I., & McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, *88*, 881.

George, E. I., & McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica Sinica*, *7*, 339.

Geys, H. M., Regan, M. M., Catalano, P. J., & Molenberghs, G. (2001). Two latent variable risk assessment approaches for mixed continuous and discrete outcomes from developmental toxicity. *J. Agric. Biol. Environ. Stat.*, *6*, 340.

Geys, H., Molenberghs, G., & Ryan, L. M. (1999). Pseudolikelihood modeling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, *94*(447), 734–745.

Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, *9*(3), 173–197.

Gomar, J. J., Bobes-Bascaran, M. T., Conejero-Goldberg, C., Davies, P., Goldberg, T. E., & Alzheimer's Disease Neuroimaging Initiative, f. t. (2011). Utility of Combinations of Biomarkers, Cognitive Markers, and Risk Factors to Predict Conversion From Mild Cognitive Impairment to Alzheimer Disease in Patients in the Alzheimer's Disease Neuroimaging Initiative. *Archives of General Psychiatry*, *68*(9), 961–969.

Griffin, J. E., & Brown, P. J. (2005). Alternative prior distributions for variable selection with very many more variables than observations [Working Paper. Coventry: University of Warwick. Centre for Research in Statistical Methodology].

Griffin, J. E., & Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.*, *5*(1), 171–188.

Griffin, J. E., & Brown, P. J. (2012). Structuring shrinkage: Some correlated priors for regression. *Biometrika*, *99*(2), 481–487.

Gueorguieva, R., & Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous response. *J. Amer. Stat. Assoc.*, *96*, 1102.

Gueorguieva, R., & Sanacora, G. (2006). Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in medicine*, *258*, 1307–1322.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, *96*(4), 835–845.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Second). Springer-Verlag New York.

Helzner, E. P., Luchsinger, J. A., Scarmeas, N., Cosentino, S., Brickman, A. M., Glymour, M. M., & Stern, Y. (2009). Contribution of Vascular Risk Factors to the Progression in Alzheimer Disease. *Archives of Neurology*, *66*(3), 343–348.

Hobson, J. (2015). The montreal cognitive assessment (moca). *Occupational Medicine*, *65*(9), 764–765.

Holmes, C. C., & Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, *1*(1), 145–168.

Holmes, L. B., Harvey, E. A., Brown, K. S., Hayes, A. M., & Khoshbin, S. (1994). Anticonvulsant teratogenesis: I. a study design for newborn infants. *Teratology*, *49*(3), 202–207.

Hu, Y., Zhao, K., & Lian, H. (2015). Bayesian quantile regression for partially linear additive models. *Statistics and Computing*, *25*, 651–668.

Huang, H., Shangguan, J., Li, Y., & Liang, H. (2020). Bi-level variable selection in high dimensional tobit models. *Statistics and Its Interface*, *13*, 151–156.

Huang, J., Breheny, P., Lee, S., Ma, S., & Zhang, C.-H. (2016). The mnet method for variable selection. *Statistica Sinica*, *26*(3), 903–923.

Johnson, B. A. (2009). On lasso for censored data. *Electron. J. Statist.*, *3*, 485–506.

Kiiveri, H. (2003). A bayesian approach to variable selection when the number of variables is very large. *Institute of Mathematical Statistics Lecture Notes-Monograph Series*, *40*, 127.

Li, F., & Zhang, N. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, *105*, 1978.

Li, Q., Xi, R., & Lin, N. (2010). Bayesian regularized quantile regression. *Bayesian Anal.*, *5*, 1.

Lindsay, J., Laurin, D., Verreault, R., Hébert, R., Helliwell, B., Hill, G. B., & McDowell, I. (2002). Risk Factors for Alzheimer's Disease: A Prospective Analysis from the Canadian Study of Health and Aging. *American Journal of Epidemiology*, *156*(5), 445–453.

Little, R. J., & Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, *72*, 496.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (Second). John Wiley & Sons, Inc.

Liu, C. (2005). Robit regression: A simple robust alternative to logistic and probit regression. In *Applied bayesian modeling and causal inference from incomplete-data perspectives* (pp. 227–238).

Liu, C., & Rubin, D. B. (1998). Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika*, *85*(3), 673–688.

