Statistica Sinica Preprint No: SS-2016-0531	
Title	A New Semiparametric Approach to Finite Mixture of Regressions using Penalized Regression via Fusion
Manuscript ID	SS-2016-0531
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0531
Complete List of Authors	Erin Austin Wei Pan and Xiaotong Shen
Corresponding Author	Erin Austin
E-mail	erin.e.austin@ucdenver.edu
Notice: Accepted version subject to English editing.	

A New Semiparametric Approach to Finite Mixture of Regressions using Penalized Regression via Fusion

Erin Austin[1], Wei Pan[2], Xiaotong Shen[3]

[1] *Department of Mathematical and Statistical Sciences, University of Colorado Denver, 80204*
[2] *Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455*
[3] *School of Statistics, University of Minnesota, Minneapolis, MN 55455*

Abstract: For some modeling problems a population may be better assessed as an aggregate of unknown subpopulations, each with a distinct relationship between a response and associated variables. The finite mixture of regressions (FMR) model, where an outcome is derived from one of a finite number of linear regression models, is a natural tool in this setting. In this article we first propose a new penalized regression approach, then we demonstrate how it can, in some types of problems, better identify subpopulations and their corresponding models than a semiparametric FMR method. Our new method fits models for each person via grouping pursuit, utilizing a new group truncated L_1- penalty (gTLP) that shrinks differences between estimated parameter vectors. The methodology causes the individuals' regression coefficients to cluster into a few common models, in turn revealing previously unknown subpopulations. In fact, by varying the penalty strength, the new method can reveal a hierarchical structure among the subpopulations that can be useful in exploratory analysis. Simulations using FMR models and real data analysis show the performance of the method is promising.

Key words and phrases: FMR, Group LASSO, Group TLP, Grouping Pursuit, Penalized Regression, Semiparametric.

1. Introduction

A traditional way to assess the association between candidate variables and an outcome of interest is to generate model estimates at a population level. However, it is often reasonable to hypothesize that for different, unknown subpopulations, an outcome results from different sets of variables (or possibly from different sized effects of the same variables). For example, a disease outcome may be a function of different sets of genetic variants for different groups of individuals within a population. Modeling approaches that don't account for subpopulation induced heterogeneity and the possibility of subpopulation specific effect sizes could easily fail to identify factors associated with a response for only some of the subpopulations.

Statistically, modeling outcomes for a population may in fact require the assumption of a distinct relationship for distinct but unknown subpopulations. One modeling frame-

work useful for this strategy is the finite mixture of regressions (FMR) model. Here, an individual's outcome is predicted from one regression model (known as a component) out of a set of possible regression models. Because the actual component is unknown for any given observation, a natural choice for fitting FMR models is the Expectation-Maximization (EM) algorithm of Dempster, Laird, and Rubin (1977). Methods based on the EM algorithm yield density estimates and component level regression coefficient estimates depending on the likelihood assumptions used when fitting the model. Wedel and DeSarbo (1995) showed how the algorithm could successfully estimate regression parameters for mixtures of common distributions such as normal or binomial. An EM-like algorithm was developed by Benaglia, Chauveau, and Hunter (2009) to allow for more generality in the error term. While able to lower error rates, it is unclear what objective function is being maximized and whether successive iterations guarantee an increase in the objective function. A maximum smoothed likelihood algorithm was created by Levine, Hunter, and Chauveau (2011) to remedy the Benaglia shortcomings. The algorithm's advantages, though, did not hold when using the Benaglia, Chauveau, and Hunter (2009) approach to updating bandwidths. Subsequently, Hunter and Young (2012) developed a semiparametric EM-like algorithm, removing the parametric assumptions on the components, that was successful when the initialization was directed towards true values. EM-based algorithms have been successful in FMR problems where it is possible to (1) specify the mixture distribution and its corresponding number of components and (2) initialize the algorithm.

