ABSTRACT

FRANKLIN, ANTHONY M. Penalized Latent Variable Estimator for Finite Mixture of Regression Models. (Under the direction of Dr. Howard Bondell and Dr. Thomas Reiland).

When data structures are assumed from multiple sub-populations, selecting and estimating impactful covariates is a difficult challenge. More specifically in mixture of regression models, there are situations in which the impact of a given covariate is exactly the same in different sub-populations, and different in others. This creates a need for a flexible modeling technique that can compare across sub-populations. This research manuscript investigates implementing a penalization term to identify and collapse regression parameters that may be "shared" across clusters for a given model, referred to as a penalized latent variable regression analysis (PLVR). As a result, one can improve estimation and develop a more parsimonious model without sacrificing predictability.

Penalized Latent Variable Estimator for Finite Mixture of Regression Models

by
Anthony M Franklin

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2019

APPROVED BY:

Dr. Dave Dickey	Dr. John Griggs
Dr. Howard Bondell Co-chair of Advisory Committee	Dr. Thomas Reiland Co-chair of Advisory Committee

DEDICATION

This thesis would only be possible with the support of my family, friends and advisors. My mother has always beens supportive of my endeavors. My advisors Dr. Howard Bondell and Dr. Reiland, Dr. Griggs and Dr. Dickey have each been encouraging during the entire process. My wife and kids are my most precious loves in my life. I dedicate this work to my family.

BIOGRAPHY

I am married to a wonderful woman, and father of 3 children (plus 1 dog). I was born in New York City and raised in small town in South Carolina. I have a loving mother and brother. My mother once said: "when you write down your dreams, they become goals." In 7^{th} grade I wrote down, I wanted to get my doctorate. Today, I can hug my mother as a dream achiever.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Howard Bondell. You were always encouraging, patient and understanding. I am forever grateful for your support and guidance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

In many analytical applications, determining the impact of covariates to a given response is very important. When data structures are assumed from a single population, variable selection and estimation is a problem that has extensive research and development. Yet, when data structures are assumed from multiple sub-populations, selecting and estimating impactful covariates is more difficult challenge. Moreover, in cases where the impact of covariates to the response varies based on the sub-population provides another challenge. More specifically, there are situations in which the impact of a given covariate is exactly the same in different sub-populations, and different in others. This creates a need for a flexible modeling technique that can compare across sub-populations. Thus accurate estimation of these effects must be addressed intentionally and efficiently. The following chapters summarize a short list of available techniques designed to address this problem and their corresponding shortcomings. Furthermore, this research manuscript investigates implementing a penalization term to identify and collapse regression parameters that may be "shared" across clusters for a given model, referred to as a penalized latent variable regression analysis (PLVR). As a result, one improves estimation and develops a parsimonious model without sacrificing predictability.

Multiple linear regression (MLR) models are simple and often provide an adequate and interpretable description of how a set of input variables affect the response variable(s). For this reason, MLR models are frequently utilized in many applications.

When MLR is used for prediction purposes, it can sometimes outperform fancier nonlinear models, especially in situations with small sample sizes, low signal-to-noise ratio or sparse data. Yet

one inherent assumption MLR makes is that the all predictor variables have a synonymous effect on all subjects in the sample. Standard MLR can not account for situations when different predictors are important for different groups of subjects. In literature this is referred to as subject heterogeneity. When subject heterogeneity is present, estimating a single set of regression coefficients may be misleading, and most likely inaccurate. Subject heterogeneity often arises in economic and marketing. For example, in marketing analytics it is common to consider different impact of demographics for customer spending behavior.

To combat subject heterogeneity it is natural to first identify similar, homogeneous, groups within the data. The concept of identifying similarities is defined based upon the goal of the study. In statistical literature this is referred to as clustering. Clustering aims to do one of two things: group similar observations, based on some specified distance or proximity measure, and/or to discover patterns in data. Classical clustering techniques are associated with a distance metric and therefore often are limited to exploratory analysis. For example consider k-means clustering method which randomly partitions the data into k groups by initializing k centers and assigning observations to group associated with the closest center. Closest is typically defined by a squared error criterion between observations and centers. This algorithm updates by reassigning centers based on the means of the new groups created and converges when the squared error criterion decreases at an arbitrary rate which typically coincides with no further reassignments of observations to different groups. Another classical clustering example is hierchical clustering (HC). The idea of HC methods is to build a hierarchy of clusters based on a similarity measure of points. In the agglomerative method, the first step will look for the two most similar points and merge them to create a "pseudo-point". Each following iterative step will merge the two closest points (or "pseudo-points") to create the next "pseudo-point". This process is continued until the desired number of clusters, consisting of points and/or "pseudo-points", have been created. Hierarchical clustering methods can be displayed as a tree diagram, showing the relationships of all points. (See diagram $\star$) K-means and HC methods are popular due to their simple implementation but also face criticism for being too simple as well. K-means, hierchical clustering and other classical proximity clustering techniques do not directly contribute to variable selection and serve only as unsupervised learning techniques. Classical methods do not directly relate the predictor variables to the response variable while identifying groups. Hence, simple clustering methods generally are too limited and do not suffice for regression problems.

Clustering methods used for pattern recognition have increased in popularity due to the rise in the number of data mining problems. Data mining can be defined as the process of extracting useful information from large amounts of data. Clustering methods of this type are typically supervised learning techniques, thus the findings can be used to assess observations outside the original dataset supporting the findings. In this setting, homogeneous groups would correspond to groups of a similar pattern or meeting similar specified criterion, not just proximitiy. Furthermore, if the

grouping structure in the data set is known a-priori, then clustering methods become classification techniques since the objective is to solely assign observations to their respective group. If the true grouping structure is not perfectly known, one can assume such a structure (e.g. a distribution of the structure) or implement a flexible data-driven method. Hence clustering and classification techniques are often interchangable.

One example of a flexible non-parametric method is a classification and regression tree (CART). CART models are popular techniques in literature for the purpose of separating observations into subgroups by creating splits on predictor variables. A CART uses logical rules to create understandable splits and often result in subgroups homogeneous in the response variable. Although classification trees cluster using relationships of the predictor and response variables, the binary nature of logical rules can often fail to capture the true pattern in the data.

One example of a parametric clustering method is a mixture model. Mixture model clustering has grown in popularity with the increasing power of computational abilities. Mixture model clustering methods allow one to assume the individual structures of the groups to follow a specified distribution (i.e. Gaussian distribution). Each group is responsible for a proportion of the data, this proportion is represented by a mixture proportion parameter. Moreover, each observation is assigned to the cluster with the highest probability. Using probabilities to associate observations to clusters is referred as fuzzy clustering in the literature, as opposed to hard clustering like k-means methods.

For the purposes of identifying sub-population heterogeneity in a regression framework, a useful clustering technique must have the ability to group observations that fit the same pattern. More specifically, finding groups in the data focusing on a similar linear association between predictor and response variables. Classical clustering techniques alone are not adequate.

A naive analyst may attempt a two stage approach to solving the subject heterogeneity problem. First, the analyst would apply a clustering analysis to the data set to divide the observations into groups with similar characteristics, then perform regression fits for each cluster. The major problem with this method is that the clustering objective functions and regression objective functions are not necessarily related. Thus it would be better to integrate the cluster analysis into the regression framework, so that the parameters can be estimated simultaneously by optimizing one single objective function. This problem is referred to as clusterwise regression (CLR).

Chapter 2 will summarize some of the more popular techniques and advancements developed to solve the clusterwise regression problem. Chapter 3 will define and summarize finite mixture of regression models and extensions aimed to improve parameter estimation. Chapter 4 will introduce the penalized latent variable regression model estimation method, the focus of the research manuscript. Chapter 5 contains the results from the numerical analysis. Chapter 6 will conclude the manuscript and findings.

CHAPTER

2

LITERATURE REVIEW

2.1 Review of Clusterwise Regression Techniques

Clusterwise linear regression techniques can be categorized by two main approaches, algorithmic and likelihood approaches. For context, the term algorithmic implies a variation of a semi-exhaustive combinatorial method. That is, the method searches for the optimal combination of subjects and groupings from a potentially long list of such combinations based on a given criterion. Where the likelihood approach involves solving a likelihood based equation, although the solution may be provided via an iterative algorithm (i.e. EM Algorithm). The early attempts to solve the problem of sub-population heterogeniety began with parametric approaches that date back to the late 1960's. In the late 1970's algorithmic approaches to the problem began to arise, due to lack of computational power and inadequete parameter estimation in the maximum likelihood framework. In recent decades there has been an increased popularity in likelihood methods, due to technological advances and computer efficiency. The original work presented in this manuscript will focus on the likelihood based, mixture of regression models. First it is important to define historical techniques and identify the shortcomings which led to this research study.

2.1.1 Algorithm Approach

The most basic approach to solve a CLR problem is to use graphical tools to explore the data and assess appropriate associations. The limitations of graphical techniques and its usefulness lie the

inability to extend beyond a couple of covariates [MB88]. Multiple covariates are commonplace in the examples of interest, and graphical techniques are not feasible in such cases. Hence the need and development of numerical methods.

2.1.1.1 Spath Exchange Algorithm

One of the original non-parametric model approaches to clusterwise regression can be credited to H. Spath [DC88]. The techinique proposed by Spath focuses on partitioning the set of observations into a prespecified number of classes and establishing a regression model within each class. In describing this method, it is convenient to use similar notation and indices above. Consider the subject pair $(y_i, \mathbf{x}_i)$ where $\mathbf{x}_i$ is a J-column vector of covariates, with n subjects, J covariates, and m disjoint partitions. Define P_r as the r^{th} partition set, $\beta_j^{(r)}$ as the j^{th} covariate regression coefficient for the r^{th} partition. Lastly, M is the set of all observations, union of partitions. One can then express the objective function as:

$$\min\{\sum_{r=1}^{m} \min_{\beta_j^{(r)}} \sum_{i \in P_r} (y_i - \sum_{j=1}^{J} \beta_j^{(r)} x_{ji})^2\}$$

Spath aimed to find a partition that minimizes the sum of squared errors across all clusters. Analytically, one aims to find a partition that minimizes the above objective function. Moreover, this is accompanied with an exchange algorithm that transfers individual observations across clusters until no further improvement in the sum of squares criterion is achievable. The general idea for the method can be described by the following steps:

- Step 1: Choose some initial partition $P_1, \ldots, P_m$ where each $|P_r| \geq J$, and a starting observation $i = i_s$.
- Step 2: For each observation $i \in P_r$ and $|P_r| \geq J$ examine whether there are clusters P_u with $r \neq u$ such that shifting the observation i from P_r to P_u reduces the objective function. If so, then choose the shift that results in the maximal reduction of the objective function. Then adjust the associated clusters, $P_r := P_r - \{i\}$, $P_u := P_u \cup \{i\}$. Otherwise move to the next observation.
- Step 3: Repeat step 2 until no further reduction or until all observations have been examined.

This algorithm is reasonable when the number of observations are relatively large compared to the number of variables or when the hidden subpopulations are well separated. It has been noted that for very large sample sets (n), even local optima of the objective function often take unreasonable computing times. This exchange algorithm is known to converge to a locally optimal solution, thus it is recommended one uses multiple random starting solutions. The results depend on the initial partition as well as the initial starting observation i_s. A standard initial partition is to uniquely

evenly separate the observations into the desired subgroups. For example $C_j = \{i : i \in M, \ \ i \equiv j(mod\ n)\}$, choosing multiple i_s values and J_n. Moreover, notice for $J = 1$ and $x_{1i}\,(i = 1,\ldots,n)$ the objective function reduces to the minimum variance criterion in one dimension. One could see the natural extension of the objective function to L_p norms. From Spath's original exchange algorithm, numerous other extensions and competing methods have been developed. Yet, the clear major shortcoming is the inefficient iterative nature as well as the inability to identify congruent impacts of covariates across different groups. [Spa82].

2.1.1.2 Cherry Picking Algorithm

Banks, House, and Killourhy propose a non-parametric approach to clusterwise regression cases referred to as the cherry-picking algortihm. The algorithm is aimed to address situations in which the observed data can be described by multiple simple models. Moreover, each of the models may be described by a subset of the proposed list of covariates. In fact, each of the models may use a different set of covariates, and if the models use the same covariates, the magnitudes are not necessarily equivalent. The authors define this case as mixture sparsity. The authors aimed to avoid the pitfalls of Bayesian and frequentist methods that require sensitive model specifications. Such specifications can lead to inaccurate analyses when minor deviations from the model assumptions arise. Hence, the motivation for a more exploratory approach. In a nutshell, the algorithm will identify a few observations as a starting base, and then "cherry-pick" other observations based on these select few until a set has been formed. If the resulting set is of a sufficient size and decent fit, then the set is removed for a separate analysis.

As an example, consider an observed set of data points that originate from three true populations with noise as follows:

$$Y \sim \left\{ \begin{array}{ll} x+\epsilon_1 & \text{w.p. } p = .4 \\ 1-x+\epsilon_2 & \text{w.p. } p = .4 \\ u \sim unif(0,1) & \text{w.p. } p = .2 \end{array} \right\}$$

where $\epsilon_1 \sim N(0, 0.04)$ and $\epsilon_2 \sim N(0, 0.09)$.

The example dataset consists of two subgroups corresponding to opposite oriented lines, distorted by random noise. The idea of the algorithm is if one can identify at least three points that come from one of true linear groups, then the remaining points in the linear group can be swept in by choosing the points that lie closest to the fitted line formed by the selected points. The algorithm doesn't solely rely on the initial estimated line from the three points, it refits the line after each iteration of sweeping points. Although some unwanted points may be included in the group, i.e. where the lines intersect, ideally the impact would be minimal, given a sufficient dataset. After the first group is selected, the associated points can be removed from the dataset and the analyst would search for the next group by initializing with new points. On the contrary, if the points selected do

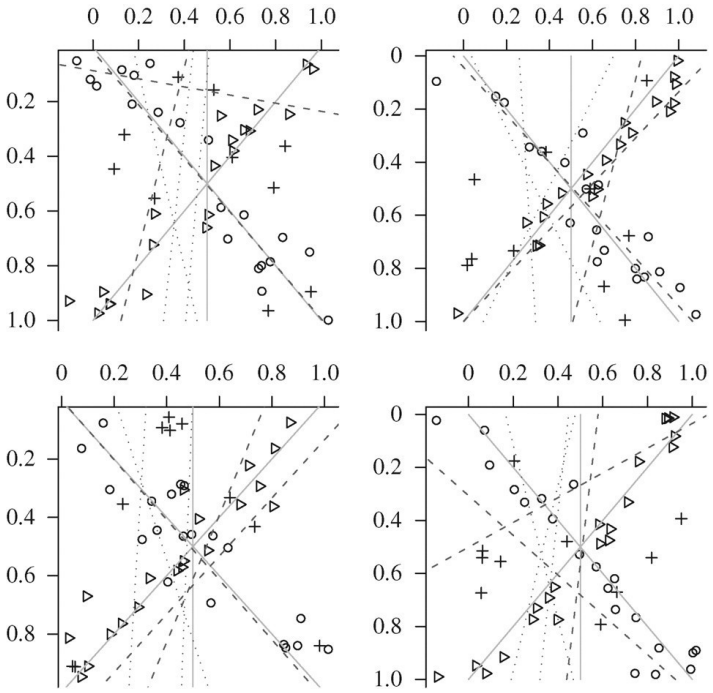

Figure 2.1 The figure shows y vs. x for four simulated datasets composed of a mixture of three components. The dashed lines indicate the fitted cherry-picking regressions, and the dotted lines are the Bayesian mixture model regressions. Points plotted as triangles are generated from the first component, points plotted as circles are generated from the second component, and points plotted as plus signs are generated from the uniform component.

not all come from the same linear group, then it is likely that the line fitted to the initial points has a residual variance too large or too small. In this case of a large residual variance, a large portion of the data points in the set will be swept into the grouping and dilute the ability to distinguish different groupings. Another issue that could arise from poor selection of initial points is the residual variance could by chance be too small. In this case, the number of points selected for the grouping would be too small. Either case would be considered a failure and as a result the algorithm suggests selecting multiple initialization points. Groupings are expanded by selecting points that do not decrease the coefficient of determination or that fall within a prespecified confidence band of the estimated line. Once the groupings have been selected, the remaining points are considered noise and ignored. Here is the breakdown of implementation:

1. Randomly select three points. Fit a line and calculate the R^2 (or some other model fit statistic).
2. Expand the initial set by adding all points that individually do not decrease the current value of R^2.
3. Refit the line to this expanded set. Expand the set again by adding all point whose residuals fall within the 80% confidence band. Repeat this step until no new points can be added.
4. If the set exceeds approximately 25% of the data, then declare this set to be a component and remove from the overall population.
5. Return to step 1 and repeat until the desired number of components is found.

For the above scenario, the authors show that the cherry picking algorithm does a better job of finding the mixture components than Bayesian mixture model with pre-specified priors. 2.1 shows nine simulated datasets composed of the above sample mixture. Visually, these set of plots indicate that the cherry-picking algorithm is competitive with Bayesian mixture modelling.

This procedure can be extended to fitting multiple linear regression models for subgroups. As the dimension of the models increase, the size of the starting subsample must also increase. For a J covariate model, $J+2$ points are needed to calculate a goodness of fit estimate. The chance of selecting a starting subsample drawn completely from the same subgroup decreases exponentially. It is important to note that this algorithm is an exploratory procedure, thus tuning is essential. For example, the authors even suggest generating thousands of initial triples and observing the distribution to help determine appropriate cutoffs for a sufficient residual variance.

The cherry-picking concept has many parallels and can be viewed as an extension of elemental set theory. [Ban09] In the framework of robustness, the cherry-picking algorithm is similar to an S-estimator, thus using a minimum volume region to subset the data and model correspondingly [McL82]. Notice also the cherry-picking algorithm inherently is addressing the complex variable

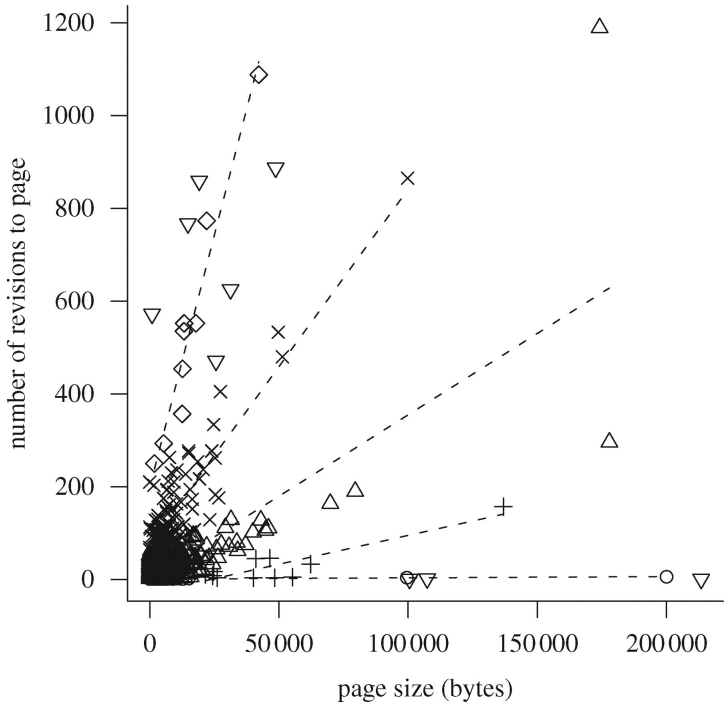

Figure 2.2 Here is an example of multiple linear regression fits using the cherry-picking approach. Superposed on the data plot are five lines identified by the cherry-picking algorithm. Different symbols distinguish the different lines; the triangles that point upwards denote points that were not assigned to any of the five lines.

selection problem by selecting only covariates that are effective in group amassing. The shortcoming of this technique is its inability to scale to problems involving large sets of covariates. In those cases, this technique can become very unstable and local minimums found are not trustworthy. Furthermore, this method does not explore cross-group relationships and commonalities.