Liu, X., Wang, Z., & Wu, Y. (2013). Group variable selection and estimation in the tobit censored response model. *Computational Statistics & Data Analysis*, *60*, 80–89.

Mallick, H., & Yi, N. (2014). A new bayesian lasso. *Statistics and its interface*, *74*, 571–582.

McLachlan, G. J., & Basford, K. E. (2004). *Mixture models: Inference and applications to clustering*. Marcel Dekker.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer.

Morris, J. C. (1993). The clinical dementia rating (cdr). *Neurology*, *43*(11), 2412–2412.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*(3), 391–411.

Nikoloulopoulos, A. K., & Karlis, D. (2008). Multivariate logit copula model with an application to dental data. *Statistics in Medicine*, *27*(30), 6393–6406.

Oliveira, R. M., & Teixeira-Pinto, A. (2015). Analyzing multiple outcomes: Is it really worth the use of multivariate linear regression? *Journal of Biometrics & Biostatistics*, *06*(04).

Olkin, I., & Tate, R. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Stat.*, *32*, 448.

O'Malley, A. J., Normand, S.-L. T., & Kuntz, R. E. (2002). Application of models for multivariate mixed outcomes to medical device trials: Coronary artery stenting. *Statistics in medicine*, *22 2*, 313–36.

Panagiotelis, A., & Smith, M. (2008). Bayesian identification, selection and estimation of semiparametric functions in high-dimensional additive models. *Journal of Econometrics*, *143*(2), 291–316.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

Pfeffer, R. I., Kurosaki, T. T., Harrah, C. H., Chance, J. M., & Filos, S. (1982). Measurement of functional activities in older adults in the community. *Journal of Gerontology*, *37*(3), 323–329.

Polson, N., & Scott, J. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian Statistics*, *9*, 501.

Polson, N. G., Scott, J. G., & Windle, J. (2014). The bayesian bridge. *JRSS: Series B (Statistical Methodology)*, *76*(4), 713–733.

Qin, Z., Walker, S., & Damien, P. (1998a). Uniform scale mixture models with applications to bayesian inference. [Working papers series, University of Michigan Ross School of Business].

Qin, Z., Walker, S., & Damien, P. (1998). Uniform scale mixture models with applications to variance regression. [Working papers series, University of Michigan Ross School of Business].

Qin, Z. S., Damien, P., & Walker, S. (2003). Scale mixture models with applications to bayesian inference. *AIP Conference Proceedings*, *690*(1), 394–395.

Ravaglia, G., Forti, P., Maioli, F., Martelli, M., Servadei, L., Brunetti, N., Pantieri, G., & Mariani, E. (2006). Conversion of Mild Cognitive Impairment to Dementia: Predictive Role of Mild Cognitive Impairment Subtypes and Vascular Risk Factors. *Dement Geriatr Cogn Disord*, *21*, 51–58.

Regan, M. M., & Catalano, P. J. (1999). Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. *Biometrics*, *55*, 760.

Ročková, V., & George, E. I. (2014). EMVS: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, *109*(506), 828–846.

Ročková, V., & George, E. I. (2016). Fast bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, *111*(516), 1608–1622.

Ročková, V., & George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, *113*(521), 431–444.

Saitz, R., Mulvey, K. P., & Samet, J. H. (1997). The substance abusing patient and primary care: Linkage via the addiction treatment system? *Substance Abuse*, *18*, 187–195.

Samet, J. H., Larson, M. J., Horton, N. J., Doyle, K., Winter, M., & Saitz, R. (2003). Linking alcohol and drug dependent adults to primary medical care: A randomized controlled trial of a multidisciplinary health intervention in a detoxification unit. *Addiction*, *98*, 509–516.

Sammel, M. D., Lin, X., & Ryan, L. (1999). Multivariate linear mixed models for multiple outcomes. *Statistics in Medicine*, *18*(17-18), 2479–2492.

Sammel, M. D., Ryan, L. M., & Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society. Series B (Methodological)*, *59*(3), 667–678.

Scott, J., & Berger, J. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, *38*, 2587.

Stingo, F., Chen, Y., Vannucci, M., Barrier, M., & Mirkes, P. (2010). A bayesian graphical modeling approach to microrna regulatory network inference. *Annals of Applied Statistics*, *4*, 2024.