The EM algorithm has served as the main statistical tool for another category of approaches to subpopulation estimation: clustering subject-specific regression models. In early work, DeSarbo and Cron (1988) used the EM algorithm for clusterwise linear regression. The methodology estimated sets (one per cluster) of linear regression parameters assuming normal densities and a given number of clusters. Interested in the model-based clustering of cyclone tracks or curves, Gaffney and Smyth (2003) used a maximum a posteriori (MAP) EM algorithm for random effects regression mixtures under the assumption that they were from one of k prespecified subpopulations that follow a normal density. While still dependent on the density and components assumptions, the work of these authors demonstrated the potential for the clustering of subject-specific models.

In settings where the number of subpopulations is unknown or the error distribution cannot be reasonably assumed, alternatives to or enhancements of the EM algorithm must be considered. Penalized regression as part of FMR model estimation has shown promise as one such improvement. Specific to the goal of variable selection, some investigators have integrated in penalized regression. An effective EM algorithm developed by Khalili and

Chen (2007) for a penalized mixture model was applied to the FMR setting for the purpose of variable selection, but estimation was based on a parametric likelihood assumption. The same authors, Khalili, Chen, and Lin, followed-up in 2010, again with an EM approach using penalized likelihood for variable selection that was effective in simulations at selecting important covariates, but this was after applying a screening method. To our knowledge the most successful approaches to date for estimating FMR models depend on methodology using some form or approximation of the EM algorithm, and thus depend on making successful likelihood assumptions or successful density estimations.

In the following work we take a novel approach towards identifying unknown subgroups and their corresponding regression models via grouping pursuit (fusion). Our approach does not depend on any likelihood assumptions or component density estimations. The key to our methodology is the application of a new type of penalized regression to simultaneous fitting of *separate* regression models for *each* subject. If there exist unknown subpopulations, then the individual fitted models should be the same within the same subpopulation but different across the subpopulations. Specifically, the subjects within a subpopulation share a common model, but the common models differ by subpopulation. Thus, a logical methodological step is the inclusion of a grouping feature to penalize differences in the estimated covariate coefficients *across* individuals. As we will elaborate on shortly, we develop just such a penalty that enables us to force the individuals' models to cluster into a few common models, corresponding to different subpopulations. The methodology can be used as an exploratory data analysis tool akin to hierarchical clustering versus model-based clustering or k-means clustering where the number of clusters is specified.

Penalized regression has been researched to specifically assess its ability to identify and/or leverage groups of variables associated with an outcome. Yuan and Lin (2006) demonstrated that when groups of variables appeared (or disappeared) together in a model, using a group LASSO (Least Absolute Shrinkage and Selection Operator) penalty to select groups of variables or factors (group LASSO) resulted in better performance than the standard LASSO. Another penalized regression approach, the fused LASSO from Tibshirani et al. (2005), added an additional penalty to LASSO specifically for differences in successive regression coefficients. In situations where the features had a natural order, the additional grouping penalty showed promise for both regression and classification. Luo, Wang, and Tsai (2008) proposed a modified EM algorithm for the FMR estimation problem that incorporated penalization of differences in component regression coefficients. The method, called MR-LASSO, demonstrated the ability to use penalization of differences in estimated regression models to identify mixture components. Shen, Huang, and Pan (2012) developed

a penalized regression method for simultaneous supervised clustering and feature selection over a given undirected graph that utilized a truncated-L_1 penalty (TLP) for grouping pursuit. Successful identification and estimation of unknown homogenous groups of effects were possible with their approach. The method used a single linear regression model for a single response, but assumed that the full coefficient vector could be partitioned into subsets of homogeneous coefficients. The new method improved parameter estimation and group identification by penalizing differences within these smaller vectors. In related work, Pan, Shen, and Liu (2013) developed a penalized regression-based clustering (PRclust) method where the TLP penalty was applied to differences in the centroids of data points. PRclust performed well in situations such as non-convex clusters where other more common methods did not. Pivotal to the current work, the success of PRclust demonstrated the potential for comparisons across subjects with a grouping penalty. Subsequently, Chi and Lange (2015) demonstrated how alternating direction method of multipliers or ADMM could be effective when solving the convex clustering problem involving penalized differences in centroids. In fact, Wu et al. (2016) recently provided a new DC-ADMM algorithm that combines difference of convex (DC) programming with ADMM to more efficiently cluster via centroid difference TLP penalization.