In the framework of non-parametric methods, Muruzabal, Vidaurre and Sanchez [Mur12] implemented the unsupervised learning technique, self-organizing maps (SOM) as well as multilayer perceptron (i.e. neural networks) in the CLR framework. The authors title the method SOMwise regression (SOMwiseR) and propose a supervised learning procedure by clustering on the concatenated data set of the predictors and response variables. The prediction results are promising, yet it does not address direct coefficient estimation.

Before mentioning parametric methods, hybrid methods have been developed as well. Ciampi proposed a locally linear regression model based on tree-growing [Cia07]. This method is a heuristic algorithm that constructs classification trees with a linear regression at each of the leaves. The aim is to partition the dataset and fit regressions on each subset, then the global fit of these local models is optimal. The local regression aspect of this technique can account for unique cluster impacts of covariates. Yet, the disjoint nature of the partitions does not account for cross-group commonalities.

2.2 Likelihood Approach

A popular parametric approach to the CLR framework is the finite mixture of regression model (FMR) formulation. FMR models are known for being flexible and powerful probabilistic modeling tools that can account for unobserved sub-population heterogeneity. One assumes the response variable measures are obtained from a mixture of conditional densities that arise in unknown proportions. These densities represent the unobserved sub-populations, these sub-populations will be called components of the model. The densities are pre-specified, and the appropriate forms will be discussed below. These distributional assumptions would imply one needs to select an appropriate approach to parameter estimation. Likelihood based approaches are most commonly used in the literature of FMR problems. Section 3 will summarize the FMR techniques [FJ02].

2.2.1 Finite Mixture of Regression Models

Finite mixture models have been extensively researched and have a wide variety of applications. FMR models are a special case of finite mixture models. In the beginning when finite mixture methods were introduced the parameters of the mixtures were estimated using method of moments techniques as well as a heavy focus on graphical approaches. Maximum likelihood estimation was later shown to be a far superior technique. More specifically, by utilizing finite conditional mixture distributions with exponential family distributions, one can estimate the regression parameters

using maximum likelihood procedures. One then forms the log-likelihood function from a sample of observations and maximizes the log-likelihood function subject to a summability constraint on the mixing proportions $\sum_k \pi_k = 1$. The optimization problem can be solved using a list of procedures, such as gradient methods, Gibbs samplers, and the EM-algorithm. While gradient methods such as the Newton-Raphson method may converge in a relatively small number of iterations and provide asymptotic variances for parameter estimates, it is important to note that convergence is not guaranteed. On the other hand the EM-algorithm ensures convergence and has other attractive characteristics. The EM-algorithm has proved to be most popular of the two and will remain the focus in this manuscript. [Dem77]

In the finite mixture model framework, when an exponential family density is used to describe the components, a generalized linear model can be implemented as a general regression stucture. The most popular distribution studied in literature is the normal density. In the case of a finite mixture of normal densities the expectations of the densities are represented by linear functions of a set of explanatory variables (covariates). Also, the likelihood approach to the normal mixtures are much more feasible to work with and have easily derivable asymptotic properties. There are two genres of likelihood mixtures, unconditional and conditional. In the unconditional approach for finite mixtures (the infinite case is not considered in this manuscript), the data is random in nature while the mean and variance of the underlying normal density are not known and thus must be estimated. Conditional mixtures allow for the estimation of regression parameters as well as a posterior classification of observations. It is important to note that in both cases the number of components is determined a-priori. Furthermore, the described framework can be used in both a multivariate and univariate setting. This manuscript only focuses on the univariate case, although extension to the multivariate case is simple in notation and most parametric methods mentioned are applicable as well.

Finite mixture of regression models were introduced as early as 1972 by Quandt and Ramsey in a 'switching regression' framework. [QR73] Switching regression models are common in the econometrics literature. Consider

$$\begin{aligned} y_i = &\quad \textstyle\sum_{j=1}^{J} \beta_{1j} x_{ji} + u_{1i}, \\ &or \\ y_i = &\quad \textstyle\sum_{j=1}^{J} \beta_{2j} x_{ji} + u_{2i}, \end{aligned}$$

where u_{1i} and u_{2i} satisfy the classical assumptions made about error terms and are distributed as $N(0,\sigma_1^2)$ and $N(0,\sigma_2^2)$ respectively. Furthermore, if one assumes $(\beta_{1.},\sigma_1^2) \neq (\beta_{2.},\sigma_2^2)$ then the equations above can represent the states of switching between two regimes. This equation is applicable when it is fair to assume no information is known of how to classify the observations between the two regimes. Note, when the regime membership is known, regardless of the number of components, that information can be included in the model and there is no need for a mixture approach.

The initial approaches to solving the switching regression problem involved using method of moment estimators and moment generating functions. As the switching regression model became generalized to more than two groups, maximum likelihood methods were seen as more feasible and hence became more popular. [DC88] Desarbo and Cron investigated a mixture approach with conditional inputs, multiple components and a univariate response. Again assuming normally distributed components, the expected means can be represented as a linear function of a set of explanatory variables. Moreover, Desarbo and Cron estimated the conditional mixture models via the EM-algorithm (which will be described later). These findings have been considered the foundation of finite mixture of regression modeling.

Numerous mixture of regression models have since been developed, many simply interchanging the expected mean functions as well as the underlying component densities. [KR89] Kamakura and Russell investigated multinomial logit and probit regression models. Univariate poisson mixture regression models were propsed by Wedel, DeSarbo, Bult and Ramaswamy [Wed93]. Also, later in 1998 conditional multivariate normal regression mixtures were developed [DR94]. A myriad of distributions from the exponential family have been used to describe the patterns in sub-populations. In fact, the conditional mixture models mentioned above can be considered a special case of a mixture likelihood approach to a generalized linear models. [DW95] Next is to describe the framework of the finite mixture of regression models.

2.2.2 Mixture Approach of Generalized Linear Models

Consider a univariate random variable $\mathbf{y}$ that is assumed to come from a population composed of K sub-classes. Since it is not known in advance to which class each subject belongs, a mixture probability π is assigned to each respective sub-class. This manuscript works with the conditional case, hence one assumes the component assignment in order to describe the corresponding sub-class. The FMR model literature can be extended to the general linear model structure. In this manuscript, the component-wise distribution function is Gaussian and the mean function is a linear combination of proposed covariates. Many other distribution functions and corresponding mean functions have been proposed in the FMR model framework. Thus, it is beneficial to briefly define the generalized linear model framework as a foundation to the FMR problem, as well as mention other works in the FMR topic. This allows one to define the Gaussian distribution as a special case of GLM approach. Given the subject pair $(y_i, \mathbf{x}_i)$ where $\mathbf{x}_i$ is a $(J+1)$-column vector of covariates, with n subjects, J covariates. and K components.

One can express the conditional exponential family probability density function of observation i given sub-class k in the following form,

$$f_{i|k}(y_i|\theta_{ik}, \lambda_k) = e^{(y_i\theta_{ik} - b(\theta_{ik})/a(\lambda_k) + c(y_i, \lambda_k))}$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are specific functions coupled with the canonical parameters θ_{ik}, mean function μ_{ik}, and dispersion parameter λ_k used to identify the appropriate density function from the exponential family for the k^{th} component. The dispersion parameter is assumed to be known and constant over the observations within sub-class k and $a(\lambda_k) > 0$. Notice both the canonical and dispersion parameter in this case depend on the conditional sub-class k. [DW95] McCullah and Nelder (1989) give a more thorough description of the exponential family characteristics. For instance, if λ_k is unknown then the distribution may not be a member of the exponential family.

A generalized linear model (GLM) was defined to generalize a suite of linear regression techniques, such as ordinary linear regression, Poisson regresssion and logistic regression. General linear models introduce a link function that uniquely relates the linear predictor to the response variable. Generalized linear models also allow models to incorporate error distributions that are not normally distributed. In reference to the link function, in the linear model framework, consider a linear predictor η_{ik} that consists of a linear combination of J covariates $\mathbf{x}$ as such:

$$\eta_{ik} = g(\mu_{ik}) = \sum_{j=1}^{J} \mathbf{x}_i^T \beta_{jk}.$$

Where $g(\cdot)$ represents the link function. This predictor represents the function of the distribution mean μ_{ik} and is conditional on the k^{th} sub-class. In the case of the normal distribution the link is the identity function. In the case of the Poisson and binomial distribution, the canonical links are the log and logit, respectively. Thus, one would formulate the generalized linear model by specifying the distribution of the random response variable y_i, a linear predictor η_{ik} and a link function $g(\cdot)$.

Now, consider a parameter vector $\boldsymbol{\theta}^T = \{\boldsymbol{\pi}^T, \boldsymbol{\beta}^T, \boldsymbol{\lambda}\}$ where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$,$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$,$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$. The resulting finite mixture model would look of the following form:

$$f_i(y_i|\boldsymbol{\theta}) \sim \quad \sum_{k=1}^{K} \pi_k f_{i|k}(y_i|\mathbf{x}_i, \boldsymbol{\beta}_k, \lambda_k),$$

where the specified normal distribution would have the following density,

$$f_{i|k}(y_i|\mathbf{x}_i, \boldsymbol{\beta}_k, \lambda_k = \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} exp\{-\frac{1}{2\sigma_k^2}(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k)^2\}.$$

After defining the density, in order to estimate the parameters of interest it is necessary to formulate the likelihood function:

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} \ln\left(\sum_{k=1}^{K} \pi_k f_{i|k}(y_i|\mathbf{x}_i, \boldsymbol{\beta}_k, \sigma_k^2)\right) \tag{2.1}$$

where n represents the number of observations. This formulation is not particularly easy to work with, thus a log transformation is often preferred.

$$\ell_n(\boldsymbol{\theta}|\mathbf{y}) = \ln(L) = \sum_{i=1}^{n} \ln\left(\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi_k \sigma_k^2}} exp\{-\frac{1}{2\sigma_k^2}(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k)^2\} \right) \tag{2.2}$$

Estimates for the parameter vector $\boldsymbol{\theta}$ can be found by maximizing the above likelihood function, still maintaining the summability constraint on the mixture proportions π_k. Now combining the estimates of $\boldsymbol{\theta}$ with the use of the Bayes rule, one can derive a posterior probability estimate for each observation which can then be used to assign an observation to a class. Define the posterior probability for the i^{th} observation and k^{th} class as the following:

$$p_{ik} = \frac{\pi_k f_{ik}(y_i|\mathbf{x}_i, \boldsymbol{\beta}_k, \sigma_k^2)}{\sum_{k=1}^{K} \pi_k f_{ik}(y_i|\mathbf{x}_i, \boldsymbol{\beta}_k, \sigma_k^2)}, \tag{2.3}$$

This implies one could assign observation i to class k using the following rule:

$$p_{ik} > p_{ij} \ \forall \ k \neq j.$$

Mixture models consisting of exponential family distributions are said to be identifiable if unique values of the parameters determine distinct members of the family of mixtures. Lack of identifiability makes estimation meaningless. In the case of mixtures of regressions, due to the nature of the mean function as a function of covariates in each component, there is a threat to assuring identifiability since many possible arrangements of the classes may exist. McLachlan and Basford (1988) have shown that the likelihood of a finite mixture is invariant under the permutation of labeling of the components. Hence, it is best practice to propose and work with only one arrangement to assure identifiability.

In mixture models the parameter list primarily consists of mixture component weights and the component specific parameters that are associated with the distributions Ω. For identifiability of the mixture model there must be a unique parameter vector in the parameter space. Consider the K component mixture, to prevent identifiability problems a couple conditions must be set for the parameter vector. The mixture component weights must be greater than zero, $\pi_k > 0 \ \forall k = 1, \ldots, K$. This condition helps prevent overfitting and eliminates problems involving empty sets, which can cause identifiability issues. Also, no two components can have equal specific parameter sets. That is $\forall k \neq u \in \{1, \ldots, K\} \Rightarrow \boldsymbol{\theta}_k \neq \boldsymbol{\theta}_u$. This condition prevents identifying a separate component that should be combined as one.

In mixture model literature, there is another concept to consider called label switching. Label switching is referring to the idea that a given K component finite mixture model has at least $K!$ parameterizations, due to the possible permutations of the K components. By definition of model identifiability, one must establish a unique representation for all equivalence classes in the model space. In this context, the analyst must specify a subset of parameterizations which contain only

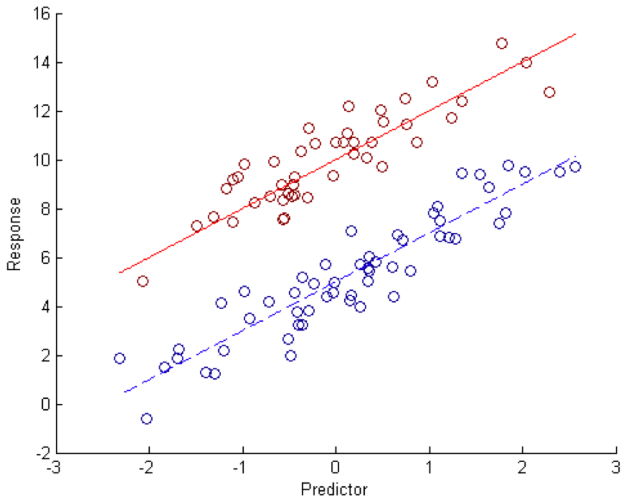

Figure 2.3 Here is an example of multiple sub-populations that each have linear patterns. The true linear equation is superimposed onto the corresponding sub-population data points. These sub-populations represent different sub-components of a mixture model of regressions.

one permutation of the set of component parameters $\mathbf{\Omega}^* \in \mathbf{\Omega}$. By imposing an ordering constraint on the components, with respect to a combination of parameters, one can help reduce the overall potential for model identifiability.

Once a subset parameterization space has been defined, one can further examine identifiability problems, in literature this is referred as generic identifiability. Generic identifiability is guaranteed for normal finite mixture regression models barring a couple considerations. Such as, the full column matrix $\mathbf{X}$ must be full rank [Tei63]. This assumption certainly holds for mixtures where the component distributions are from the same family, as used in this research. It is worth noting, all components are not required to share the same density function. The scenario above using the univariate conditional normal density in the finite mixture of regression model is a method introduced and explained in great detail by [DC88]. The EM-algorithm is a common technique used to obtain maximum likelihood estimates for the regression and variance parameters of each component simultaneously. The details of the algorithm will be described in the next subsection.

2.2.3 EM Algorithm

To solve the mixture model equations, the most common technique is the EM-algorithm. The EM-algorithm was first introduced in 1977 by Dempster Laird and Rubin [Dem77]. The EM-algorithm is an iterative method for solving likelihood functions that depends on unobserved latent variables. The key to implementing the EM-algorithm is identifying how to incorporate the latent variable. In a mixture regression framework, the non-observed variable z_{ik} represents the observation's

appropriate component association. More specifically, z_{ik} is the indicator for the i^{th} observation belonging to the k^{th} component. If observation i belongs to component k then $z_{ik} = 1$. In general, consider a random vector $\mathbf{z}_i$ that is independently and identically distributed multinomial with n observations and probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$. That is $\mathbf{z}_i \sim multi(K, \pi_1, \pi_2, \ldots, \pi_K)$, where $\sum_u \pi_u = 1$. Then the density can be written

$$f(\mathbf{z}_i|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{ik}},$$

where $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iK})$. Incorporating this latent variable allows the "conditional" property to be expressed notationally. That is, the response variables $y_i|\mathbf{z}_i$ are conditionally independent and the corresponding density can be expressed:

$$f(y_i|\mathbf{z}_i) = \prod_{k=1}^{K} f_{i|k}(y_i|\boldsymbol{\beta}_k, \sigma_k^2)^{z_{ik}},$$

Now, incorporating all of the data, the complete log-likelihood function can be written:

$$\ell_c(\boldsymbol{\theta}; \mathbf{y}, \mathbf{Z}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln(f_{i|k}(y_i|\boldsymbol{\beta}_k, \sigma_k^2)) + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln(\pi_k). \tag{2.4}$$

To solve the complete data log-likelihood one would use the iterative EM-algorithm. Before diving into the details, here is how the algorithm is set up. The algorithm is broken into two steps, expectation (E-step) and maximization (M-step) step. The E-step incorporates the expectation of the complete log-likelihood derived above. The expectation of the likelihood is conditional in nature, such that it uses suggested estimates of the parameter vector $\boldsymbol{\theta}$ while the response vector is held fixed as well. The M-step of the derived expectation equation from the E-step is maximized with respect to the parameter vector $\boldsymbol{\theta}$ to calculate an updated vector of estimates. The algorithm continuously alternates between the E and M-steps until a certain criterion is met in relation to optimal solution of the likelihood function. It has been shown the EM-algorithm produces log-likelihood values that increase monotonically. Solving the generalized linear mixture model above is very similar to solving the ordinary mixture model equation except for the individual component wise likelihood functions used in the M-step. Next, a summarized explanation of the E and M-steps below.

2.2.3.1 E-Step

In the E-step of the algorithm, one derives the expectation of the complete likelihood function with respect to the unobserved latent variable $\mathbf{z}$ conditioned on the observed responses $\mathbf{y}$ and the vector of suggested estimates $\boldsymbol{\theta}$. That is, one maximizes $E[\ln_c(\mathbf{Z}|\boldsymbol{\theta}, \mathbf{y})]$ which is expressed as $E[\ell_c(\boldsymbol{\theta}; \mathbf{y}, \mathbf{Z})]$ in common likelihood function notation.

$$
\begin{aligned}
E[\ell_c(\boldsymbol{\theta};\mathbf{y},\mathbf{Z})] = & \quad E\left[\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\ln(f_{i|k}(y_i|\boldsymbol{\beta}_k,\sigma_k^2)) + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\ln(\pi_k)\right] \\
= & \quad \sum_{i=1}^{n}\sum_{k=1}^{K} E[z_{ik}|\boldsymbol{\theta},\mathbf{y}]\ln(f_{i|k}(y_i|\boldsymbol{\beta}_k,\sigma_k^2) \\
& + \sum_{i=1}^{n}\sum_{k=1}^{K} E[z_{ik}|\boldsymbol{\theta},\mathbf{y}]\ln(\pi_k).
\end{aligned} \tag{2.5}
$$

The above equation shows there is a need for an estimated conditional expected value of the latent variable z_{ik}, the individual observation component membership probability. To calculate this expectation, Bayes Theorem can be applied and the conditional distribution of the response $\mathbf{y}$ given the latent variable $\mathbf{Z}$ must be included. The Bayes Rule can be used as follows:

$$
E[z_{ik}|\boldsymbol{\theta},\mathbf{y}] = \frac{\pi_k f_{i|k}(y_i|\boldsymbol{\beta}_k,\sigma_k^2)}{\sum_{k=1}^{K}\pi_k f_{i|k}(y_i|\boldsymbol{\beta}_k,\sigma_k^2)}
$$

The current estimates of this posterior probability use the previous iteration estimates of β_k, σ_k^2. Notice the component membership probability is synonymous with the posterior probabilities (2.3) denoted π_{ik}.

2.2.3.2 M-Step

To begin the M-step, one replaces the expected latent variable equation with the estimated posterior probabilites $\hat{p}_{ik}$ from equation (2.3). The mixture probabilities must also be replaced. In this framework, one would solve for the mixture proportions $\hat{\pi}_k$ by using the posterior probability estimates $\hat{p}_{ik}$.

Desarbo and Cron show that one way of solving for the mixture proportions is appending a Lagrange multiplier to the expected complete likelihood equation (2.5) and imposing the summability constraint on the mixture proportions.

$$
E[\ell_c(\boldsymbol{\theta};\mathbf{y},\mathbf{x},\mathbf{Z})] - \mu[\sum_{k=1}^{K}\pi_k - 1] \approx \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{p}_{ik}\ln(\pi_k) - \mu[\sum_{k=1}^{K}\pi_k - 1].
$$

where μ is the Lagrangian multiplier. Notice the first term of the complete likelihood function can be dropped with respect to the mixture proportions. Optimizing the above equation with respect to the mixture proportions yields

$$
\hat{\pi}_k = \frac{\sum_{i=1}^{n}\hat{p}_{ik}}{n}.
$$

One thing to notice is the derivations of both the posterior probabilites and mixture proportions are not dependent upon the distributional assumption of the latent variable z. Hence, it will not be referenced after this point.