Stingo, F., & Vannucci, M. (2011). Variable selection for discriminant analysis with markov random field priors for the analysis of microarray data. *Bioinformatics*, *27*, 495.

Sukthuayat, C., & Chaimongkol, S. (2018). A study of the performance for variable selection of ordinal regression models under multicollinearity using reversible jump algorithm. *Journal of Applied Statistics and Information Technology*, *3*(2), 19–30.

Tan, M., Qu, Y., & Rao, S. J. (1999). Robustness of the latent variable model for correlated binary data. *Biometrics*, *55*(1), 258–263.

Tang, Z., Shen, Y., Li, Y., Zhang, X., Wen, J., Qian, C., Zhuang, W., Shi, X., & Yi, N. (2018). Group spike-and-slab lasso generalized linear models for disease prediction and associated genes detection by incorporating pathway information. *Bioinformatics*, *34*(6), 901–910.

Tang, Z., Shen, Y., Zhang, X., & Yi, N. (2017a). The spike-and-slab lasso generalized linear models for prediction and associated genes detection. *Genetics*, *205*(1), 77–88.

Tang, Z., Shen, Y., Zhang, X., & Yi, N. (2017b). The spike-and-slab lasso Cox model for survival prediction and associated genes detection. *Bioinformatics*, *33*(18), 2799–2807.

Tate, R. F. (1954). Correlation between a discrete and a continuous variable. point-biserial correlation. *The Annals of Mathematical Statistics*, *25*(3), 603–607.

Teixeira-Pinto, A., & Mauri, L. (2011). Statistical analysis of noncommensurate multiple outcomes. *Circulation. Cardiovascular quality and outcomes*, *4*, 650–656.

Teixeira-Pinto, A., & Normand, S.-L. T. (2009). Correlated bivariate continuous and binary outcomes: Issues and applications. *Statistics in medicine*, *28*(13).

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267.

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

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

Tutz, G., & Pössnecker, W. (2012). *Variable selection for the multinomial logit model.*, *Conference: Proceedings of the 27th international workshop on statistical modelling, prag.*

Ueda, N., & Nakano, R. (1998). Deterministic annealing em algorithm. *Neural Networks*, *11*, 271.

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50.

Walker, S., & Gutierrez-Pena, E. (1999). Robustifying bayesian procedures (with discussion) (J. Bernardo, J. Berger, A. Dawid, & A. Smith, Eds.). In J. Bernardo, J. Berger, A. Dawid, & A. Smith (Eds.), *In bayesian statistics 6*. Oxford: University Press.

Walker, S. D., Damien, P., & Meyer, M. (1997). *On scale mixtures of uniform distributions and the latent weighted least squares method* [Working paper, University of Michigan Business School]. Working paper, University of Michigan Business School.

Wang, X., Wypij, D., Gold, D., Speizer, F., Ware, J. H., Ferris, B., & Dockery, D. (1994). A longitudinal study of the effects of parental smoking on pulmonary function in children 6-18 years. *American Journal of Respiratory and Critical Care Medicine*, *149*(6), 1420–1425.

Wang, Z., Ma, S., & Wang, C. Y. (2015). Variable selection for zero-inflated and overdispersed data with application to health care demand in germany. *Biometrical journal*, *57*(5), 867–884.

Weiss, R. E., Jia, J., & Suchard, M. A. (2011). A bayesian model for the common effects of multiple predictors on mixed outcomes. *Interface Focus*, *1*(6), 886–894.

Wu, T. T., & Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.*, *2*(1), 224–244.

Wu, Y., & Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, *19*, 801–817.

Zhanfeng, W., Yaohua, W., & Lincheng, Z. (2010). A lasso-type approach to variable selection andestimation for censored regression model. *Chinese Journal of Applied Probability and Statistics*, *26*(1), 66.

Zhang, C.-H., & Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, *36*(4), 1567–1594.

Zhang, H., Huang, X., Gan, J., Karmaus, W., & Sabo-Attwood, T. (2016a). A two-component g-prior for variable selection. *Bayesian Anal.*, *11*(2), 353–380.

Zhang, H., Maity, A., Arshad, H. S., Holloway, J. W., & Karmaus, W. J. J. (2016b). Variable selection in semi-parametric models. *Statistical Methods in Medical Research*, *25*, 1736–1752.

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

Zou, H., & 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.