The best performing penalized regression-based strategies have forgone explicit use of FMR and have instead compared either subject-level differences in vectors of numerical variables to aggregate into clusters or compared differences in regression model coefficients in known subgroups to collapse into clusters. Post our original 2014 submission, we learned of the closely related work of Ma and Huang (2017). In it the authors presented a subgroup identification method based on pairwise penalization of subject-specific intercepts via a fusion approach with convex penalties. Here, the intercept-only penalization was designed to identify a subgroup structure. Their work is exciting to us as it demonstrated the potential value of our approach because they had success using a methodology that is approximately one case within our more general framework. Ma and Huang emphasized the estimation of the intercept term in subgroup analysis, in contrast to our method which emphasizes a slope parameter in mixture regression. That is, the following work incorporates a grouping pursuit framework to shrink differences between the full subject-specific models for problems similar to FMR. Our approach to penalized regression uses grouping pursuit when simultaneously fitting *separate* models for *each* subject. To be explicit, we penalize only the differences in corresponding parameter estimates between each pair of subject-specific regression models. We study both the LASSO penalty developed by Tibshirani (1996) and the TLP invented by Shen, Pan, and Zhu (2012b) in two ways. First, we penalize without using a group feature

by applying the penalty to the individual coefficient differences. In a sense we are grouping the subjects for each coefficient separately. This approach shrinks differences in the subjects' models parameter by parameter and does not explicitly shrink differences between the full models. Therefore, we next apply two group penalties based on LASSO and TLP to the differences in the estimated parameter vectors for each pair of samples' regression model. Our work extends the research introduced above and, in particular, that of Ma and Huang in a few critical ways. First, we are allowing the subgroups to be defined by differences in observable factors; we feel this is important as it accounts for the possibility that any set of variables might only affect a subset of the population. Second, we incorporate the non-convex TLP penalty in our clustering approach. Previous work, e.g. that of Pan, Shen, and Liu in 2013, demonstrated the advantages of TLP over L_1 penalties when performing subject-level penalization.

When applied, it is our hypothesis that we will see a hierarchical clustering of individual models depending on the magnitude of the penalty and thresholding parameters. In turn we reveal the increasingly granular partitions of the population into subpopulations that result from monotonically changing parameter values. It is important that the method provide a means to choose the number of subpopulations; thus, we describe a generalized cross-validation method to select a best set of subgroups. However, the discussion below focuses on the larger question of finding the clustering paths that arise from our penalized regression-method. In this way we provide a fuller view of the problems where our new method would be valuable in application. The following discussion uses simulated FMR models to permit comparison to previous methods and is followed by application to a relevant true genetics data setting. The intent of the following is to show the new penalized regression-based method can handle FMR models and the clustering of subject-level regression models. Because of this we use this article primarily to establish its efficacy in the cornerstone case of one covariate problems, a necessary step before building to higher dimensions in subsequent work. The resulting estimates are compared to the very successful Hunter and Young (2012) semiparametric FMR which uses an EM-like approach.

2. Methods

In this section we first detail the FMR model. A subsequent section delineates our penalized regression approach and its computation.