Next is to maximize the final derived expectation of the complete log-likelihood equation with respect to the regression parameters given the provisional estimates. That is:

$$\max_{\boldsymbol{\beta},\sigma^2} E[\ell_c(\boldsymbol{\beta},\sigma^2;\mathbf{y},\mathbf{Z},\hat{\boldsymbol{\beta}},\hat{\sigma^2})]$$

The maximization of the above equation results in a series of equations that can be simultaneously solved, one for each respective component k. More specifically, one would solve the following K equations with respect to the corresponding regression parameters:

$$G_k = \sum_{i=1}^{n} \hat{p}_{ik} \ln(f_{i|k}(y_i|\boldsymbol{\beta}_k,\sigma_k^2)).$$

In a regression framework, solving this equation is equivalent to a single group generalized linear model problem but incorporating the complete data. Moreover, each observation is weighted using the posterior probabilities to the corresponding component k. From this point, each regression parameter can be derived individually by solving partial derivative equations. One would solve for the regression parameters as follows:

$$\frac{\partial G_k}{\partial \boldsymbol{\beta}_{jk}} = \sum_{i=1}^{n} \hat{\pi}_{ik} \frac{\partial \ln(f_{i|k}(y_i|\boldsymbol{\beta}_k,\sigma_k^2)}{\partial \boldsymbol{\beta}_{jk}} = 0.$$

and

$$\frac{\partial G_k}{\partial \sigma_k^2} = \sum_{i=1}^{n} \hat{\pi}_{ik} \frac{\partial \ln(f_{i|k}(y_i|\boldsymbol{\beta}_k,\sigma_k^2)}{\partial \sigma_k^2} = 0.$$

The GLM notation and details for the resulting equations can be found in [DW95]. At this point the general notation of the EM algorithm has been introduced. Next, it is appropriate to introduce the iterative nature of the algorithm. All of the previous steps described in the algorithm start with an initial estimate. Each cycle through the steps updates the estimates in efforts to reach an "optimal" solution. Here is a summarized explanation of the algorithm:

1. At the first iteration, $t = 0$, the procedure begins with a set number of components, and an initial set of parameter estimates $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(0)}$. The initial parameter estimates are used to compute the observation component probabilities $\hat{p}_{ik}$. Component probabilities could be computed first, or a partition could be used to find the initial parameter estimates of each component using the resulting equations from the maximization step above.

2. Now given an initial set of estimates and the sample set, the next objective is to maximize the updated expectation equation from the E-step using the current set of estimates. That is,

$$\max_{\boldsymbol{\beta}^{(t+1)},\boldsymbol{\sigma}^{2(t+1)}} E[\ell_c(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\sigma}^{2^{(t+1)}};\mathbf{y},\mathbf{Z},\hat{\boldsymbol{\beta}}^{(t)},\hat{\boldsymbol{\sigma}}^{2^{(t)}})].$$

The superscripts depict the iterative nature of the algorithm. So deriving the regression esti-

mates in the iterative algorithm becomes a psuedo re-weighted least squares problem. The parameter estimates are found via a maximum likelihood method as described above in the previous subsection.

3. After each iteration and updated estimates are computed, a convergence criterion is calculated based on the resulting likelihood values as such: $\|\ell(\theta^{(t+1)}|\mathbf{y}) - \ell(\theta^{(t)}|\mathbf{y})\| < \epsilon$ where ϵ is sufficiently small.

4. If the convergence criterion is not met, then return to step two and recompute the posterior probabilities for the observations and proceed as instructed.

There have been many advancements to the EM algorithm since its inception. Many extensions are adaptations of the maximization approach. The following sections describe the common EM algorithm extensions.

2.2.4 EM Algorithm Extensions

2.2.4.1 Classification EM

One EM approach is referred to as the classification version (CEM). The CEM algorithm essentially removes the fuzzy weighting in the GLM equations above and replaces them with hard assignments. Then the maximization step is continued with specific component groupings and sizes that correspond to the number of observations assigned to the component. [Mar00] Moreover, within the pseudo partitions, the posterior probabilities are still used as weights when computing the corresponding parameter estimates. In the original EM approach, the hard assignment is done once when the algorithm has converged and the user wants to measure the misclassification error rates or other diagnostics. The E-step of the CEM algorithm is identical to the E-step of the original EM algorithm described above.

In the classification step the observed sample set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ is partitioned into K mutually exclusive groups $\mathbf{P} = (P_1, \ldots, P_K)$ where the groups are created by assigning observations based on the individual largest posterior probability at the given iteration. That is:

$$P_u^{(r+1)} = (\mathbf{x}_i, y_i) : \hat{p}_{iu}^{(r)} = a\,r\,g_h \max \hat{p}_{ih}^{(r)}$$

where r is the iteration and if $\hat{p}_{iu}^{(r)} = \hat{p}_{ih}^{(r)}$ and $u < h$ then $(\mathbf{x}_i, y_i) \in P_u^{(r+1)}$. If the maximum posterior probability is tied with component, the component with the smallest index is chosen as the tie breaker. This type of classification now has the potential for empty sets. If a component does not have any obervations, then problem is reduced to $K-1$ components and the algorithm starts with the new mixture structure.

The M-step has the same objective to maximize the objective function consisting of the expectation equation, except instead of a total sample weighted least squares type problem, each parameter is estimated using the sub-sample consisting of the assigned partitions above. The mixture proportions can be calculated as such,

$$\hat{\pi}_u = \frac{n_u}{n}$$

where n_u is the number of observations assigned to component u. One can then show that solving for component level parameter estimates can be written explicitly as

$$\hat{\beta}_k^{(t+1)} = (X_k^T W_k X_k)^{-1} X_k^T W_k \mathbf{y}_k$$

$$\hat{\sigma}_k^{2^{(t+1)}} = \frac{\sum_{i=1}^{n_k} \hat{p}_{ik}^{(t)} (y_i - x_i^T \beta_k^{(t+1)})^2}{\sum_{i=1}^{n_k} \hat{p}_{ik}^{(t)}}$$

where X_k is a $n_k \times (J+1)$ matrix of predictors for the kth component, W_k is a $n_k \times n_k$ diagonal matrix with diagonal entries $\hat{p}_{ik}^{(t)}$ and $\mathbf{y}_k$ is a $n_k \times 1$ vector of responses for the k^{th} component. The CEM algorithm cycles through the three steps repeatedly until the prespecified convergence criterion is met.

The addition of the classification step in the algorithm makes this approach have a K-means like technique, where the computed posteriors are used as similarity measures and hard assignments to components like clusters. Simulations studies have shown that when the true values are used as starting values, the CEM method consistently converges in fewer iterations than the original EM algorithm. The authors of this technique use the method to show improved accuracy of parameter estimation in mixture of regression models. More specifically, when a subset of proposed covariates have similar effects or parallel linear subspaces, this method improves accuracy of parameter estimation. These situations are the focus of this research and hence the attention to this approach. Yet this technique does not explicitly share information across components. [Bie03]

2.2.4.2 Stochastic EM

Faria and Soromenho also investigate the stochastic version of the EM algorithm (SEM). This technique incorporates a random component by implementing a stochastic step between the E-step and M-step of the algorithm. This random part consists of simulating the unobserved latent variable $\mathbf{z}_i$ and randomly drawing the observations from their updated conditional distribution. That is, a partition is formed $\mathbf{P} = (P_1, \ldots, P_K)$ where observations are assigned to the components at random according to the multinomial distribution of the current latent observation variable $\mathbf{z}_i^{(t)}$. So instead of directly using the derived posterior probabilities, the observational component probabilities are drawn at random using the posteriors as distributional inputs. This can be expressed

as:

$$\mathbf{z}_i^{(t+1)} \sim multi(\hat{p}_{i1}^{(t)}, \hat{p}_{i2}^{(t)}, \ldots, \hat{p}_{iK}^{(t)}).$$

The resulting $\mathbf{z}_i$ draw represents the observational assignment to the associated component. The maximization step is the same as the CEM step described above. The partitions are used to define K separate weighted least squares equations with a sub-sample size n_u for $u = 1, \ldots, K$.

It is important to note that the SEM approach does not converge point-wise. It does create a stationary distribution that has Markov chain properties and is concentrated around the MLE. Simulations have shown when random starting values are used, the SEM outperforms the CEM and traditional EM for observation misclassification errors, when a subset of proposed covariates have similar effects.

2.2.5 Number of Components

When applying mixture models to actual data, the true number of components, K, is rarely known. Therefore, it must be estimated. The inference behind estimating the number of components is not a well developed problem and has some challenges. Throughout the literature, the inference on the number of components K in mixture models has separated the estimation from the model fitting. This manuscript focuses on a likelihood approach, thus in reference to that framework there are two main approaches to estimating the number of components in a mixture regression model. One method is a penalized form of the likelihood. Another method is to perform a hypothesis test on the number of components. There are challenges to both approaches. Before describing both approaches, one must consider the general perspective of this problem. Conceptually, there are an infinite number of distributions to describe any given fixed number of components for a finite mixture model. Thus, it is reasonable to consider addressing the problem as the minimum number of components in the mixture compatible with the data and with sufficiently large mixture proportions.

Consider performing a hypothesis test on the number of components where the null-hypothesis $\mathbf{H}_0$ is that the population consists of K components and the alternative $\mathbf{H}_A$ asserts $K+1$ components. That is:

$$\mathbf{H}_0 = K$$
$$\mathbf{H}_A = K+1$$

Using MLEs to estimate the parameters of each given space respectively, the likelihood ratio statistic is equivalent to $LR = -2\log(\lambda) = -2\log(\theta_0 - \theta_A)$. One issue with this test is that the null hypothesis is at the boundary of the parameter space of the alternative hypothesis, which implies that the regularity conditions do not hold at the null hypothesis. Therefore, the generalized likelihood ratio statistic is potentially not asymptotically distributed chi-square with degrees of freedom equal to the difference in number of parameters under the two hypotheses spaces. Thus, this traditional

test is not dependable [Tit90].

Using the penalized approach, the log likelihood function is penalized typically by subracting a term representing the complexity of the model. This complexity can represent the function of the number of parameters in it, like the AIC or BIC. The Akaike's Information Criterion (AIC) selects the model that minimizes the following expression

$$-2\log(L(\hat{\theta}+2d))$$

where d is the number of parameters in the model. Although the asymptotic assumptions used to derive the expression do not always hold for the same reasons the LRTs are not always valid, (basically the regularity conditions breakdown) the AIC remains popular in literature. Many authors have studied and noted that the AIC tends to overestimate the correct number of components. [CS96].

The Bayesian Information criterion (BIC) developed by [Sch78] is expressed as

$$-2\log(L(\hat{\theta})+d\log(n)))$$

where n is the number of observations in the sample. Again, with this criterion, regularity conditions breakdown due to the mixture model nature. Despite this, literature still supports the use of the BIC as a legitimate criterion. Research has shown that the BIC does not underestimate the true number of components asymptotically. Research has also shown that the penalty term of BIC penalizes complex models more heavily than AIC. Although, it has also been shown that for small sample sizes, the BIC may suggest too few components.

In general, penalized likelihood approaches are less demanding than the likelihood ratio tests. There have been many other tests that have been proposed for estimating the number of components, including but not limited to methods incorporating score statistics or Monte Carlo estimations. In this study, penalized likelihood approach is used for estimating the number of components.

2.2.6 FMR Extensions

This section will list and briefly describe some extended work in the finite mixture of regression model framework.

The maximum likelihood approach to mixture regression models is based on the assumption of normally distributed errors. Similar to a traditional regression case, the normality assumption is sensitive to outliers or heavy-tailed error distribution behavior. In fact, a single outlier, if far enough away, can dramatically affect the MLE. To combat this, robust methods have been proposed to enhance estimation. In general, robust fitting for mixtures of location and scale family distributions are well developed in literature. Two popular approaches to be mentioned are fitting mixtures using

the trimmed likelihood estimator and mixture of regression modeling using t distributions.

2.2.6.1 Trimmed Likelihood Estimator

Neykov et al developed the trimmed likelihood approach, an extension of the traditional maximum likelihood estimation method. In short, the trimmed likelihood estimator (TLE) looks for a subset of observations out of the sample whose likelihood is maximal. The basic premise being to remove a subset of observations whose values would be highly unlikely to occur if the fitted model were indeed true. Below is the expression for the weighted trimmed likelihood estimator (WTLE) [HL97]

$$\hat{\theta}_{WTLE} \equiv \arg\min_{\theta \in \boldsymbol{\theta}^J} \sum_{i=1}^{m} w_{\nu(i)} f(y_{\nu(i)}; \theta),$$

where $f(y_{\nu(1)};\theta) \leq f(y_{\nu(2)};\theta) \leq \ldots \leq f(y_{\nu(m)};\theta)$ for a fixed θ, $f(y_{\nu(i)};\theta) = -\log(h(\mathbf{y};\theta)$, for $i = 1,\ldots,m$ i.i.d. observations that have a probability density $h(\mathbf{y};\theta)$ with $\theta \in \boldsymbol{\theta}^J \subset \mathscr{R}^J$ as the unknown parameter. The vector corresponding to the permutation indices $\nu = (\nu(1),\ldots,\nu(n))$ depends on the component parameter vector θ, while m, $(m \leq n)$ is the trimming parameter and $w_i \geq 0, \forall i$ are nondecreasing functions dependent on the mixture model $f(\cdot)$.

The idea of the TLE is to identify and remove the $n-m$ observations with smallest likelihood values. With this in mind, all possible $\binom{n}{m}$ partitions of the data must be fitted by the MLE. Hence, the WTLE is given by the partition with the resulting negative log-likelihood is minimal. The WTLE expression is a more general expression and the TLE is a special case such that $w_{\nu(i)} = 1$ for $i = 1,\ldots,m$ and $w_{\nu(i)} = 0$ otherwise. Moreover, it can be shown that when $h(\mathbf{y};\theta)$ is the normal regression error density, the TLE coincides with the least trimmed squares estimators. It is important to note, directly computing the TLE is expensive for large datasets because of the exhaustive combinatorial nature.

2.2.6.2 Mixture of T-Distributions

It is well known that normally distributed errors lack in capturing heavier tailed activities. In traditional linear regression, if the variance is unknown or if the normality assumption is not sufficient, then a popular alternative is the t distribution. [MP00] proposed extending the mixture model estimation using the t-distribution. Then Yao, Wei and Yu implemented the t distribution in a finite mixture of regressions framework [Yao14]. Consider a mixture of t distributions given covariate vectors $\mathbf{x}_i$,

$$h(y_i|\mathbf{x}_i,\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_i f(y_i|\mathbf{x}_i^T \boldsymbol{\beta}_k, \sigma_k^2, \tau_k),$$

where

$$f(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k,\sigma_k^2,\tau_k)=\frac{\Gamma(\frac{\tau_k+1}{2})|\boldsymbol{\sigma}_k|^{-1}}{(\pi_k\tau_k)^{\frac{1}{2}}\Gamma(\frac{\tau_k}{2})[1+g(y_i,\mathbf{x}_i^T\boldsymbol{\beta}_k,\sigma_k^2)/\tau_k]^{.5(\tau_k+1)}}$$

and $g(y_i,\mathbf{x}_i^T\boldsymbol{\beta}_k,\sigma_k^2)=(y_i-\mathbf{x}_i^T\boldsymbol{\beta}_k)^2/\sigma_k^2$. Where $\mathbf{x}_i^T\boldsymbol{\beta}_k$ represents the location function, σ_k^2 is the scale parameter and τ_k represents the degrees of freedom. This results in the complete log likelihood

$$\log(L_c(\boldsymbol{\theta};\mathbf{X},\mathbf{y},\mathbf{Z}))=\sum_{i=1}^{n}\sum_{k=1}^{K}z_{ik}\log(\pi_k f(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k,\sigma_k^2,\tau_k))$$

It is also well known that the t-distribution can be considered as a scale mixture of normal distributions. Thus, one can represent the error distribution ϵ

$$\epsilon|u\sim N(0,\tfrac{\sigma^2}{u})$$
$$u\sim gamma(\tfrac{\tau}{2},\tfrac{\tau}{2}).$$

So, one can then simplify and show that marginally ϵ has a t-distribution with scale parameter σ^2 and degrees of freedom τ. By introducing another latent variable u, independent of z, and substituting into the likelihood. The complete log-likelihood for $(\mathbf{X},\mathbf{y},\mathbf{z},\mathbf{u})$ then can be written as

$$\log(L_c(\boldsymbol{\theta};\mathbf{X},\mathbf{y},\mathbf{Z},\mathbf{u}))=\sum_{i=1}^{n}\sum_{k=1}^{K}z_{ik}\log(\pi_k\phi(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k,\sigma_k^2/u_i)f(u_i;\frac{\tau_i}{2},\frac{\tau_i}{2}))$$

where $\mathbf{u}=(u_1,\ldots,u_n)$. Implementing this method now involves calculating two independent expectations with respect to each unobserved variable. So in the expectation step at the $(t+1)$ iteration

$$E(z_{ik}|\mathbf{X},\mathbf{y},\boldsymbol{\theta}^{(t)})=p_{ik}^{(t+1)}=\frac{\pi_k^{(t)}f(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k^{(t)},\boldsymbol{\sigma}_k^{2^{(t)}},\tau_k^{(t)})}{\sum_{k=1}^{K}\pi_k f(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k^{(t)},\boldsymbol{\sigma}_k^{2^{(t)}},\tau_k^{(t)})}$$

$$E(u_i|\mathbf{X},\mathbf{y},\boldsymbol{\theta}^{(t)},z_{ik}=1)=u_{ik}^{(t+1)}=\frac{\tau_k^{(t)}+1}{\tau_i^{(t)}+\delta(y_i,\mathbf{x}_i^T\boldsymbol{\beta}_k^{(t)},\boldsymbol{\sigma}_k^{2^{(t)}},\tau_k^{(t)})}$$

Then in the corresponding M-step, the analyst can optimize for closed form solutions to the parameter vector $(\pi_k,\boldsymbol{\beta}_k,\sigma_k^2,\tau_k)$. This method has shown good results in studies of cases when outliers are present and impactful, yet it struggles with the presence of high leverage outliers. This study controls for outliers and leverage points, hence robust approaches are not considered.

2.2.6.3 Random Effects Regression Mixtures

Xu and Hedeker ([XH01]) describe a random-effects mixture model to determine whether treatment responses are from distinct subgroups. Xu et. al use both an EM approach as well as the Fisher scoring algorithm. Xu and Hedeker encourage the use of both sequentially as opposed to one method

alone. Consider the following notation,

$$\mathbf{y}_i = \mathbf{E}_i\boldsymbol{\alpha} + \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{e}_i,$$

where $\mathbf{y}_i$ is the $n_i \times 1$ vector of responses for subject i, $\mathbf{E}_i$ is a known $n_i \times p$ matrix containing explanatory variables, $\boldsymbol{\alpha}$ is a $p \times 1$ vector of regression unknown paramters, $\mathbf{X}_i$is a known $n_i \times q$ design matrix, $\boldsymbol{\beta}_i$ is the $q \times 1$ independent vector of unknown individual effects distributed as $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and also independent of the $n_i \times 1$ error vector $\boldsymbol{\epsilon}_i$ where the errors are distributed independently as $N(\mathbf{0}, \sigma^2\boldsymbol{I}_i)$. Xu and Hedeker then introduce a mixture model on the random effects to account for potential random subject effects and/or subject-varying grouping.

$$h(\boldsymbol{\beta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{\beta})$$

where $f_k(\boldsymbol{\beta}) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and K is the prespecified number of components. Xu and Hedeker represent the problem as one multivariate space by exploiting the joint normality conditioned on the component membership. The maximum likelihood equations are broken down into marginal components, thus the authors focus on solving for maximum marginal likelihood estimators. To begin the EM process, starting values must be input and the authors simply suggest using the ordinary homogeneous model and equal component weights. The E-step and M-step are similar to methods described above and the details can be found in [XH01].