2.1 Finite Mixture of Regressions Model

To motivate and contrast with our new method, we briefly review the Finite Mixture of Regressions Model. Using the language of McLachlan and Peel (2000) and notation of

Khalili and Chen (2007), suppose Y_i represents the value of a continuous random variable, or response, for subject $i = 1, ..., n$. Let X_{ji} equal subject i's value for covariate $j = 1, ..., p$; therefore, $X_i = (x_{1i}, x_{2i}, ..., x_{pi})$ is the vector of covariates for subject i. Next, let $f(y; \theta_k(x), \phi_k)$ for $k = 1, \ldots, K$ represent K conditional parametric densities of y given x as a function of a canonical parameter, θ_k, and a dispersion parameter, ϕ_k. Utilize the identity link function $g(\mu) = \mu$ such that $\theta = x\beta = \mu$, and (x, Y) follows a FMR model of order K where the conditional density function of Y given x has the form:

$$f(y; x, \Psi) = \sum_{k=1}^{K} \pi_k f(y; \theta_k(x), \phi_k). \tag{2.1}$$

The FMR model has order $K < \infty$ as it is a mixture of K densities (known as component densities). In this equation the unknown parameters are $\Psi = (\beta_1, \beta_2, ..., \beta_K, \phi, \pi)$, where $\beta_k = (\beta_{1k}, \beta_{2k}, ..., \beta_{pk})^T$, $\phi = (\phi_1, \phi_2, ..., \phi_K)^T$, $\pi = (\pi_1, \pi_2, ..., \pi_{K-1})^T$ such that both $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$.

Parametric approaches by specifying a parametric form of $f(\theta, \phi)$ and estimating $f(\hat{\theta}, \hat{\phi})$ are most common. As described in the introduction, though, parametric approaches can be too restrictive; therefore, we compare our penalized regression approach to a semiparametric method developed by Hunter and Young (2012). Their method estimates each of the component densities by a nonparametric kernel estimate $\hat{f}(\cdot)$ and provides component level regression coefficients based on a specific K. The Hunter and Young method generates K sets of regression coefficient estimates, partly depending on the specified and estimated likelihoods in an EM-like algorithm. To be described in the next section, our method starts with over-specified n sets of regression coefficients and uses grouping pursuit with group penalties to find a hierarchical clustering of the individual regression models without specifying or estimating a parametric model or likelihood.

2.2 A New Semiparametric Approach Based on Penalized Regression

Model

We begin by hypothesizing that the parameters of the underlying model for a response can vary by subpopulation. To capture this we estimate a model for each subject in the study using penalized regression with a group feature intended to reveal subpopulations via clustering among these models.

As before suppose Y_i represents the value of a continuous response for subject $i = 1, \ldots, n$. Again, let $X_i = (x_{1i}, \ldots, x_{pi})$ be the vector of p covariates for subject i. Assume for each subject i there is a subject-specific linear model:

$$Y_i | X_i = \beta_{0i} + X_i \beta_i + \epsilon_i \tag{2.2}$$

where $\beta_i = (\beta_{1i}, \ldots, \beta_{pi})^T$ and $\mathrm{E}(\epsilon_i) = 0$. Please note how we initially allow for a sample-dependent (β_{0i}, β_i^T) for each subject, and we at no time specify or estimate a density function for ϵ_i. Our method is semiparametric as we specify the linear form of the relationship, but we do not use $f(\cdot)$ in the FMR model. That is, no specific parametric distribution is assumed, but note here that asymptotically we require sub-Gaussian tails as will be explained shortly in a discussion of conditions for identifiability.

Observe from our model how the covariates associated with an outcome would have non-zero values in β_i, but we do not assume the set of non-zero coefficients are identical for all i. For example, a set of covariates might affect the responses of only a subset of the populations (affect only a subpopulation). Even in cases where the same set of covariates affect multiple subpopulations, the magnitude and/or direction of effect can vary. That is, a set of covariates might impact the outcome of interest for several subpopulations, but impact each differently. In each of these scenarios there is one overarching principle: if multiple subjects' outcomes result from the group of covariates in the same functional way, then the (β_{0i}, β_i^T)'s for this subset of the population should be identical. In this way we can partition our population into groups defined by identical (β_{0i}, β_i^T)'s. Our method provides estimates for β_{0i} and β_i by minimizing

$$(1/2) \parallel Y - X\beta \parallel_2^2 + \lambda P(\beta)$$

$$\text{with} \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & 1 & X_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & 1 & X_n \end{bmatrix} \quad \text{with} \quad \mathbf{0}^T = \left.\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right\} p+1 \quad , \text{and} \quad \beta = \begin{bmatrix} \beta_{01} \\ \beta_1 \\ \beta_{02} \\ \beta_2 \\ \vdots \\ \beta_{0n} \\ \beta_n \end{bmatrix}.$$