It has been shown that the Fisher scoring solution is often faster than the EM algorithm. The Fisher scoring algorithm provides an estimate of the information matrix, which is used to provide the large-sample variances and covariances of the maximum marginal likelihood estimators. The authors note a couple issues with the solution such as the challenges of model testing, and the breakdown of the assumption of standard regularity conditions. Furthermore, the extent of the research does not include mixtures in random-effects models for categorical outcomes.

2.2.6.4 Assignment Dependence (Covariate Dependent Mixture Proportions)

In the traditional mixture of regression models, the mixture proportions are assumed to be constant with respect to the changes of covariates included in the model. That is, the mixture proportions as well as the other parameters are estimated using the EM algorithm. Jacobs et. al proposed a supervised learning approach to estimating the mixture proportions while accounting for changes in the covariates. [Jac91] In the machine learning literature, a mixture approach is often referred to as a mixture of experts model. The new model of interest is the following:

$$f(y_i|\mathbf{x}_i) = \sum_{k=1}^{K} \hat{\pi}_k(\mathbf{x}_i) h_{i|k}(y_i|\mathbf{x}_i^T\boldsymbol{\beta}_k, \sigma_k^2). \tag{2.6}$$

where K is the number of components and $h(\cdot)$ is the individual component density function. Jacobs et. al. suggests a parametric approach to estimating the mixture proportions

$$\pi_{ik}(\mathbf{x}_i) = \frac{exp(\beta_k^T \mathbf{x}_i)}{\sum_{k=1}^{K} exp(\beta_k^T \mathbf{x}_i)}.$$

where β_k is a vector of coefficients usually estimated with an iterative re-weighted least squares routine. Other methods have been proposed for estimating the mixture proportions, such as kernel methods. This approach aims to replace the estimate of π_k in the regular mixture of regressions with a local estimate.

Each technique mentioned above aims to improve the parameter estimation of the finite mixture of regression models, and have shown promise in certain situations. Each technique alone does not address the variable estimation challenges at the focus of this research.

2.2.6.5 Variable Selection

A recent extension of the FMR models and one that is relevant to the work of this manuscript was proposed by Khalili and Chen involving implementing penalty functions for variable selection within components. [KC07] Khalili and Chen attempt to solve the issue of irrelevant portions of the matrix of covariates. That is, the identification of sub-models that may exclude some variables in the covariate space. Intuitively, this may be done using some information criterion such as AIC or BIC. Although not impossible, such an approach would be computationally expensive and inefficient due to the nature of the mixture model problem. Hence, solving the parameter estimation and variable selection in one step is most attractive. Khalili and Chen propose a penalty function that is appended to the FMR likelihood function as follows

$$\tilde{\ell}_n(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta}) - \mathbf{P}_n(\boldsymbol{\theta})$$

with the penalty function

$$\mathbf{P}_n(\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \sum_{j=1}^{J} \left\{ P_{nk}(\boldsymbol{\beta}_{jk}) \right\}.$$

where the $P_{nk}(\boldsymbol{\beta}_{jk})$ values are nonnegative and nondecreasing functions in $|\boldsymbol{\beta}_{jk}|$. By subtracting the penalty function, when maximizing $\tilde{\ell}_n(\boldsymbol{\theta})$, the optimization will search for estimates of $\boldsymbol{\beta}$ that are as small as possible or zero, hence identifying a sub-model. In standard regression penalty methods, the penalty functions are scaled by the sample size. In the FMR setting, the penalty functions are proportional to the mixture proportions π_k, thus incorporating a function of the sample size.

Khalili and Chen work with three specific penalty functions: the L_1-norm (LASSO) used by Tibshirani, as well as the HARD and SCAD penalty functions discussed by Fan and Li [FL01]. The

authors also prove consistency of the penalized likelihood estimators for variable selection under certain conditions. Numerically, the authors describe a traditional EM algorithm approach, as described above, with a revised maximization step. The penalty function may pose issues with differentiability of the β's. More specifically, when $\beta = \mathbf{0}$, the penalty function $P_{nk}(\beta)$ may not be differentiable. Thus to avoid this issue, the authors take an approach similar to Fan and Li and replace $P_{nk}(\beta)$ with a local quadratic approximation in the neighborhood of the given value β_s,

$$P_{nk}(\beta) \approx P_{nk}(\beta_s) + \frac{P'_{nk}(\beta_s)}{2\beta_s}(\beta^2 - \beta_s^2)$$

Consider after the t^{th} iteration the parameter vector $\mathbf{\Omega}^{(t)}$, the new penalty function

$$\tilde{\mathbf{P}}_n(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) = \sum_{k=1}^{K} \pi_k \sum_{j=1}^{J} \left\{ P_{nk}(\beta_{jk}^{(t)}) + \frac{P'_{nk}(\beta_{jk}^{(t)})}{2\beta_{jk}^{(t)}}(\beta_{jk}^2 - \beta_{jk}^{(t)^2}) \right\}.$$

The resulting conditional expectation of the penalized complete likelihood function is the following

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) = E[\ell_c(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}, \mathbf{X}, \mathbf{y}, \mathbf{Z})] - \tilde{\mathbf{P}}_n(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$$

The E-step results in calculating posterior probabilities for the latent variable z as previously described. In the M-step, the mixture proportions are calculated the same as above, yet updating the regression coefficients and dispersion parameter is not a closed form solution and requires a numerical solution. Khalili and Chen perform a simulation study comparing a BIC method to the three proposed penalty functions. The authors report that the performance of the new method is dependent on the penalty function, size of the regression coefficients and the mixture structure.

This method is a significant step in the complex variable selection problem for mixture of regression models. The proposed penalty functions allow for more parsimonious models and improves efficiency over traditional stepwise methods. Yet, this method does not account for possile correlations or similarities among separate subgroups. In cases where the effects of covariates are the same in some groups and others are unique, this proposed model cannot leverage the relationship. Hence, the hypothesis of this research: In cases where a potential subset of proposed covariates have similar effects, a model that is flexible enough to share information across different components can achieve more accurate estimations than traditional FMR models.

CHAPTER

3

PENALIZED LATENT VARIABLE ESTIMATOR FOR FINITE MIXTURE OF REGRESSION MODELS

3.0.1 The Procedure

When data structures are assumed from multiple sub-populations, selecting and estimating impactful covariates is a difficult challenge. More specifically in mixture of regression models, there are situations in which the impact of a given covariate is exactly the same in different sub-populations, and different in others. The penalized latent variable estimation method attempts to identify and equate identical corresponding coefficient parameters across different mixture components. When the impact of covariates is identical across a pair of components, this implies the difference of coefficients, from the associated pair(s) of mixture components, would be approximately zero. This scenario is the primay focus of this manuscript. This research proposes a unique penalization approach to identifying and collapsing regression parameters that may be "shared" across mixture components for a given model.

To establish notation, consider the subject pair $(y_i, \mathbf{x}_i)$ where y_i is the response variable of interest and $\mathbf{x}_i$ is a p-column vector of covariates, with n subjects, p covariates. for each of the $K < \infty$ components.

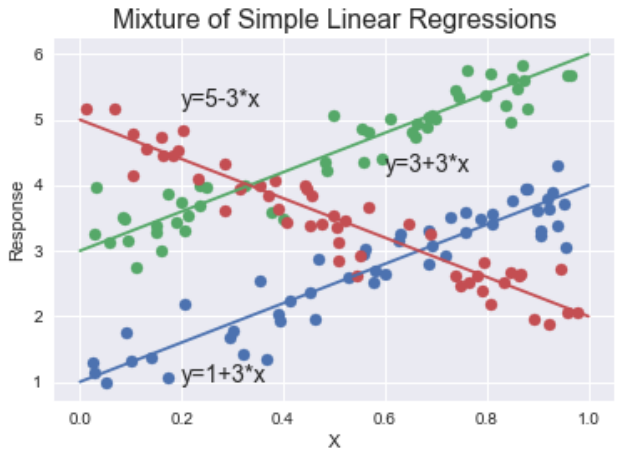

Figure 3.1 Here is an example of multiple sub-populations that each have linear patterns. The true linear equation is superimposed onto the corresponding sub-population data points. In this example, the slope of two linear equations are equivalent. The corresponding coefficients of the sub-component parameter vector would reflect this equality. This plot give a visual example of the described scenario of interest.

Define the FMR model of order K, and the conditional density of $(Y|\boldsymbol{X})$ as the following:

$$f_i(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_{i|k}(y_i|\mathbf{x}_i,\boldsymbol{\theta}_k),$$

where $\boldsymbol{\theta}^T = [\boldsymbol{\pi}^T, \beta^T, \boldsymbol{\sigma}^T]$, $\boldsymbol{\pi} = [\pi_1,\ldots,\pi_K]$,$\boldsymbol{\sigma} = [\sigma_1,\ldots,\sigma_K]$, $\boldsymbol{\beta} = [\beta_1,\ldots,\beta_K]$ with $\beta_k = [\beta_{0k},\ldots,\beta_{pk}]$. Substituting the normal distribution would produce the following k^{th} component density:

$$f_{i|k}(y_i|\mathbf{x}_i,\pi_k,\boldsymbol{\beta}_k,\sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} exp(-\frac{1}{2\sigma_k^2}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k)^2).$$

Given a sample of observations $(y_i,\mathbf{x}_i)$, the conditional log-likelihood function of $\boldsymbol{\theta}$ is given by:

$$\ell_n(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{X}) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} exp(-\frac{1}{2\sigma_k^2}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k)^2)\right), \tag{3.1}$$

with n observations and K components.

The resulting maximum likelihood estimates of $\hat{\boldsymbol{\theta}}$ are used to obtain the posterior probability for the i^{th} observation and k^{th} sub-population as the following:

$$\hat{p}_{ik} = \frac{\hat{\pi}_k f_{ik}(y_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}_k,\hat{\sigma}_k^2)}{\sum_{k=1}^{K} \hat{\pi}_k f_{ik}(y_i|\mathbf{x}_i,\hat{\boldsymbol{\beta}}_k,\hat{\sigma}_k^2)}, \tag{3.2}$$

This implies one could assign observation i to class k using the following rule: $p_{ik} > p_{ij} \;\forall\; k \neq j$.

In context of the proposed problem, this study assesses the differences of a given covariate across multiple mixture components. Consider the j^{th} covariate $\boldsymbol{x}_j$ and its corresponding coefficient parameter for the k^{th} mixture component, β_{jk}. If the effect of a corresponding covariate in a different mixture component is exactly the same, then the estimated difference should be close to zero, but will not be exactly zero. That is, $\hat{\beta}_{jk} - \hat{\beta}_{jl} \approx 0$ for $k \neq l$. This manuscript explores the impact of equating a given covariate effect across different mixture components. By equating the regression coefficients, this creates a submodel with one less parameter to estimate. To test the multiple possible sub-models with equal effects, for each unique covariate is a model selection problem. As literature has shown, studying these sub-models in the traditional iterative approach, using information criteria such as AIC and BIC can be computationally cumbersome and should be avoided. Thus, this manuscript proposes a penalization approach.

First, to ensure identifiability, consider the following definition: the "last" component $K \in \{1,\ldots,K\}$ as the baseline mixture component used as a reference for difference parameters. Given the baseline component, a unique representation can be established in the notation, and ambiguous permutations are avoided. Moreover, $\pi_k > 0 \; \forall k = 1,\ldots,K$ for all proposed order-K FMR models. This prevents vanishing components during estimation and asymptotic behavior.

Let $\boldsymbol{\beta}$ denote the vector of all mixture regression coefficients β_{jk} ($\forall\, j = 0,\ldots,p$ and $k = 1,\ldots,K$). That is $\boldsymbol{\beta}$ has length $(K(p+1))$, with $\boldsymbol{\beta} \subset \boldsymbol{\theta}$. Let L denote a linear transformation matrix such that $L\boldsymbol{\beta} = \boldsymbol{\beta}^*$ and $L\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}^*$. Choose L as follows:

$$L = [\; I_{K(p+1)} \;\; D^T \;]^T \tag{3.3}$$

where the $((K-1)(p+1)) \times (K(p+1))$ matrix D consists of $\{1,-1\}$ positioned to create pairwise differences with the baseline component q, for each component and each covariate. The resulting $\boldsymbol{\beta}^* = [\boldsymbol{\beta}^T, \boldsymbol{\beta}_z^T]^T$ where $\boldsymbol{\beta}_z^T = [\beta_{0z_1},\ldots,\beta_{pz_1},\beta_{0z_2},\ldots,\beta_{pz_2},\ldots,\beta_{pz_{K-1}}]^T$ is a $(K-1)(p+1)$ vector of pairwise difference parameters. And each $\beta_{jz_u} = (\beta_{ju} - \beta_{jK})$, for $u \neq K$, represents the difference effect relative the base component K. The complete parameter vector $\boldsymbol{\beta}^*$ is overparameterized and has length $(K(p+1)) + (K-1)(p+1)$.

Now, define the penalty function

$$\boldsymbol{P}_L(\boldsymbol{\beta}^*) = \quad \{\textstyle\sum_{k=1}^{(K-1)} \sum_{j=0}^{p} |\beta_{jz_k}| + \sum_{j=0}^{p} \sum_{1 \leq r < k \leq K-1} |\beta_{jz_k} - \beta_{jz_r}|\} \tag{3.4}$$

where $\beta_{jz_r} = (\beta_{jr} - \beta_{jK})$ and $\beta_{jz_k} = (\beta_{jk} - \beta_{jK})$ which implies $(\beta_{jz_k} - \beta_{jz_r}) = (\beta_{jk} - \beta_{jr})$, that is the difference from the k^{th} and r^{th} group, respectively. If these effects are the same, then their estimated difference should be close to zero. Furthermore, the penalized differences aims to set the closely estimated differences to exactly zero. Notice the second term in the penalty function accounts for all pairwise differences across groups not involving the specified base group K. In

order to combine the model selection and parameter estimation in one step, one can impose the penalty function on the log-likelihood function (3.1) above. Thus define the solution to constrained optimization problem can be expressed as:

$$\tilde{\beta}_L^* = \quad \arg\min_{\beta^*} \{\ell_n(\beta^*|\mathbf{Y},\mathbf{X}) - \lambda_L \boldsymbol{P}_L(\beta^*)\} \tag{3.5}$$

The first penalization function $\boldsymbol{P}_L(\beta^*)$ implemented uses the L_1-norm penalty similar to the LASSO technique by Tibshirani ([Tib96]). The L_1-norm penalty has been studied extensively in literature and has shown promising ability to shrink estimates to zero, in regression settings. Yet, the traditional LASSO technique often struggles to handle naturally larger coefficients and does not have the strongest asymptotic properties. In response, we also consider an adaptive penalty, similar to the adaptive LASSO. The adaptive penalty allows for a scalable amount of shrinkage for different coefficients (or differences of coefficients). The traditional LASSO uses the same amount of shrinkage for each coefficient and differences, which can produce inconsistent results. The adaptive LASSO uses a weighted penalty of the form $\sum_{j=1}^{p} \omega_j |\beta_j|$ where $\omega_j = 1/|\hat{\beta}_j|^{\nu}$, and $\hat{\beta}_j$ is the least squares estimate with $\nu > 0$. For this problem, we only consider the $\nu = 1$ case. For actual computations, a single least squares fit is done initially and the weight for each difference parameter is adjusted accordingly. The solution to adaptive LASSO constrained optimization problem can be expressed as:

$$\tilde{\beta}_A^* = \quad \arg\min_{\beta^*} \{\ell_n(\beta^*|\mathbf{Y},\mathbf{X}) - \lambda_A \boldsymbol{P}_A(\beta^*)\}\} \tag{3.6}$$

with

$$\boldsymbol{P}_A(\beta^*)\} = \{ \sum_{k=1}^{(K-1)} \sum_{j=0}^{p} \omega_j^{(k)} |\beta_{jz_k}| + \sum_{j=0}^{p} \sum_{1 \le r < k \le K-1} \omega_j^{(kr)} |\beta_{jz_k} - \beta_{jz_r}| \}$$

where $\omega_j^{(rk)} = |\hat{\beta}_{jz_k} - \hat{\beta}_{jz_r}|^{-1}$ and $\omega_j^{(k)} = |\hat{\beta}_{jz_k}|^{-1}$.

The proposed approach will be called the penalized latent variable estimator for finite mixture of regression models (PLV-FMR). In general, shrinking the proposed difference parameters β_{jz_k} does not always imply collapsing components. Yet, in a special case that all of the covariate effects for a particular component are estimated to have difference parameters of zero, for the same corresponding pairwise component, this would present a case for collapsing a mixture component and combining the two corresponding components into one. More specifically, given a pair of mixture components $k_1,\ k_2 \in \{1,\ldots,K\}\ \ k_1 \neq k_2$, if $\tilde{\beta}_{jz_{k_1}} = \tilde{\beta}_{jz_{k_2}}\ \forall\, j \in \{0,\ldots,p\}$, then the mixture components k_1 and k_2 have been estimated to be equal, and should be combined.

3.0.1.1 Penalized Variance Components

In the spirit of identifying similar characteristics across groups, this section explores the differences of variance components, $\tau = \sigma^2$. First, consider the expanded parameter vector $\boldsymbol{\theta}^* = [\beta^{*T} \boldsymbol{\tau}^T]^T$

where $\boldsymbol{\tau} = [\tau_1 \ \ldots \ \tau_K]^T$. An intuitive penalization approach would suggest an absolute value penalty of differences of variance components. Yet, after further investigation, there are advantages to considering a penalty on the squared differences of variance components. For instance, in the following section we will show that the quadratic representation allows for simple reformulation of the objective function for numerical estimation. Moreover, despite the inability to shrink to 0, the quadratic nature of the penalty easily conforms to the quadratic approximation approach proposed in the study. Furthermore, the quadratic representation does not impact the asymtotic properties of latent variable coefficient estimator. Thus one can include the following penalization term:

$$\lambda_\tau(\sum_{k=2}^{K}\sum_{1\leq j<k}\omega_\tau^{(kj)}(\tau_k-\tau_j)^2)$$

where $\tau_k = \sigma_k^2$, $\omega_\tau^{(kj)} = 1/|\hat{\tau_k} - \hat{\tau_j}|$ and new penalty parameter λ_τ. Given a mixture of normal densities, the same variance parameter constraints are $\tau > 0$, and allow for direct two-way combinations to be considered in the penalty.

To combine the two penalties into a single constrained optimization problem, one should first notice, the two penalty functions are independent. Thus, the dual penalty parameters can be combined or remain separated. Consider the following combined optimization function:

$$\tilde{\boldsymbol{\beta}}_D^* = \quad \arg\min_{\beta^*}\{\ell_n(\boldsymbol{\beta}^*|\mathbf{Y},\mathbf{X}) - \lambda_A \boldsymbol{P}_A(\boldsymbol{\beta}^*) - \lambda_\tau \boldsymbol{P}_\tau(\boldsymbol{\tau})\} \tag{3.7}$$

In the above equation, only the adaptive penalty function is considered. The next section describes the computational approach to the proposed methodologies.

3.0.2 Computation and Tuning

In this section I will address the primary steps and/or adjustments to compute the PLV-FMR optimization problem. Notice the presented log-likelihood function (3.1) is a relatively complex objective function for optimization, that is also subject to the proposed penalty constraints. One alternative approach is to leverage the functions asymptotic properties and transform the objective function (3.1) into a more feasible expression that allows for direct computation, i.e. a quadratic representation. The least squares approximation method proposed by Wang and Leng ([WL07]) is used to reformulate the optimization problem which then sets up the usage of the quadratic programming problem. The next subsection discusses the least squares approximation and setup for the quadratic programming problem for estimation of the PLV-FMR equation.

3.0.2.1 Least Squares Approximation

When applying a LASSO type penalty to regression model cases, Wang and Leng showed that many penalization problems can be unified into one simplified theoretical framework. More specifically, Wang and Leng proposed a method of least squares approximation (LSA) for LASSO regression problems. That is, the LSA transforms a complex LASSO type objective function into an asymptotically equivalent least squares, quadratic problem. Hence, the numerical approximation of the original problem can be simplified to an extended standard least squares approach.