Here, we use an objective function form like Wu et al. (2016) and Ma and Huang (2107) where the penalty parameter λ is applied to a specified penalty, $P(\beta)$. We consider two penalty forms and require $\lambda > 0$ for identifiability: Tibshirani's convex LASSO penalty (Tibshirani, 1996) and Shen, Pan, and Zhu's non-convex truncated L_1- penalty (TLP) (Shen, Pan, and Zhu, 2012b). For our two approaches with respect to the LASSO penalty and grouping pursuit:

1. $P_L(\beta) := LASSO(\beta) := \sum_{i<j} \|\beta_{0i} - \beta_{0j}\|_1 + \sum_{m=1}^{p} \sum_{i<j} \|\beta_{mi} - \beta_{mj}\|_1$
2. $P_{gL}(\beta) := gLASSO(\beta) := \sum_{i<j} \| \binom{\beta_{0i}}{\beta_i} - \binom{\beta_{0j}}{\beta_j} \|_2$

where $\|\cdot\|_1$ is the L_1 norm and $\|\cdot\|_2$ is the L_2 norm. The nongroup version, $P_L(\beta)$, bases selection on the between sample differences in individual coefficient estimates. Depending on

the size of λ, the nongroup version chooses the nonzero differences between final estimated sample models by comparing corresponding parameters separately. In contrast, the group version, $P_{gL}(\beta)$, will shrink differences between the full estimated parameter sets and more likely have $(\beta_{0i}, \beta_i^T) = (\beta_{0j}, \beta_j^T)$.

The LASSO penalty will shrink all coefficient differences. However, if there are in fact multiple groups, then group LASSO will encourage shrinkage between and not just within groups. To better maintain between group while reducing within group differences, one strategy is to truncate the penalty for large coefficient differences. Potentially, this could lessen the between group shrinkage, thus maintaining between group differences for better clustering or subpopulation identification. The TLP does exactly this by implementing a thresholding parameter, $\tau > 0$. For our two approaches with respect to TLP:

1. $P_{TLP}(\beta) := TLP(\beta) := \sum_{i<j} \min(\|\beta_{0i} - \beta_{0j}\|_1/\tau, 1) + \sum_{m=1}^{p} \sum_{i<j} \min(\|\beta_{mi} - \beta_{mj}\|_1/\tau, 1)$

2. $P_{gTLP}(\beta) := gTLP(\beta) := \sum_{i<j} \min(\|\binom{\beta_{0i}}{\beta_i} - \binom{\beta_{0j}}{\beta_j}\|_2/\tau, 1)$

In comparing the LASSO and TLP versions, there is no further penalty for differences greater than τ for the TLP version, but there is with LASSO. Overall, LASSO parameter estimates are known to be biased, and TLP corrects this adaptively by combining shrinkage and thresholding (Shen, Pan, and Zhu; 2012b).

Computation

Given λ and τ (TLP only), estimates using the nongroup penalties P_L and P_{TLP} were obtained from slight modifications of the `gflasso` and `ncTLF` functions in FGSG: Feature Grouping and Selection Over an Undirected Graph in Matlab engineered by Yang et al. (2013).