The LSA method exploits the standard Taylor series expansion at the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ obtained by minimizing $\ell_n(\boldsymbol{\beta})$ defined in (3.1). A Taylor series expansion at $\hat{\boldsymbol{\beta}}$, the mle, gives

$$n^{-1}\mathscr{L}_n(\boldsymbol{\beta}) \approx n^{-1}\mathscr{L}_n(\hat{\boldsymbol{\beta}}) + n^{-1}\mathscr{L}'_n(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\{\frac{1}{n}\mathscr{L}''_n(\hat{\boldsymbol{\beta}})\}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}).$$

where $\mathscr{L}$ represents the objective function of interest and one assumes the existence of the continuous second-order derivative, $\mathscr{L}''_n$, with respect to $\boldsymbol{\beta}$. Note, since $\hat{\boldsymbol{\beta}}$ is the minimizer of $\mathscr{L}_n(\boldsymbol{\beta})$, that is, $\mathscr{L}'_n(\hat{\boldsymbol{\beta}}) = 0$ then this simplifies to the following quadratic expression:

$$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\{\frac{1}{n}\mathscr{L}''_n(\hat{\boldsymbol{\beta}})\}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}).$$

The resulting least squares function is very familiar. The asymptotic behavior of the least squares estimator is well developed in literature. The quantity $n^{-1}\mathscr{L}''_n(\hat{\boldsymbol{\beta}})$ is commonly used to approximate the asymptotic covariance of $\hat{\boldsymbol{\beta}}$. So define $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}} \approx E(n^{-1}\mathscr{L}''_n(\hat{\boldsymbol{\beta}}))$ as the covariance of $\hat{\boldsymbol{\beta}}$, and the LSA function simplifies to:

$$n^{-1}\ell_n(\boldsymbol{\beta}|\mathbf{X},\mathbf{Y}) \approx (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\hat{\Sigma}^{-1}_{\hat{\boldsymbol{\beta}}}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}).$$

This covariance estimator will be discussed in the next subsection. The LSA method assumes all potentially penalized parameters are included in the objective function.

For the PLV-FMR method, the LSA helps simplify the FMR log-likelihood objective function, defined in (3.1). Moreover, since the difference parameters are a linear transformation of the original vector of mixture coefficients $\boldsymbol{\beta}$, covariance properties do not need to be directly derived for the difference parameters. Instead, the corresponding linear transformations can be applied post analysis. Furthermore, to simplify the estimation approach, the variance component parameters are penalized independent of the regression coeffients. This implies the complete parameter vector can be expressed modularly. That is, the separate penalty functions can be optimized independently also. First, we consider the computation of the regression coefficients only.

First consider the complete vector of regression coefficient parameters $\boldsymbol{\beta}^* \in \mathbb{R}^{(2K-1)(p+1)}$.

$$\boldsymbol{\beta}^* = [\beta_{01},\beta_{11},\ldots,\beta_{p1},\ldots,\beta_{0K},\beta_{1K},\ldots,\beta_{pK},\beta_{0z_1},\ldots,\beta_{pz_1},\ldots,\beta_{pz_{K-1}}]$$

Let U be a linear transformation matrix such that $U\boldsymbol{\beta}^* = \boldsymbol{\beta}$. We choose

$$U = \begin{bmatrix} \boldsymbol{I}_{K(p+1)} & \mathbf{0}_{(K(p+1))\times(K-1)(p+1)} \end{bmatrix}$$

where $\boldsymbol{I}_{K(p+1)}$ is an identity matrix with $K(p+1)$ elements on its diagonal, and $\mathbf{0}_{(K(p+1))\times(K-1)(p+1)}$ is a $(K(p+1))\times(K-1)(p+1)$-matrix of zeros.

Using least squares approximation, the new formulated objective function can be expressed as:

$$n^{-1}\ell_n(\boldsymbol{\beta}^*|\mathbf{X},\mathbf{Y}) \approx (U(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^*))^T \hat{\Sigma}^{-1}_{\hat{\boldsymbol{\beta}}} (U(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^*)) \tag{3.8}$$

where $\hat{\Sigma}^{-1}_{\hat{\boldsymbol{\beta}}}$ is the estimated covariance matrix for the vector of mle's $\hat{\boldsymbol{\beta}}$.

Given the asymptotically quadratic formulation (**??**), it is feasible to represent the optimization problem as a quadratic programming problem. First, for each $j = 1,\ldots,p$ and $k = 1,\ldots,(K-1)$, set $\beta_{jz_k} = \beta^+_{jz_k} - \beta^-_{jz_k}$ where $\beta^+_{jz_k} \geq 0$ or $\beta^-_{jz_k} \geq 0$ and at least one is strictly zero. Then $|\beta_{jz_k}| = \beta^+_{jz_k} + \beta^-_{jz_k}$. Now the full $a+2b$ dimensional parameter vector can be expressed $\boldsymbol{\eta} = [\boldsymbol{\beta}^T \ \boldsymbol{\beta}^{+T}_z \ \boldsymbol{\beta}^{-T}_z]^T$. To account for the expanded parameter vector, denote

$$U_{\boldsymbol{\eta}} = \begin{bmatrix} \boldsymbol{I}_{K(p+1)} & \mathbf{0}_{(K(J+1))\times 2(K-1)(p+1)} \end{bmatrix}$$

The resulting optimization problem can be written as

$$\begin{gathered} \tilde{\boldsymbol{\eta}} = \arg\min (U_{\boldsymbol{\eta}}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}))^T \hat{\Sigma}^{-1}_{\hat{\boldsymbol{\beta}}} (U_{\boldsymbol{\eta}}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})) \\ s.t.\ M\boldsymbol{\eta} = \mathbf{0}, \\ \textstyle\sum_j \sum_k \omega^{(k)}_j (\beta^+_{jz_k} + \beta^-_{jz_k}) + \sum_j \sum_{1\leq r<k\leq K-1} \omega^{(kr)}_j [(\beta^+_{jz_k} - \beta^+_{jz_r}) - (\beta^-_{jz_k} - \beta^-_{jz_r})] \leq t \\ \text{with } \beta^+_{jz_k}, \beta^-_{jz_k} \geq 0 \ \ \forall k = 1,\ldots,(K-1) \text{ and } j = 0,\ldots,p \end{gathered} \tag{3.9}$$

where

$$M = \begin{bmatrix} D_K & -\boldsymbol{I}_{(K-1)(p+1)} & \boldsymbol{I}_{(K-1)(p+1)} \\ \mathbf{1}_{K(p+1)} & \mathbf{0}_{(K-1)(p+1)} & \mathbf{0}_{(K-1)(p+1)} \end{bmatrix}.$$

Notice the penalization function $p(\boldsymbol{\beta}^*)$ is represented as a set of linear constraints, as opposed to the Lagrangian formulation. This is consistent with the quadratic programming notation. Note, the maximum likelihood estimates are initially computed once, then used for adaptive weights and as a starting values. The final quadratic programming problem (3.9) has $a+2b$ parameters and $a+3b$ linear constraints.

For the double penalty method including the variance components, the above computational framework can be directly extended. Recall $\boldsymbol{\theta}^* = [\boldsymbol{\beta}^{*T}\boldsymbol{\tau}^T]^T$. Notice the quadratic nature of the variance penalty function, this formulation can be combined with the quadratic function used in (3.9). In fact, this function will be reformulated to resemble a ridge regression penalty and combined

in the quadratic programming problem.

Unlike the latent difference coefficient parameters, new difference parameters do not need to be created for the variance components. Since the variance components and coefficients are penalized independently, one can rewrite the quadratic objective function to include variances as such,

$$n^{-1}\ell_n(\boldsymbol{\theta}^*|\mathbf{X},\mathbf{Y}) \approx [U(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^*)]^T \hat{\Sigma}^{-1}_{\hat{\boldsymbol{\beta}}} U(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^*) + (\boldsymbol{\tau} - \hat{\boldsymbol{\tau}})^T \hat{\Sigma}^{-1}_{\hat{\boldsymbol{\tau}}} (\boldsymbol{\tau} - \hat{\boldsymbol{\tau}})$$

where $\hat{\boldsymbol{\tau}}$ and $\hat{\Sigma}^{-1}_{\hat{\boldsymbol{\tau}}}$ contain the corresponding maximum likelihood estimates and estimated covariances for the variance components, respectively. Now consider the proposed variance component penalty function $\boldsymbol{P}_\tau(\boldsymbol{\tau})$,

$$\lambda_\tau(\sum_{k=2}^{K} \sum_{1 \le j < k} \omega_\tau^{(kj)} (\tau_k - \tau_j)^2) = \lambda_\tau[\sum_{i=1}^{K} \tau_i^2 (e_{ik}) - 2 \sum_{j \neq k} \omega_\tau^{(ij)} \tau_j \tau_k]$$

where $e_{ik} = \sum_{i \neq k} \omega_\tau^{(ik)}$

$$\boldsymbol{p}_\tau(\boldsymbol{\tau}) = \begin{bmatrix} \tau_1 & \tau_2 & \dots & \tau_K \end{bmatrix} \times \begin{bmatrix} e_{1k} & -\omega_\tau^{(12)} & \dots & -\omega_\tau^{(1K)} \\ -\omega_\tau^{(12)} & e_{2k} & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -\omega_\tau^{(1K)} & \dots & \dots & e_{Kk}. \end{bmatrix} \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_K \end{bmatrix}$$

$$= \boldsymbol{\tau}^T \boldsymbol{\Gamma} \boldsymbol{\tau}.$$

Next consider the terms that contain $\boldsymbol{\tau}$ in the objective function and the proposed variance component penalty function:

$$(\boldsymbol{\tau} - \hat{\boldsymbol{\tau}})^T \hat{\Sigma}^{-1}_{\hat{\boldsymbol{\tau}}} (\boldsymbol{\tau} - \hat{\boldsymbol{\tau}}) + \lambda_\tau \boldsymbol{\tau}^T \Gamma \boldsymbol{\tau}$$

$$= \boldsymbol{\tau}^T (\hat{\Sigma}^{-1}_{\hat{\boldsymbol{\tau}}} + \lambda_\tau \Gamma) \boldsymbol{\tau} - 2 \hat{\boldsymbol{\tau}}^T \hat{\Sigma}^{-1}_{\hat{\boldsymbol{\tau}}} \boldsymbol{\tau} + C_1$$

$$= (\boldsymbol{\tau} - \hat{\boldsymbol{\tau}}^*)^T (\hat{\Sigma}^{-1}_{\hat{\boldsymbol{\tau}}} + \lambda_\tau \Gamma)(\boldsymbol{\tau} - \hat{\boldsymbol{\tau}}^*) + C_2.$$

For constants C_1 and C_2 not dependent on τ. Where

$$\hat{\boldsymbol{\tau}}^* = (\hat{\Sigma}^{-1}_{\hat{\boldsymbol{\tau}}} + \lambda_\tau \Gamma)^{-1} \hat{\Sigma}^{-1}_{\hat{\boldsymbol{\tau}}} \hat{\boldsymbol{\tau}}$$

The resulting double-penalty optimization problem can be expressed as:

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\beta},\boldsymbol{\tau}} \{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T \hat{\Sigma}_{\hat{\boldsymbol{\beta}}}^{-1} (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + (\boldsymbol{\tau}-\hat{\boldsymbol{\tau}}^*)^T (\hat{\Sigma}_{\hat{\boldsymbol{\tau}}}^{-1} + \lambda_\tau \Gamma)(\boldsymbol{\tau}-\hat{\boldsymbol{\tau}}^*) + \lambda_A P_A(\boldsymbol{\beta}^*)\}. \tag{3.10}$$

Note, the number of constraints grow linearly as the number of predictors p and the number of mixture components K increases. In practice the number of mixture components may not get too large, meanwhile the number of predictors may prove exceedingly large.

3.0.2.2 Covariance and Information Matrix

Estimating the covariance matrix defined in (3.10) is very important to the proposed analysis. The proposed method is associated with the incomplete data problem, reflected by leveraging the EM algorithm for parameter estimation.

Numerous techniques have been developed, aimed to compute the observed information matrix within the EM algorithm framework. This manuscript assumes the existence and computation of the complete data gradient and second derivative matrix. Consider the unobserved complete sample $(\mathbf{X},\mathbf{Z}) = [(\mathbf{x}_1,\mathbf{z}_1),\ldots,(\mathbf{x}_n,\mathbf{z}_n)]^T$, with $\boldsymbol{Z}$ is the matrix of latent variable values representing the observation's mixture component association. That is, $\mathbf{z}_i = [z_{i1},\ldots z_{iK}]$ where

$$z_{ik} = \begin{cases} 1 & i^{th} \text{ obs in component } k \\ 0 & \text{otherwise} \end{cases}$$

Let Y represent the incomplete observed sample. Next define $\ell_c(\mathbf{X},\mathbf{Z},\boldsymbol{\theta})$ and $\ell_n(\mathbf{Y},\boldsymbol{\theta})$ as the likelihood functions for the complete and incomplete data, respectively. These functions were previously defined (2.4) and (2.2), respectively. To compute the observed information in the EM algorithm, define $S_c(\mathbf{X},\mathbf{Z},\boldsymbol{\theta})$ and $S(\mathbf{y},\boldsymbol{\theta})$ as the gradient vectors of ℓ_c and ℓ_n respectively, and define $B_c(\mathbf{X},\mathbf{Z},\boldsymbol{\theta})$ and $B(\mathbf{y},\boldsymbol{\theta})$ as the negatives of the associated second derivative matrices. Then one can define the information matrix as the following:

$$I_Y(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[B_c(\mathbf{X},\mathbf{Z},\boldsymbol{\theta})|\mathbf{Y}] - E_{\boldsymbol{\theta}}[S_c(\mathbf{X},\mathbf{Z},\boldsymbol{\theta})S^T(\mathbf{X},\mathbf{Z},\boldsymbol{\theta})|\mathbf{Y}] + S(\mathbf{Y},\boldsymbol{\theta})S(\mathbf{Y},\boldsymbol{\theta})^T.$$

Notice the gradient vector at the converged mle, $S(\mathbf{Y},\hat{\boldsymbol{\theta}}) = \mathbf{0}$, can be dropped from the equation. The resulting equation can be expressed in a simplified notation.

$$I_Y(\boldsymbol{\theta}) = I_{\mathbf{X}} - I_{\mathbf{X}|\mathbf{Y}}$$

this formulation is straight forward to compute. Define

$$S(\mathbf{Y},\boldsymbol{\theta}) = \left[\ \frac{\partial \ell_c}{\partial \beta_{01}}, \ \frac{\partial \ell_c}{\partial \beta_{jk}}, \ \cdots \frac{\partial \ell_c}{\partial \beta_{jz_k}}, \ \cdots \frac{\partial \ell_c}{\partial \beta_{Jz}}, \ \frac{\partial \ell_c}{\partial \tau} \ \right]$$

and the second derivative matrix

$$B(\mathbf{Y},\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2 \ell_c}{\partial^2 \beta_{01}} & \frac{\partial^2 \ell_c}{\partial \beta_{01}\partial \beta_{11}} & \cdots & \frac{\partial^2 \ell_c}{\partial \beta_{01}\partial \beta_{0z}} & \cdots & \frac{\partial^2 \ell_c}{\partial \beta_{01}\partial \tau} \\ \frac{\partial \ell_c}{\partial \beta_{11}\partial \beta_{01}} & \frac{\partial \ell_c}{\partial^2 \beta_{11}} & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \ddots & \cdots & \cdots & \vdots \\ \frac{\partial^2 \ell_c}{\partial \beta_{0z}\partial \beta_{01}} & \cdots & \cdots & \frac{\partial \ell_c}{\partial^2 \beta_{1z}} & \vdots & \\ \vdots & & & & \ddots & \\ \frac{\partial \ell_c}{\partial \tau \partial \beta_{01}} & \cdots & \cdots & \cdots & \cdots & \frac{\partial \ell_c}{\partial^2 \tau} \end{bmatrix}$$

Now the information matrix can be constructed by substituting the resulting mle values and plugging into the final equation. Lastly, by inverting the information matrix I_Y, this quantity can be used to estimate the desired covariance matrix. Displaying the gradient vector and second derivative matrix above was simply to show the direct nature of this method.

3.0.2.3 Degrees of Freedom and Tuning

The performance of the proposed PLVR model is dependent on the choice of tuning parameters. Thus, the choice of tuning parameters is essential both theoretically and practically. Similar to other penalized likelihood problems, the tuning parameter for this study is focused on optimizing a given information criteria, e.g. AIC or BIC, both which require an estimate for degrees of freedom. Thus, the optimal tuning parameter t aims to balance the trade-off between model complexity and model fit. For the traditional LASSO, the number of degrees of freedom is estimated by the number of nonzero coefficients. In this study, the estimate of degrees of freedom for the overall mixture model is directly proportional to the number of unique coefficients across components for each given covariate. That is,

$$\hat{df} \propto \sum_{j=0}^{p} r_j^* \tag{3.11}$$

where $r^* \leq K$ represents the number of estimated unique coefficients for covariate j. The proportional notation is used since mixture proportions and variance parameters can be accounted for, but equal out. Typically an unpenalized approach would result in the full model of estimated parameters, which is proportional to $df_{full} = (K(p+1)+K)$. This essentially represents the maximum number degrees of freedom. Hence, a direct calculation of (3.11) is df_{full} less the number of terms in the penalty function that would equate to zero. It has been shown that for fixed dimensional cases, the BIC can identify the true model consistently, whereas the AIC may fail to do so. ([Aka09]) For this study, both methods are considered and reported. As mentioned earlier, the PLVR model framework is aimed to identify the true model structure while maintaining prediction accuracy, thus the BIC method is theoretically advantageous.

3.0.3 Penalized Latent Variable for FMR Estimator Asymptotics

This section addresses the asymptotic properties of the PLV-FMR estimator. The vector of coefficients, β, is the focus of the derived asymptotic inference. Similar to the original adaptive-LASSO used in a homogenous regression case or used for variable selection in finite mixture of regression models, the adaptive PLV-FMR estimator in (3.7) shares the oracle property. That is, the performance of the adaptive PLV-FMR estimator is asymptotically equivalent to knowing the true shared covariate effects across sub-populations beforehand, and then applying maximum likelihood estimation to the identified design. Thus as the sample size increases, the probability of selecting the correct model structure and shared covariate effects tends to one.

Let $\mathscr{A} = \{(j,r,u) : \beta_{jk} \neq \beta_{jm}\}$ represent the set of indices for the true nonzero differences. Moreover, let $\mathscr{A}_n$ denote the set of indices for the estimated nonzero differences. Let $\beta_z^{\mathscr{A}}$ denote the vector of difference parameters corresponding to the set $\mathscr{A}$, therefore $\beta_z^{\mathscr{A}} \in \beta^*$. Note, for notation purposes, a default baseline group is pre-specified, thus the vector $\beta_z^{\mathscr{A}}$ corresponds to the unique true model structure. As a result, define $\tilde{\beta}_z^{\mathscr{A}}$ denote the "oracle" estimator of $\beta_z^{\mathscr{A}}$. As an "oracle" estimator, given standard conditions, $\sqrt{n}(\tilde{\beta}_z^{\mathscr{A}} - \beta_z^{\mathscr{A}}) \rightarrow \mathscr{N}(\mathbf{0}, \Sigma_{\hat{\beta}_z^{\mathscr{A}}})$. It is important to note $\Sigma_{\hat{\beta}_z^{\mathscr{A}}}$ is a singular matrix due to the overparamterized difference parameter vector $\beta_z^{\mathscr{A}}$, in the mixture model setting. Consider the following theorem that states the adaptive PLV-FMR estimator achieves the oracle property.

THEOREM 1: Let the order-K FMR design be such that $\pi_k > 0$, $k = 1,\ldots,K < \infty$ and $\lambda \rightarrow \infty$ and $\lambda_A = o(n^{1/2})$ then, the adaptive PLV-FMR estimator $\tilde{\beta}^*$ and its corresponding estimate of differences $\tilde{\beta}_z^{\mathscr{A}}$ have the following properties:

$$
\begin{aligned}
&(i) \quad \mathscr{P}[\mathscr{A}_n = \mathscr{A}] \rightarrow 1 \\
&(ii) \quad \sqrt{n}(\tilde{\beta}_z^{\mathscr{A}} - \beta_z^{\mathscr{A}}) \rightarrow \mathscr{N}(\mathbf{0}, \Sigma_{\beta_z^{\mathscr{A}}})
\end{aligned}
$$

The theorem 1 part (i) corresponds to the consistency in variable selection. That is, the guarantee for selecting the correct structure and identifying the covariates that share impacts across sub-populations, with probability tending to 1. Part (ii) assures the asymptotic distribution of the estimator converges to the true vector difference parameters β_z corresponding to mixture structure $\mathscr{A}$. The asymptotic covariance matrix contains the parameters that remain under the true mixture structure after removing unnecessary coefficients and variance components, detailed by $\mathscr{A}$. Note the design condition requires non-vanishing mixture proportions and a pre-specified finite number of sub-populations. The proof for theorem 1 can be found in the appendix to the manuscript.

The next section summarizes the numerical study for the PLV regression procedure.

CHAPTER

4

NUMERICAL ANALYSIS

This simulation study is designed to examine the performance of the PLV-FMR estimator and its ability to identify a simplified submodel without sacrificing predictive accuracy. Ten scenarios have been constructed to display the proposed methods' ability to collapse variables via identifying true difference parameters while maintaining accurate estimates of the remaining parameters and minimizing the predictive error rate. The first 4 examples reflect homogenous variance components. As part of the assessment of performance, many different sets of conditions were observed. These 4 examples are intended to summarize the consistent findings over these simulations. The remaining examples are samples with heterogenous variance components. Furthermore, the examples will reflect the impact of an increased number of covariates, increased number of mixture components as well as mixture proportion imbalances. There are three penalization techniques used in this analysis. First the L_1 linear constraint, next the adaptive L_1 penalty; both on the difference coefficient parameters only in the homogenous variance examples. The third method is the adaptive double penalty, applied on scenarios with cross component unique variance components, by incorporating the penalty on both coefficients and the variance components.

For each scenario, the data was simulated from an order-K finite mixture of normal densities, as such:

$$y_i \sim \sum_{k=1}^{K} \pi_k N(\mathbf{x}_i^T \boldsymbol{\beta}_k, \sigma_k^2)$$

where π_k are pre-specified mixture proportions, and σ_k^2 represents the true heterogeneous variance

parameters. The covariates **x** are generated from a multivariate normal with mean **0**, covariance equal to the identity matrix I such that independent correlation structure, $\rho_{ij} = 0$, for $i \neq j$. The mixture proportions, sample sizes, and number of replications vary for a given scenario. All simulations and calculations are conducted using MATLAB codes. The starting values are computed using the Flexmix package in R, built by Leisch et. al [Lei04].

To compare the performance of the model selection techniques, a few measures are calculated. First, the AIC and BIC criterion are calculated and used to tune the penalty parameters for each method. The mean squared errors (MSE) are aggregates over generated samples and are used to compare the overall accuracy for the vector of coefficients. Standard errors for the individual coefficients and MSEs are computed using Bootstrap samples and compared across techniques. Furthermore, to assess the performance of the penalty techniques, the average number of collapsed variables (MCV) is recorded as well as the proportion of models selected that chose the entirely correct model (ACM). The ACM metric only considers samples that correspond to the correct mixture order being identified. That is, for each generated sample, the proposed method is assessed using a range of mixture model order values K. For each generated sample, it is recorded whether the estimator captures the true model order. For each instance the correct mixture model order is selected, the proportion of estimates with the true model structure is reported. This assures the comparison of correct parameter estimation is based on the same parameter sub-space. To assess the predictive ability of the estimator, an out-of-sample set of 10,000 test observations is generated to compare misclassification rates. Lastly, the study reports the ability to choose the correct order of the mixture distribution K.

For this simulation study the complete coefficient vector will consist of the parameters for all of the groups where the first (m+1) elements are the "base" group and each block of following (m+1) elements represent differences from the base group to the corresponding variables in other groups. For example consider a two group example with one covariate. The true coefficient vectors for the first and second group respectively are $\beta_1 = (5\ 10)$ and $\beta_2 = (10\ 10)$. This results in a complete coefficient vector of $\beta^* = [5\ 10\ 5\ 0] = [\beta_{01}\ \beta_{11}\ (\beta_{02} - \beta_{01})\ (\beta_{12} - \beta_{11})]$

The starting values used for each scenario are the original maximum likelihood estimates. These estimates also correspond to the unpenalized standard used to compare the performance of the proposed method. The examples included in this manuscript were chosen to best reflect the overall findings.

4.0.1 Single Variance Examples

2 Groups

Example 1: In this example, the data are generated using two groups, $K = 2$, with true mixture proportions of $\pi = (0.5, 0.5)$, each group containing five covariates plus an intercept, $p = 5$, and a homo-

geneous population variance, $\sigma^2 = 1$. The complete coefficient vector is $\beta^* = (6\,5\,4\,5\,1\,3\,,3\,3\,0\,0\,0\,0)$. Each dataset contained a sample size of 100 observations.

Table 4.1 (Row 1-2): Estimated mean squared errors of θ for the MLE, PLV-FMR L_1 penalty estimator and PLV-FMR adaptive L_1 penalty estimator given 2 Groups, 5 covariates. (Row 3): Proportion of samples with correctly estimated mixture order. (Row 4): OOS misclassification rate. (Row 5): Average number of collapsed variables. (Row 6): Proportion of correctly identified model structures. This table reflect samples of 100 observations.

	Unpenalized	L_1 Penalty		Adaptive Penalty	
MSE		AIC	BIC	AIC	BIC
β	0.7822(0.048)	0.6306(0.044)	0.6267(0.036)	0.4913(0.040)	0.4563(0.043)
σ^2	0.0705(0.009)	0.0400(0.009)	0.0255(0.004)	0.0385(0.008)	0.0360(0.005)
K	0.9540 (0.014)	0.9548 (0.031)	0.9551 (0.014)	0.9580 (0.032)	0.9640 (0.015)
Misclass (%)	15.00(0.005)	15.00(0.008)	15.00(0.009)	13.33(0.008)	13.33(0.009)
MCV	0.0	4.00(0.43)	4.00(0.00)	4.00(0.00)	4.00(0.00)
ACM (%)	0.0	52.0	82.0	80.0	96.0

This example displays a simple, balanced mixture model with 13 parameters to estimate. The columns in table 4.1 correspond to different methods, and sub-selections. For each of the model parameters, the unpenalized method significantly under-performed compared to the proposed penalty methods. Observe the relative MSE decreases nearly 40% for the regression coeffients and nearly 50% for the variance component. For this simple scenario, the BIC outperforms the AIC in each case and each parameter error metric. The true complete coefficient vector indicates there are 4 covariates with shared effects. The results from Table 4.1 indicate the unpenalized method was unable to identify the shared effects. Meanwhile both penalty methods consistently identified all 4 shared coefficient parameters($MCV = 4.0$). In fact, the adaptive penalty with BIC selection was able to identify the correct model structure with collapsed difference coefficients up to ($ACM = 96\%$) of the samples. The OOS misclassifcation metric implies the adaptive penalty method consistently out-performed the other methods. Lastly, each method was able to identify the correct mixture order ($K = 2$) in over 95% of the samples.

Example 2: In this example, the data are generated using two groups, $K = 2$, with true mixture proportions of $\pi = (0.5, 0.5)$, each group containing eleven covariates plus an intercept, $p = 11$, and a homogeneous population variance, $\sigma^2 = 1$. The complete coefficient vector is $\beta^* = (5\,5\,5\,5\,5\,5\,5\,5\,5\,5\,5\,,5\,0\,5\,0\,0\,5\,0\,0\,0\,0\,0\,0)$. Each dataset contained a sample size of 100 observations.

This example is a balanced mixture model with nearly double the amount of covariates within each mixture component compared to example 1. That is, there are 25 model parameters to estimate.

Table 4.2 (Row 1-2): Estimated mean squared errors of θ for the MLE, PLV-FMR L_1 penalty estimator and PLV-FMR adaptive L_1 penalty estimator given 2 Groups, 11 covariates. (Row 3): Proportion of samples with correctly estimated mixture order. (Row 4): OOS misclassification rate. (Row 5): Average number of collapsed variables. (Row 6): Proportion of correctly identified model structures. This table reflect samples of 100 observations

	Unpenalized	L_1 Penalty		Adaptive Penalty	
MSE		AIC	BIC	AIC	BIC
β	0.9085(0.052)	0.6507(0.041)	0.7306(0.055)	0.5357(0.032)	0.4708(0.031)
σ^2	0.0902(0.009)	0.0392(0.007)	0.0223(0.004)	0.0472(0.009)	0.0348(0.006)
K	0.9540 (0.013)	0.9480 (0.030)	0.9360 (0.016)	0.9490 (0.031)	0.9540 (0.014)
Misclass (%)	8.00(0.003)	8.00(0.002)	8.00(0.004)	8.00(0.004)	8.00(0.005)
MCV	0.0	8.00(0.28)	9.00(0.00)	9.00(0.00)	9.00(0.00)
ACM (%)	0.0	34.0	70.0	55.0	96.0

In this example, the two sub-components are pretty well separated. Similar, to the previous example, the unpenalized method failed to estimate the shared effects. The unpenalized method was consistently less accurate in the coefficent parameter estimation, compared to the proposed penalty methods. The relative MSE is significantly lower for the regression coeffients and the variance components when using the adaptive penalty method. This example also doubles the amount of shared effects, compared to example 1. Both penalty methods consistently identified at least 8 shared coefficient parameters($MCV = 8.0$). In fact, the adaptive penalty with BIC selection was able to identify the correct model structure with collapsed difference coefficients up to ($ACM = 96\%$) of the samples. The OOS misclassifcation metric was consistently under 10% for each method compared. Again, for this well separated sub-group problem, each technique was able to identify the correct mixture order ($K = 2$) in over 93% of the samples.

3 Groups

Example 3: In this example, the data are generated using three groups, $K = 3$, with true mixture proportions of $\pi = (0.33, 0.33, 0.34)$, each group containing four covariates and an intercept, $p = 4$, and a homogeneous population variance, $\sigma^2 = 1$. The complete coefficient vector is $\beta^* =$ (3 3 3 3 3 , 5 0 0 0 0 , 0 0 5 0 0). Each dataset contained a sample size of 100 observations.

By increasing the order of the mixture model, this increases the number of potential parameters in the penalty function for the model and allows for more shrinkage. Table 4.3 reflects the significant decrease in mean squared error for the coefficient and variance parameters. In this case, the error for the unpenalized method is more than double the amount found in the adaptive penalty term using the BIC criterion. Similar, to the previous examples, the unpenalized method failed to estimate

Table 4.3 (Row 1-2): Estimated mean squared errors of θ for the MLE, PLV-FMR L_1 penalty estimator and PLV-FMR adaptive L_1 penalty estimator given 3 Groups, 4 covariates. (Row 3): Proportion of samples with correctly estimated mixture order. (Row 4): OOS misclassification rate. (Row 5): Average number of collapsed variables. (Row 6): Proportion of correctly identified model structures. This table reflect samples of 100 observations.

	Unpenalized	L_1 Penalty		Adaptive Penalty	
MSE		AIC	BIC	AIC	BIC
β	0.7761(0.042)	0.5575(0.046)	0.5640(0.041)	0.4028(0.025)	0.3340(0.027)
σ^2	0.0498(0.009)	0.0249(0.004)	0.0237(0.005)	0.0259(0.004)	0.0175(0.005)
K	0.9140 (0.014)	0.7800 (0.041)	0.9020 (0.012)	0.7900 (0.041)	0.9190 (0.011)
Misclass (%)	15.00(0.005)	14.00(0.007)	15.00(0.006)	14.00(0.006)	14.00(0.007)
MCV	0.0	8.00(0.38)	8.00(0.00)	8.00(0.00)	8.00(0.00)
ACM (%)	0.0	52.0	77.0	78.0	100.0

the shared effects, meanwhile both penalty methods consistently identified at least 8 shared coefficient parameters($MCV = 8.0$). In this example the shared effects are different across the mixture component combinations. Notice for the 4^{th} and 5^{th} covariate in the coefficient vector, the true effect is the same for all 3 groups. This did not impact the proposed penalty methods, as non-base group differences are also accounted for in the penalty term. In fact, the adaptive penalty with BIC selection was able to identify the correct model structure with collapsed difference coefficients for all of the samples ($ACM = 100\%$). The OOS misclassifcation rate was slightly improved using the adaptive penalty compared to the unpenalized estimator. Due to the relatively small number of observations within a sub-sample, this example reflects a more challenging prediction problem. By increasing the number of sub-groups, the ability to identify the correct order became more challenging. Using the AIC criterion and the L_1 penalty, the method correctly estimated $K = 3$ in 78% of the samples. Also, the best performing method (BIC, adaptive penalty) correctly estimated 91% of the samples.

4 Groups

Example 4: In this example, the data are generated using four groups, $K = 4$, with equal true mixture proportions of $\pi = (0.25, 0.25, 0.25, 0.25)$, each group containing five covariates plus an intercept, $p = 4$, and a homogeneous population variance, $\sigma^2 = 1$. The complete coefficient vector is $\beta^* = (1\,1\,1\,1\,1\,,0\,0\,5\,0\,0\,,5\,5\,0\,0\,0\,,0\,0\,0\,5\,5)$. Each dataset contained a sample size of 100 observations.

This example displays the impacts of increasing the number of mixture components. First, as the mixture order increases, to create a well separated problem, one must also increase the number of covariates and magnitudes of the effects. This example displays a limited number of covariates

Table 4.4 (Row 1-2): Estimated mean squared errors of $\boldsymbol{\theta}$ for the MLE, PLV-FMR L_1 penalty estimator and PLV-FMR adaptive L_1 penalty estimator given 4 Groups, 4 covariates. (Row 3): Proportion of samples with correctly estimated mixture order. (Row 4): OOS misclassification rate. (Row 5): Average number of collapsed variables. (Row 6): Proportion of correctly identified model structures. This table reflect samples of 100 observations

	Unpenalized	L_1 Penalty		Adaptive Penalty	
MSE		AIC	BIC	AIC	BIC
$\boldsymbol{\beta}$	1.3896(0.101)	1.1496(0.114)	1.1810(0.116)	0.6060(0.069)	0.5077(0.042)
σ^2	0.0736(0.013)	0.0387(0.008)	0.0389(0.013)	0.0372(0.009)	0.0182(0.004)
K	0.9440 (0.014)	0.4800 (0.031)	0.9440 (0.014)	0.4800 (0.031)	0.9440 (0.014)
Misclass (%)	24.17(0.005)	24.17(0.004)	24.17(0.005)	22.92(0.004)	22.50(0.007)
MCV	0.0	7.00(0.42)	9.00(0.31)	9.00(0.00)	10.00(0.00)
ACM (%)	0.0	8.0	20.0	45.0	83.0

and relatively small magnitudes for the effects. There are 21 model parameters to estimate. First notice the misclassification rates for each technique is consistently above 20%. Yet, as there are more parameters in the model, there is more opportunity for overall improved parameter estimation. Table 4.4 reflects the significant decrease in mean squared error for the coefficient and variance parameters as previously shown in examples 1-3. In this case, the error for the unpenalized method is more than double the amount of compared to the adaptive penalty term using the BIC criterion for the coefficient vector. For the unpenalized variance component, the corresponding error decreases four-fold. In this example, there are 10 true shared effects across the 4 mixture components. Table 4.4 displays the gradual improvement of identifying shared effects across the different techniques. In this case, the adaptive penalty method significantly out-performed the L_1 penalty in estimation and ability to collapse difference parameters. Notice the best L_1 penalty models were able to identify the correct model structure 20% of the samples, while the adaptive penalty reached the true structure nearly 83% of the samples. Note, this example only uses samples of 100 observations, and increasing the number of subgroups makes the classification problem more challenging. This example reflected the challenge of identifying the correct mixture order, as the order increases. In this example, the AIC criterion often chose the over parameterized model, $\hat{K} \geq 5$. Meanwhile, the adaptive penalty technique was able to identify the correct mixture order ($K = 4$) in over 94% of the samples.

4.0.2 Unique Variance Examples

The next set of examples assume a heterogenous variance across mixture components. Thus, for the proposed penalty method, the penalized variances are included in the penalty function. The examples reflect the impacts of varying the number of model parameters, mixture proportions, and

sample sizes. Heterogeneous variances are more common and present more challenging scenarios to estimate.

2 Groups

Example 5 (Limited Covariates): In this example, the data is generated using two groups, $K = 2$, with equal true mixture proportions of $\pi = (0.5, 0.5)$, each group containing five covariates plus an intercept, $p = 5$ and a heterogeneous population variance, $\sigma_1^2 = \sigma_2^2 = 1$. The complete coefficient vector is $\beta^* = (3\,3\,3\,3\,3\,3\,,3\,3\,0\,0\,0\,0)$. Each dataset contained a sample size of 150 observations.

Table 4.5 (Row 1-2): Estimated mean squared errors of θ for the MLE, PLV-FMR L_1 penalty estimator and PLV-FMR adaptive L_1 penalty estimator given 2 Groups, 5 covariates (Heterogeneous). (Row 3): Proportion of samples with correctly estimated mixture order. (Row 4): OOS misclassification rate. (Row 5): Average number of collapsed variables. (Row 6): Proportion of correctly identified model structures. This table reflect samples of 150 observations.

MSE	Unpenalized	Adaptive Double Penalty AIC	Adaptive Double Penalty BIC
β	0.2267 (0.010)	0.1713 (0.008)	0.1579 (0.005)
σ^2	0.0894 (0.009)	0.0830 (0.012)	0.0672 (0.010)
K	0.9440 (0.014)	0.8800 (0.031)	0.9440 (0.014)
Misclass (%)	0.1357 (0.000)	0.1342 (0.000)	0.1331 (0.000)
MCV	0.0000 (0.000)	4.0000 (0.479)	4.0000 (0.000)
ACM (%)	0.0000 (0.000)	0.3400 (0.030)	0.5520 (0.031)

This example is a balanced mixture model with 5 total shared effects, including the shared variance component. This example displays a limited number of covariates and relatively small magnitudes for the effects. First notice the impact of the corresponding metrics when the variance components are included in the penalty function. Table 4.5 displays a consistent decrease in mean squared error for the coefficient and variance parameters as previously shown in the homogeneous examples above. Yet in this case, the decrease in error for the BIC criterion is not as dramatic for the coefficient vector and variance component vector. The double penalty method performed well for identifying shared effects across the mixture components. On average, the penalty method identified up to 4 shared effects for both AIC and BIC selection criterion. Yet this penalty model was only able to identify the correct full model structure < 10% of the samples. For the heterogeneous cases, the ACM β compares the proportion of samples with correct coefficient vector. The BIC criterion was able to identify the correct coefficient vector for over 55% of the samples. A positive note, the penalty method with the BIC criterion able to consistently identify the correct number of mixture components K

over 94% of the samples, nearly equivalent to the unpenalized method. While the AIC criterion, typically over estimated the number of components. Lastly, the out-of-sample misclassification rate showed very little evidence of improvement for the penalty method.

Example 6 (Imbalanced Mixture): In this example, the data is generated using two groups, $K = 2$, with unequal mixture proportions of $\pi = (0.25, 0.75)$, each group containing five covariates plus an intercept, $p = 5$ and a heterogeneous population variance, $\sigma_1^2 = \sigma_2^2 = 1$. The complete coefficient vector is $\beta^* =$ (3 3 3 3 3 3 ,3 3 0 0 0 0). Each dataset contained a sample size of 150 observations.

Table 4.6 (Row 1-2): Estimated mean squared errors of θ for the MLE, PLV-FMR L_1 penalty estimator and PLV-FMR adaptive L_1 penalty estimator given 2 Groups, 5 covariates (Heterogeneous). (Row 3): Proportion of samples with correctly estimated mixture order. (Row 4): OOS misclassification rate. (Row 5): Average number of collapsed variables. (Row 6): Proportion of correctly identified model structures. This table reflect samples of 150 observations.

MSE	Unpenalized	Double Penalty AIC	Double Penalty BIC
β	0.3835 (0.021)	0.2749 (0.013)	0.2212 (0.011)
σ^2	0.1814 (0.020)	0.1501 (0.019)	0.1476 (0.017)
K	0.9440 (0.014)	0.8400 (0.031)	0.9440 (0.014)
Misclass (%)	0.1693 (0.020)	0.1681 (0.020)	0.1645 (0.019)
MCV	0.0000 (0.000)	3.0000 (0.000)	4.0000 (0.000)
ACM (%)	0.0000 (0.000)	0.2560 (0.028)	0.5360 (0.032)

This example is designed to observe the impact of an unbalanced mixture model. The model parameter space is the same as example 5, thus there are 5 total shared effects, including the shared variance component. The penalty method performs similarly with respect to the MSE for the coefficient vector and variance component parameters. The table 4.6 shows there are improvements in the accuracy of parameter estimates. Similarly, there is limited improvement with the misclassification rates compared to the unpenalized case. Due to the mixture model imbalance, the misclassification increased compared to the balanced case. The proportion of correct models identified also decreased. The penalty method, using the BIC criterion was able to properly identify the correct model $< 6\%$ of the samples, compared to $\approx 10\%$ for the balanced case. The proportion of coefficient parameter vectors correctly identified decreased to $\approx 53\%$. Lastly, the unbalanced case displayed similar abilities to correctly identify the correct mixture order.

Example 7 (Increased Covariates): In this example, the data is generated using two groups, $K = 2$, with equal true mixture proportions of $\pi = (0.5, 0.5)$, each group containing ten covariates

plus an intercept, $p = 10$ and a heterogeneous population variance, $\sigma_1^2 = \sigma_2^2 = 1$. The complete coefficient vector is $\beta^* = (3\,3\,3\,3\,3\,3\,3\,3\,3\,3\,3\,,3\,3\,0\,0\,0\,0\,0\,0\,0\,0\,0)$. Each dataset contained a sample size of 150 observations.

Table 4.7 (Row 1-2): Estimated mean squared errors of θ for the MLE, PLV-FMR L_1 penalty estimator and PLV-FMR adaptive L_1 penalty estimator given 2 Groups, 14 covariates (Heterogeneous). (Row 3): Proportion of samples with correctly estimated mixture order. (Row 4): OOS misclassification rate. (Row 5): Average number of collapsed variables. (Row 6): Proportion of correctly identified model structures. This table reflect samples of 150 observations.

MSE	Unpenalized	Adaptive Double Penalty AIC	BIC
β	0.5010 (0.016)	0.3411 (0.013)	0.2743 (0.012)
σ^2	0.1663 (0.016)	0.1684 (0.014)	0.1721 (0.016)
K	0.9440 (0.014)	0.8800 (0.031)	0.9440 (0.014)
Misclass (%)	0.1430 (0.000)	0.1389 (0.000)	0.1368 (0.001)
MCV	0.000 (0.000)	7.000 (0.135)	9.000 (0.115)
ACM (%)	0.0000 (0.000)	0.1240 (0.021)	0.3480 (0.030)

This example was designed to measure the impact of significantly increasing the number of model parameters. The model parameter space contains the β^* used in the previous example, with added covariates. There are now 30 model parameters to estimate, with 10 shared effects. Table 4.7 shows the effectiveness of the penalty method for the coefficient parameter vector. There is a significant decrease in MSE for the coefficient parameter vector, but not for the variance component parameters. In fact, the penalty method is less accurate than the unpenalized method for variance components. The misclassification rate for the penalty method slightly improves. Moreover, the BIC criterion for the penalty method is able to identify the full correct model structure nearly 10% of the samples. And when only considering the coefficient parameters, the BIC method chooses the correct structure $\approx 35\%$ of the samples. For this sample, the unpenalized and BIC selector are both able to select select the correct mixture order $\approx 94\%$ of the samples generated. Meanwhile the AIC chooses a less parsimonous model nearly 12% of the samples.

3 Groups

Example 8 (Increased Shared Effects): In this example, the data is generated using three groups, $K = 3$, with true mixture proportions of $\pi = (0.3, 0.3, 0.4)$, each group containing eight covariates plus an intercept, $p = 8$ and a heterogeneous population variance, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$. The complete coefficient vector is $\beta^* = (5\,5\,5\,5\,0\,5\,0\,0\,0\,,5\,5\,0\,0\,0\,5\,0\,5\,5\,,0\,0\,0\,5\,0\,5\,0\,0\,0\,)$. Each dataset contained a

sample size of 250 observations.

Table 4.8 (Row 1-2): Estimated mean squared errors of θ for the MLE, PLV-FMR L_1 penalty estimator and PLV-FMR adaptive L_1 penalty estimator given 3 Groups, 8 covariates (Heterogeneous). (Row 3): Proportion of samples with correctly estimated mixture order. (Row 4): OOS misclassification rate. (Row 5): Average number of collapsed variables. (Row 6): Proportion of correctly identified model structures. This table reflect samples of 250 observations.

	Unpenalized	Adaptive Double Penalty	
MSE		AIC	BIC
β	1.1126 (0.034)	0.7739 (0.030)	0.6074 (0.026)
σ^2	0.5590 (0.045)	0.5040 (0.043)	0.5001 (0.032)
K	0.9320 (0.015)	0.9720 (0.010)	0.9320 (0.015)
Misclass (%)	0.1432 (0.001)	0.1379 (0.001)	0.1344 (0.000)
MCV	0.0000 (0.000)	9.0000 (0.224)	12.000 (0.045)
ACM (%)	0.0000 (0.000)	0.1100 (0.062)	0.1300 (0.012)

This example is a balanced mixture model with 14 total shared effects, including the 2 shared variance components. In this example notice the second and third group share a similar effect in the 5^{th} covariate. This can be observed from the β^* vector, represented by the same difference from the baseline group. Also notice, the base group has four covariate effects that are irrelevant. The corresponding parameters are not expected to be estimated as zero, since the penalty function is designed to collapse similar effects across mixture proportions. Table 4.8 displays a consistent decrease in mean squared error for the coefficient parameters. Again, the variance component parameter estimates are not significantly different when incorporating the adaptive penalty. The double penalty method performed as expected with identifying shared effects across the mixture components. On average, the penalty method identified up to 12 shared effects for the BIC selection criterion. Yet this penalty model was only able to identify the correct full model structure $< 3\%$ of the samples. The BIC criterion was able to identify the correct coefficient vector for nearly 13% of the samples. An interesting observation, the penalty method with the AIC criterion was able to consistently identify the correct number of mixture components K over 97% of the samples. Meanwhile the unpenalized method and BIC criterion often chose a more parsimonous model with fewer mixture components. As for the out-of-sample misclassification rate, the BIC selection technique showed a small improvement compared to the unpenalized case.

Example 9 (Imbalanced Mixture): In this example, the data is generated using two groups, $K = 3$, with imbalanced true mixture proportions of $p = (0.6, 0.2, 0.2)$, each group containing eight covariates plus an intercept, $p = 8$ and a heterogeneous population variance, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$. The

complete coefficient vector is $\beta^* = (5\,5\,5\,5\,0\,5\,0\,0\,0\,,5\,5\,0\,0\,0\,5\,0\,5\,5\,,0\,0\,0\,5\,0\,5\,0\,0\,0)$. Each dataset contained a sample size of 150 observations.

Table 4.9 (Row 1-2): Estimated mean squared errors of θ for the MLE, PLV-FMR L_1 penalty estimator and PLV-FMR adaptive L_1 penalty estimator given 3 Groups, 8 covariates (Heterogeneous). (Row 3): Proportion of samples with correctly estimated mixture order. (Row 4): OOS misclassification rate. (Row 5): Average number of collapsed variables. (Row 6): Proportion of correctly identified model structures. This table reflect samples of 100 observations.

MSE	Unpenalized	Adaptive Double Penalty AIC	BIC
β	15.0585 (1.557)	15.2150 (1.547)	15.1453 (1.461)
σ^2	4.2202 (0.618)	1.2495 (0.194)	1.1398 (0.196)
K	0.5120 (0.031)	0.7320 (0.028)	0.5120 (0.031)
Misclass (%)	0.1758 (0.004)	0.2479 (0.012)	0.2493 (0.011)
MCV	0.0000 (0.000)	7.0000 (0.530)	10.0000 (0.802)
ACM (%)	0.0000 (0.000)	0.0080 (0.003)	0.0080 (0.003)

In this example, the model parameter space is the same as example 8, thus there are 14 total shared effects, including the shared variance components. In this case, the penalty method performs similarly with respect to the MSE for the coefficient vector. The table 4.9 reflects the challenge of estimating the coefficient parameters accurately. As for the variance components, the penalty method is clearly more accurate than the unpenalized case. This is the opposite impact compared to the balanced example above. Furthermore, the penalty method struggles to accurately classify the out-of-sample observations. The penalty method still displays an ability to identify a number of shared effects. Yet the proportion of correct models identified significantly decreased compared to the balanced case. The penalty method, using the BIC criterion was able to properly identify the complete correct model $< 1\%$ of the samples. The proportion of coefficient parameter vectors correctly identified was also $< 1\%$. The penalty methods weak performance in this case is clearly reflected in its inability to estimate the correct number of mixture components. The unpenalized case only estimated the correct mixture order $\approx 50\%$ of the samples. This makes it very challenging to apply penalty methods when the starting values clearly are not as reliable.

Example 10 (Increased Covariates): In this example, the data is generated using three groups, $K = 3$, with true mixture proportions of $\pi = (0.35, 0.35, 0.3)$, each group containing fourteen covariates plus an intercept, $p = 14$ and a heterogeneous population variance, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$. The

complete coefficient vector is:

$$\beta^{(c)} = (5\,5\,5\,5\,0\,5\,0\,0\,0\,5\,5\,5\,5\,5\,5\,, 5\,5\,0\,0\,0\,0\,5\,5\,5\,5\,0\,0\,5\,0\,5\,, 0\,0\,0\,5\,5\,5\,0\,0\,0\,0\,5\,0\,5\,0\,5\,)$$

. Each dataset contained a sample size of 250 observations.

Table 4.10 (Row 1-2): Estimated mean squared errors of θ for the MLE, PLV-FMR L_1 penalty estimator and PLV-FMR adaptive L_1 penalty estimator given 3 Groups, 14 covariates (Heterogeneous). (Row 3): Proportion of samples with correctly estimated mixture order. (Row 4): OOS misclassification rate. (Row 5): Average number of collapsed variables. (Row 6): Proportion of correctly identified model structures. This table reflect samples of 250 observations.

MSE	Unpenalized	Adaptive Double Penalty AIC	BIC
β	0.9140 (0.026)	0.6704 (0.018)	0.5754 (0.014)
σ^2	0.3569 (0.017)	0.3347 (0.022)	0.3160 (0.020)
K	0.9280 (0.016)	0.9120 (0.017)	0.9120 (0.017)
Misclass (%)	0.0980 (0.000)	0.0951 (0.000)	0.0936 (0.000)
MCV	0.0000 (0.000)	13.0000 (0.330)	17.0000 (0.032)
ACM (%)	0.0000 (0.000)	0.0000 (0.000)	0.0000 (0.000)

In this example, the mixture components are balanced and there are 51 potential parameters for estimation. Observing the β^* vector, there are 24 shared effects including the 2 shared variance component parameters. Table 4.10 shows the effectiveness of the penalty method for the coefficient parameter vector. There is a significant decrease in MSE for the coefficient parameter vector, but not for the variance component parameters. In fact, the penalty method improves in estimation of variance components, but only marginally. Again the misclassification rate for the penalty method slightly improves. The BIC criterion was able to consistently identify 17 shared effects. Considering there are a total of 24 shared effects, the penalty method was not able to identify the full correct model structure for any of the samples. The penalty method was not able to identify the correct coefficient parameter vector. While the unpenalized method was able to select the correct mixture order $\approx 93\%$ of the samples generated, the penalty methods accurately identified the correct mixture order $\approx 91\%$ of samples. It is clear for mixture models with a larger number of model parameters, the sample size must increase correspondingly.

Numerical Summary

The PLVR model has displayed clear effectiveness in the ability to identify shared effects across multiple mixture components. The numerical analysis conducted examined many factors and impacts for the proposed methodology. For the homogeneous examples, the penalty methods consistently outperformed the unpenalized approach for each factor studied. In cases with sufficient sample size, the parameter estimate errors were reduced over 50%. The first four examples displayed the performances of the adaptive penalty compared to the L_1 penalty technique. For each of the examples provided, the adaptive penalty was as effective or more effective than the L_1 penalty method. The penalty methods were most effective at estimating the model paramters and identifying the shared effects. Within the penalty methods, the BIC consistently outperformed the AIC in parameter estimation. The BIC selection criterion was more consistent in identifying the proper model structure and shared parameter effects. Given a sufficient sample size, the adaptive penalty with BIC selection criterion could identify the true model structure over 90% of the samples. For each homogeneous example, the penalty method was most accurate for determining the correct mixture order.

When the individual mixture model variance components parameters were introduced, the performance of the proposed penalty method was not always as effective. Given the double adaptive penalty method was compared directly to the unpenalized method, the double penalty often showed improvements in parameter estimation and idenification of shared effects, yet the performance was dependent on model assumptions and factors. For the balanced and well separated case, the BIC, penalty method remained the most accurate approach. Yet, when sample sizes waned, or mixture proportions were unbalanced, the double penalty method resulted in ballooning errors. More specifically, in unbalanced mixture settings, the double penalty performed poorly for the coefficient parameter estimation, but maintained accuracy for the variance component parameters. Conversely, for balanced designs, the double penalty method produced minimal errors for the coefficient parameters, but only moderate results for the variance component parameters. The ability to identify shared effects proved more challenging for limited sample sizes and increased parameter space. As the number of coefficient parameters increased, a corresponding increase in sample size is needed to maintain the same efficient level of model identification. For the heterogeneous designs, the double penalty method was less consistent identifying the true full model structure. This became more evident as the parameter space increased with a restricted sample size. Lastly, with the heterogeneous design, estimating the appropriate number of mixture components proved more challenging. The proposed BIC, double penalty method often chose a more parsimonous model, and thus would under estimate the mixture model order. Where as, the AIC often outperformed the BIC and unpenalized approach, due to it's nature of choosing a less parsimonous model.

Overall, given balanced, feature rich and data rich designs, the penalized LVR models proved

very effective and accurate compared to the unpenalized approach.

CHAPTER

5

CONCLUSION

This chapter summarizes the key findings and identifies problems with the PLV-FMR estimation approach that would benefit from further research and investigation.

5.0.1 Key Findings

The objective of this research was to evaluate and summarize the effectiveness of a new estimation method (PLV-FMR) for the finite mixture of regression model. The PLV-FMR estimation method was designed to focus on scenarios where the impacts of a given covariates are presumed nearly the same in different sub-populations, but unique in others. We have accomplished the goal by demonstrating the PLV-FMR estimation method can improve parameter estimation and produce a more parsimonous model than the original MLE. Moreover, the PLV-FMR method didn't sacrifice predictability.

We observed that with a balanced mixture design, and sufficient sample size, the PLV-FMR significantly improved the coefficient parameter estimation. In some cases, reducing MSEs up to 50% in simulated examples. Futhermore, we showed that the PLV-FMR estimator asymptoticly achieved the oracle property. That is, as sample size increases, the probability of selecting the correct model structure and shared covariate effects tends to one. This was reflected in the simulation by the PLV-FMR estimator's ability to identify the correct mixture order at a higher proportion than the MLE and its ability to identify the correct model structure at a higher proportion than the MLE. Despite the PLV-FMR estimator's ability to perform in healthy data examples, the estimator and the

MLE still struggled with accuracy for imbalanced data examples.

The PLV-FMR estimation approach uses a penalization approach to avoid the excessively iterative stepwise techniques to traditional model selection. In doing so, this allowed us to reformulate our proposed function to leverage a quadratic programming agorithm. In light of the effectiveness as a result of the PLV-FMR estimation approach, there are still some areas that could benefit from further research.

5.0.2 Further Research

The PLV-FMR estimator has proven effective in many ways. Yet there are some areas for improvement and opportunities for further research. Here are a few possible problems to address:

Data Dependent Mixture Proportions: Through out this manuscript, we assumed the mixture proportions were constant with respect to changes of covariates included in the model. Moreover, the mixture proportions were estimated parametrically as a function of weighted mixture densities. One extension to this research would be to incorporate changes in the covariates when estimating the mixture proportions. This implies the mixture distribution would resemble (2.6) and allow for flexible estimation approaches for the mixture proportions.

Variable Selection: The PLV-FMR estimation method leverages common penalization techniques that are traditionally used for variable selection methods. Yet, the PLV-FMR is not a variable selection method. In fact, each of the covariates introduced in the orginal input dataset are assumed to be effective and hence used as part of the analysis. In many real world cases, many covariates introduced to the model will lack relevance for given mixture components. This opens the door for variable selection techniques to supplement the PLV-FMR analysis. The variable selection steps would best occur before the PLV-FMR estimator is calculated. Khalili and Chen implemented penalty functions for variable selection within mixture components [KC07]. It would be interesting to incorporate a variable selection component independently, or concurrently with the PLV-FMR estimation method.

Enhanced Tuning: The simulation study evaluated the effectiveness of the PLV-FMR estimator. The penalization function natively required significant tuning for the optimal penalty parameters. During this tuning process, we used a suboptimal grid search across a granular two dimensional range of values. Although grid searches are popular, in the case of this analysis, there are certainly more efficient methods for finding the optimal tuning parameters. For instance, k-fold cross-validation or random grid search techniques. Implementing a more efficient tuning approach could significantly reduce the computation time of the proposed methodology.

Alternative Mixture Densities: This research solely focused on the Guassian density for the mixture model structure. Although most basic regression techniques heavily depend on Guassian-like behavior, there are other density functions that could generalize the linear covariate structure.

Such as, quantile, logistic or poisson regressions. The PLV-FMR estimation method could naturally extend to alternative mixture densities. This will likely require distinct and creative numerical estimation approaches, but would be intriguing to assess its performances.

Bibliography

[Aka09] Akaike, H. "Information Theory and an Extension of the Maximum Likelihood Principle". *2nd Int. Symp. on Information Theory* **367** (2009), pp. 4339–4359.

[Ban09] Banks, D. L. et al. "Cherry-picking For Complex Data: Robust Structure Discovery". *The Royal Society* **367** (2009), pp. 4339–4359.

[Bie03] Biernacki, C. et al. "Choosing Starting Values For The EM Algorithm For Getting The Highest Likelihood In Multivariate Gaussian Mixture Models". *Computational Statistics and Data Analysis* **41**.3-4 (2003), pp. 561–575.

[BR09] Bondell, H. & Reich, B. "Simultaneous Factor Selection and Collapsing Levels in ANOVA". *Biometrics* **65** (2009), pp. 169–177.

[CS96] Celeux, G. & Soromenho, G. "An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model". *Journal of Classification* **13**.2 (1996), pp. 195–212.

[Cia07] Ciampi, A. et al. *Locally Linear Regression and the Calibration Problem for Micro-Array Analysis*. 2007, pp. 549–555.

[Dem77] Dempster, A. P. et al. "Maximum Likelihood Estimation From Incomplete Data Via The E-M Algorithm". *Journal of Royal Statistical Society, Series B* **39** (1977), pp. 1–38.

[DW95] DeSarbo, W. & Wedel, M. "A Mixture Likelihood Approach for Generalized Linear Models ". *Journal of Classification* **12** (1995), pp. 21–155.

[DC88] DeSarbo, W. S. & Cron, W. L. "A Maximum Likelihood Methodology for Clusterwise Linear Regression". *Journal of Classification* **5** (1988), pp. 249–282.

[DR94] Diebolt, J. & Robert, C. P. "Estimation of Finite Mixture Distributions Through Bayesian Sampling". *Jornal of Royal Statistical Society, Series B* **56**.2 (1994), pp. 363–375.

[FL01] Fan, J. & Li, R. "Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties". *Journal of the American Statistical Association* **96**.456 (2001), pp. 1348–1360.

[FJ02] Figueiredo, M. A. T. & Jain, A. K. "Unsupervised Learning of Finite Mixture Models". *IEEE Transactions On Pattern Analysis and Machine Intelligence* **24**.3 (2002), pp. 381–396.

[HL97] Hadi, A. & Luceno, A. "Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms". *Computational Statistics and Data Analysis* **25**.3 (1997), pp. 251–272.

[Jac91] Jacobs, R. et al. "Adaptive mixtures of local experts". *Journal of Neural Computation* **3**.1 (1991), pp. 79–87.

[KR89] Kamakura, W. & Russell, G. “A Probabilistic Choice Model for Market Segmentation and Elasticity Structure”. *Journal of Marketing Research* **26**.4 (1989), pp. 379–390.

[KC07] Khalili, A. & Chen, J. “Variable Selection in Finite Mixture of Regression Models”. *Journal of the American Statistical Association* **102**.479 (2007), pp. 1025–1038.

[Lei04] Leisch, F. “FlexMix: A general framework for finite mixture models and latent class regression in R”. *Journal of Statistical Software* **11**.8 (2004), pp. 1–18.

[Mar00] Markatou, M. “Mixture models, robustness, and the weighted likelihood methodology”. *Biometrics* **56**.2 (2000), pp. 483–486.

[McL82] McLachlan, G. J. “The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis”. *Handbook of Statistics* **2** (1982), pp. 199–208.

[MB88] McLachlan, G. J. & Basford, K. E. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, 1988.

[MP00] McLachlan, G. J. & Peel, D. *Finite Mixture Models*. John Wiley and Sons, Inc., 2000.

[Mur12] Muruzabal, J. et al. “SOMwise Regression: A New Clusterwise Regression Method”. *Neural Computing and Applications* **21** (2012), pp. 1229–1241.

[QR73] Quandt, R. E. & Ramsey, J. B. “Estimating Mixtures of Normal Distributions and Switching Regressions”. *Journal of the American Statistical Association* **73** (1973), pp. 730–738.

[Sch78] Schwarz, G. “Estimating the Dimension of a Model”. *The Annals of Statistics* **6**.2 (1978), pp. 461–464.

[Spa82] Spath, H. “Algorithm 48: A Fast Algorithm for Clusterwise Linear Regression”. *Computing* **29** (1982), pp. 175–181.

[Tei63] Teicher, H. “Identifiability of Finite Mixtures”. *Annals of Mathematical Statistics* **34** (1963), pp. 1265–1269.

[Tib96] Tibshirani, R. “Regression Shrinkage and Selection via the Lasso”. *Journal of the Royal Statistical Society* **58**.1 (1996), pp. 267–288.

[Tit90] Titterington, D. M. “Some Recent Research in the Analysis of Mixture Distributions”. *A Journal of Theoretical and Applied Statistics* **21**.4 (1990), pp. 619–641.

[WL07] Wang, H. & Leng, C. “Unified LASSO Estimation by Least Squares Approximation”. *Journal of the American Statistical Association* **102**.479 (2007), pp. 1039–1048.

[Wed93] Wedel, M. “A Latent Class Poisson Regression Model for Heterogeneous Count Data”. *Journal of Applied Econometrics* **8**.4 (1993), pp. 397–411.

[XH01] Xu, W. & Hedeker, D. “A random-effects mixture model for classifying treatment response in longitudinal clinical trials”. *Journal of Biopharmaceutical Statistics* **11**.4 (2001), pp. 253–273.

[Yao14] Yao, W. et al. “Robust Mixture Regression Using the T-distribution”. *Computational Statistics and Data Analysis* **71**.C (2014), pp. 116–127.

[Zou06] Zou, H. “The Adaptive Lasso and Its Oracle Properties”. *Journal of the American Statistical Association* **101**.476 (2006), pp. 1418–1429.

APPENDIX

Proofs

THEOREM 1: Let the order-K FMR design be such that $\pi_k > 0$, $k = 1, \ldots, K < \infty$ and $\lambda \to \infty$ and $\lambda = o(n^{1/2})$ then, the adaptive PLV-FMR estimator $\tilde{\boldsymbol{\beta}}^*$ and its corresponding estimate of differences $\tilde{\boldsymbol{\beta}}_z^{\mathscr{A}}$ have the following properties:

$$
\begin{aligned}
&(i) \quad \mathscr{P}[\mathscr{A}_n = \mathscr{A}] \to 1 \\
&(ii) \quad \sqrt{n}(\tilde{\boldsymbol{\beta}}_z^{\mathscr{A}} - \boldsymbol{\beta}_z^{\mathscr{A}}) \to \mathscr{N}(\mathbf{0}, \Sigma_{\boldsymbol{\beta}_z^{\mathscr{A}}})
\end{aligned}
$$

The theorem 1 part (i) corresponds to the consistency in variable selection. That is, the guarantee for selecting the correct structure and identifying the covariates that share impacts across mixture components, with probability tending to 1. Part (ii) assures the asymptotic behavior is the same as the oracle estimator. The asymptotic covariance matrix contains the parameters that remain under the true mixture structure after removing unnecessary coefficients, detailed by $\mathscr{A}$. Note the design condition requires non-vanishing mixture proportions and a pre-specified finite number of sub-populations. The proof for theorem 1 can be found in the appendix to this manuscript.

Proof of Theorem 1 (i)

The flow of this proof follows that of Bondell and Reich ([BR09]) with appropriate adaptations to the proposed methodology. Moreover, for purposes of this proof, W.L.O.G. only differences of a single covariate, say $j = 1$, across multiple mixture components is considered in the design. In a full design, each covariate and its corresponding inter-component differences are included in the analysis.

First, we reintroduce the appropriate notations. Let $\boldsymbol{\beta}^* = [\boldsymbol{\beta}^T, \boldsymbol{\beta}_z^T]^T$ where $\boldsymbol{\beta}^T$ and $\boldsymbol{\beta}_z^T$ are the order-K mixture model vector regression coefficients and vector of coefficient difference parameters, respectively. Moreover, set $\mathscr{A} = \{(j,k,m) \ : \ \beta_{jk} \neq \beta_{jm}\}$ as the set of indices for the differences that are truly non-zero (note: for this proof $j = 1$ is assumed). Define $\mathscr{A}_n$ as the set of indices for differences estimated to be non-zero. This implies, if one is given the indices of set $\mathscr{A}$, then the analyst can simply reduce the parameter vector of differences, which is the same as replacing equivalent cross-component paramters with a single parameter.

By the given definition, $\mathscr{A}^c$ and $\mathscr{A}_n^c$ represent the set of indices for differences to be true zero and estimated to be zero, respectively. Let $B_n = \mathscr{A}_n \cap \mathscr{A}^c$, that is the set of differences that should be zero, but were incorrectly estimated as non-zero. Then we aim to show that $P[B_n \neq 0] \to 0$. That is, the probability of the set of incorrect non-zero differences, remaining non-empty converges to zero. Thus, we aim to show the set of true zero difference parameters will eventually be set to zero.

One must first note that a given non-baseline difference parameter set to zero could have a cascading effect on other differences and corresponding coefficients. In short, multiple non-baseline difference parameters that are simultaneously zero represent a set of complex cases. To

avoid this complexity in the proof, consider a set of ordered mixture coefficients for a given covariate $\tilde{\beta}_{d_1} \leq \tilde{\beta}_{d_2} \leq \ldots < \tilde{\beta}_{d_K}$ where $\tilde{\beta}_d \equiv \tilde{\beta}_{1k}$ for some $k \in \{1,\ldots,K\}$. That is, there exists at least one non-baseline difference greater than zero. Next we assume the set B_n is NOT empty. For this proof, it is assumed there is a minimum of one mixture coefficient element in the set B_n. More specifically, let $\tilde{\beta}_{1m}$ represent the estimator of the largest coefficient contained in the set of difference parameters indexed by B_n. That is, $m = \max\{l : (k,l) \in B_n \text{ for some } k\}$. Let $\tilde{\beta}_{1q}$ represent the estimator of the smallest coefficient such that the pair $(q,m) \in B_n$, $\tilde{\beta}_{1q} < \tilde{\beta}_{1m}$ and $\tilde{\beta}_{1m} - \tilde{\beta}_{1q} > 0$ with $q < m$. Due to multiple permutations, consider the design matrix that corresponds to component q of the mixture as the baseline.

The aforementioned difference parameters of $\boldsymbol{\beta}_z$ can be calculated with a full-rank matrix of differences D, s.t. $D\boldsymbol{\beta} = \boldsymbol{\beta}_z$, where $D = [D_1,\ D_2,\ldots,D_K]$ is defined in (3.3). Also notice, the subscript distinction for individual parameters. Next define $\boldsymbol{\beta}_{z_k} = \beta_k - \beta_q$ for $k \neq q$. This parameter β_{jz_k} is equivalent to difference parameters defined earlier 3.5. By assumption of this proof, $\beta_{1z_m} = \beta_{1m} - \beta_{1q} = 0$ but $\tilde{\beta}_{1z_k} = \tilde{\beta}_{1k} - \tilde{\beta}_{1q} \neq 0$. This implies for all $(k,m) \in B_n$, $\tilde{\beta}_{1z_m} - \tilde{\beta}_{1z_k} > 0$. Given this parameterization, the solution for any $k < l$ is as follows:

$$|\beta_{1l} - \beta_{1k}| = \begin{cases} \beta_{1z_l} - \beta_{1z_k}, & k \neq q, l \neq q \\ \beta_{1z_l}, & k = q, l > q \\ -\beta z_k, & k < q, l = q \end{cases}$$

Using the parameterization differences, define the design matrix s.t. $\mathbf{X}_k\boldsymbol{\beta}_k = \mathbf{R}_k\boldsymbol{\beta}_{z_k}$ where $\boldsymbol{\beta}_{z_k}$ is the vector of difference coefficients corresponding to mixture component k and newly parameterized design matrix $R_k = X_k(D_k^T D_k)^{-1}\boldsymbol{\beta}_k^T$ a function of the k^{th} mixture component design matrix and difference of coefficients transformer D_k. Thus one can express the re-parameterized objective function

$$\tilde{\boldsymbol{\beta}}_z^{\mathscr{A}} = argmin\left[\sum_i \sum_k (\boldsymbol{y}_i - \mathbf{r}_{ik}\boldsymbol{\beta}_{z_k})\hat{\pi}_{ik} + p_\lambda(\beta_z)\right]$$

where

$$p_\lambda(\boldsymbol{\beta}_z) = \lambda\left\{\sum_{1\leq k<l\leq K} \omega^{(lk)}(\beta_{z_l} - \beta_{z_k}) + \sum_{q<k\leq K} \omega^{(qk)}\beta_{z_k} - \sum_{1\leq k<q} \omega^{(kq)}\beta z_k\right\}$$

In the above equation $\omega^{(lk)} = \frac{1}{|\hat{\beta}_l - \hat{\beta}_k|}$ with corresponding least squares estimates $\hat{\beta}$.

Under this new parameterization, given $\tilde{\beta}_{1z_m} \neq 0$, the optimization objective function is differentiable with respect to β_{1z_m}, evaluated at the solution $\tilde{\beta}_{1z_m}$. The basis of this proof focuses on observing the derivative in a neighborhood of β_{1z_m} and obtaining a contradiction. Taking the

derivative and solving results in the following expression:

$$\sum_k 2\boldsymbol{r}_{(m),k}^T[(\boldsymbol{y}-\mathbf{R}_k\boldsymbol{\beta}_{z_k})\hat{\boldsymbol{\pi}}_k] = \lambda \sum_{k\neq m,\,(k,m)\in\mathscr{A}} (-1)^{I_{m<k}}\omega^{(km)} + \lambda \sum_{k\neq m,\,(k,m)\in\mathscr{A}^c} \omega^{(km)}$$

where $\boldsymbol{r}_{(m),k}$ is the m^{th} column of the design matrix R. Also, $\sum_{k\neq m,\,(k,m)\in\mathscr{A}^c} = \sum_{(k,m)\in\mathscr{A}_n\cap\mathscr{A}^c}\omega^{(km)} + \sum_{(k,m)\in\mathscr{A}_n^c\cap\mathscr{A}^c}\omega^{(km)}$. The second term on the right-hand side vanishes since it would be a zero vector in the original function. Multiplying by $n^{-1/2}$ to the equation leaves

$$\frac{2}{\sqrt{n}}\sum_k \boldsymbol{r}_{(m),k}^T[(\boldsymbol{y}-\mathbf{R}_k\boldsymbol{\beta}_{z_k})\hat{\boldsymbol{\pi}}_k] = \frac{\lambda}{\sqrt{n}}\sum_{\mathscr{A}}(-1)^{I_{m<k}}\omega^{(km)} + \frac{\lambda}{\sqrt{n}}\sum_{\mathscr{B}}\omega^{(km)}.$$

Observe the RHS of the above equation. $\frac{\lambda}{\sqrt{n}}\omega^{(km)} = \frac{\lambda}{\sqrt{n}}|\hat{\beta}_k - \hat{\beta}_m|^{-1} > 0$. For all $(k,m)\in\mathscr{A}$, due to continuous transformations of continuous functions, it follows $|\hat{\beta}_k-\hat{\beta}_m|^{-1} = O_p(1)$ and by assumption $\frac{\lambda}{\sqrt{n}}\to\lambda_0<\infty$. Moreover, for $(k,m)\in\mathscr{A}^c$ and the properties of $\sqrt{n}$-consistency of the least squares estimators, $\sqrt{n}^{-1}|\hat{\beta}_k-\hat{\beta_m}|^{-1} = O_p(1)$. This implies the second term on RHS is $O_p(\lambda_n)$, which by assumption $\lambda\to\infty$. Observing the LHS, at the solution $\tilde{\boldsymbol{\beta}}_z$: $2\sum_k \boldsymbol{r}_{(m),k}^T[\sqrt{n}\frac{(\boldsymbol{y}-\mathbf{R}_k\tilde{\boldsymbol{\beta}}_{z_k})\hat{\boldsymbol{\pi}}_k}{n}]\to_d$ to some order-K mixture of normal distributions. As a result, the LHS is $O_p(1)$, but the RHS is $\to\infty$ thus we have a contradiction.

Proof of Theorem 1(ii)

This proof of asymptotic normality is an adaptation of the proof provided by Zou (2006) [Zou06]. The argument presented continues with the previously defined penalty of differences parameterization introduced in proof of Theorem 1(i). First consider true difference parameter vector $\boldsymbol{\beta}_{z0}$ and let $\tilde{\boldsymbol{\beta}}_z = \boldsymbol{\beta}_{z0} + \frac{\tilde{\boldsymbol{u}}}{\sqrt{n}}$ this implies $\tilde{\boldsymbol{u}} = \sqrt{n}(\tilde{\boldsymbol{\beta}}_z - \boldsymbol{\beta}_{z0}$. Define

$$\Psi_n(\boldsymbol{u}) = \left[\sum_i\sum_k (\boldsymbol{y}_i - \mathbf{r}_{ik}(\boldsymbol{\beta}_{z0}+\frac{\boldsymbol{u}}{\sqrt{n}}))\hat{\pi}_{ik} + p_A(\boldsymbol{\beta}_{z0}+\frac{\boldsymbol{u}}{\sqrt{n}})\right]$$

where

$$p_\lambda(\boldsymbol{\beta}_z) = \lambda\left\{\sum_{1\le k<l\le K}\omega^{(lk)}(\beta_{z_l}-\beta_{z_k}) + \sum_{q<k\le K}\omega^{(qk)}\beta_{z_k} - \sum_{1\le k<q}\omega^{(kq)}\beta_{z_k}\right\}$$

Thus $\tilde{\boldsymbol{u}} = \arg\min\{\Psi_n(\boldsymbol{u})\}$. Next define $V_n(\boldsymbol{u}) = \Psi_n(\boldsymbol{u}) - \Psi_n(\mathbf{0})$. After refactoring, one can express V with three terms as follows: $V_n(\boldsymbol{u}) = A_1 + A_2 + A_3$.

$$A_1 = \frac{1}{\sqrt{n}}\sum_k \boldsymbol{u}^T\boldsymbol{R}^T\boldsymbol{P}_k\boldsymbol{R}\boldsymbol{u}$$

$$A_2 = \frac{-2}{\sqrt{n}} \sum_k \boldsymbol{\epsilon}^T \boldsymbol{R} \boldsymbol{P}_k \boldsymbol{u}$$

$$A_3 = \frac{\lambda}{\sqrt{n}} \boldsymbol{P}_V(\boldsymbol{u})$$

where P_k is the diagonal matrix of individual observation mixture posterior probabilities. Also define $P_V(\boldsymbol{u})$ as

$$\begin{aligned} P_V(\boldsymbol{u}) = &\textstyle\sum_{1 \le k < l \le K,\, k,l \neq q} \omega^{(kl)} (|\beta_{z_l 0} - \beta_{z_k 0} + \frac{u_l - u_k}{\sqrt{n}}| - |\beta_{z_l 0} - \beta_{z_k 0}|) \\ &+ \textstyle\sum_{k \neq q} (-1)^{k<q} \omega^{(qk)} \sqrt{n} (|\beta_{z_k 0} + \frac{\boldsymbol{u}_k}{\sqrt{n}}| - |\beta_{z_k 0}|) \end{aligned}$$

Next we observe the limiting behavior of V_n. Similar to Zou's argument, the terms A_1 and A_2 have well defined limiting behavior, based on finite mixture model framework. Since $\frac{\lambda}{\sqrt{n}} \to 0$, the last term $\frac{\lambda}{\sqrt{n}} P_V(\boldsymbol{u})$ will go to zero, unless under the correct structure; $\mathscr{A}$. Thus if one assumes the correct structure, set $\mathscr{A}$, and collapses the appropriate difference parameters and identify the corresponding design matrix. In the mixture model setting, this also includes recognizing the true component structure and observation component assignment. Now define $\boldsymbol{R}_0$ as the full design matrix corresponding to $\boldsymbol{\beta}_z$ and set $\mathscr{A}$. Then $\frac{1}{\sqrt{n}} \sum_k \boldsymbol{R}_{0k}^T \boldsymbol{P}_k \boldsymbol{R}_{0k} \to \boldsymbol{M}$, a positive definite matrix, and $\frac{\sum_k \boldsymbol{\epsilon}^T \boldsymbol{R}_{0k} \boldsymbol{P}_k}{\sqrt{n}} \to W_{\mathscr{A}} = \sum_k \pi_k N(\boldsymbol{0}, \tau_k \boldsymbol{M})$, an order-K mixture of normals, where τ_k is the variance parameter for the mixture component $k \in \{1, \ldots, K\}$. It is important to note the assumption of non-vanishing mixture components, $\sum_i \pi_{ik} \to \pi_{k0} > 0$. Then one can use Slutsky's Theorem to show $V_n(\boldsymbol{u}) \to V(\boldsymbol{u})$ where,

$$V(\boldsymbol{u}) = \begin{cases} \boldsymbol{u}_0^T \boldsymbol{R} \boldsymbol{u}_0 - 2\boldsymbol{u}_0^T \boldsymbol{W}_{\mathscr{A}}, & \text{under } \mathscr{A} \\ \infty, & \text{otherwise} \end{cases}$$

Zou's proof then notes since $V(\boldsymbol{u})$ is convex, the unique minimizer of $V(\boldsymbol{u})$ is $[\boldsymbol{M}^{-1} \boldsymbol{W}_{\mathscr{A}}, \boldsymbol{0}]^T$. This implies $\tilde{\boldsymbol{u}} \to \sum_k \pi_k N(\boldsymbol{0}, \sigma_k^2 \boldsymbol{M}^{-1})$ a mixture of normals. This completes the proof.