We develop an alternating direction method of multipliers (ADMM) to fit the models when using group penalties. The ADMM form introduces another variable, Z, reflecting how the objective function can be separated and subsequently solved in parallel. In ADMM the problem with respect to group LASSO (gLASSO) is stated as:

$$\begin{aligned} &\text{minimize} \;\; f(\beta) = (1/2)\|Y - X\beta\|_2^2 + \lambda P_{gL}(Z) \\ &\text{subject to} \;\; F\beta - Z = 0 \end{aligned}$$

where F is the linear transformation matrix comparing vectors of coefficients for all pairs of samples $(1 \le i < j \le n)$. That is, $F = \left[F_{1,2}^T, \;\; F_{1,3}^T, \;\; \cdots \;\; , F_{n-1,n}^T\right]^T$ where each $F_{i,j}$ is a $(p+1) \times n(p+1)$ matrix

$$F_{i,j} = \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \overset{\text{(i(p+1)-1)}^{th}}{\underset{\downarrow}{\text{column}}} \qquad \overset{\text{(j(p+1)-1)}^{th}}{\underset{\downarrow}{\text{column}}} \\ \begin{bmatrix} \cdots & 0 & 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & 0 & 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & 0 & 1 & 0 & \cdots & 0 & -1 & 0 & \cdots \end{bmatrix} \end{array}$$

The corresponding gLASSO objective function, derived as in the method of multipliers from an augmented Lagrangian with u the scaled dual variable, is

$$L_\rho(\beta, z, u) = (1/2)\|Y - X\beta\|_2^2 + \lambda P_{gL}(Z) + (\rho/2)\|F\beta - z + u\|_2^2.$$

Boyd et al. (2011) showed the ADMM algorithm then iterates three steps until converging to coefficient estimates:

1. $\beta^{(h+1)} = (X^TX + \rho F^TF)^{-1}\big(X^TY + \rho F^T(z^{(h)} - u^{(h)})\big)$
2. $z^{(h+1)} = \begin{bmatrix} S_{\lambda/\rho}\big(F_{1,2}\beta^{(h+1)} + u_{1,2}^{(h)}\big) \\ \vdots \\ S_{\lambda/\rho}\big(F_{n-1,n}\beta^{(h+1)} + u_{n-1,n}^{(h)}\big) \end{bmatrix}$
3. $u^{(h+1)} = u^{(h)} + F\beta^{(h+1)} - z^{(h+1)}$.

In the above, the notation "(h)" is for the h^{th} iteration. S is the vector *soft thresholding operator*: $S_\kappa(a) = \big(1 - \kappa/\|a\|_2\big)_+ a$, and a_+ is equal to the positive part of a. Remark how $S_\kappa(a)$ can shrink a whole vector to 0 if the coefficient vectors being compared are the same, which is in contrast to the individual soft thresholding used in $LASSO(\beta)$. Finally, u is partitioned corresponding to the pairwise differences in coefficient vectors; thus, $u_{i,j}$ represents the subvector of u corresponding to the comparison made with $F_{i,j}$. For our estimation we set ρ, the augmented Lagrangian parameter, equal to 1.

The group TLP (gTLP) penalty is not convex, an important distinction from gLASSO; therefore, we use a difference convex method to facilitate computation. First, define the objective function:

$$S(\beta) = (1/2)\left\|Y - X\beta\right\|_2^2 + \lambda \sum_{i<j} \min\left(\left\|\begin{pmatrix}\beta_{0i}\\ \beta_i\end{pmatrix} - \begin{pmatrix}\beta_{0j}\\ \beta_j\end{pmatrix}\right\|_2/\tau, 1\right).$$

Similar to Shen, Huang, and Pan (2012), $S(\beta)$ can be written as a difference of two convex functions $S_1(\beta) - S_2(\beta)$ with

$$S_1(\beta) = (1/2)\left\|Y - X\beta\right\|_2^2 + (\lambda/\tau)\sum_{i<j}\left\|\begin{pmatrix}\beta_{0i}\\ \beta_i\end{pmatrix} - \begin{pmatrix}\beta_{0j}\\ \beta_j\end{pmatrix}\right\|_2$$

$$S_2(\beta) = (\lambda/\tau)\sum_{i<j}\left(\left\|\begin{pmatrix}\beta_{0i}\\ \beta_i\end{pmatrix} - \begin{pmatrix}\beta_{0j}\\ \beta_j\end{pmatrix}\right\|_2 - \tau\right)_+.$$

As demonstrated by those authors, a sequence of upper approximations can be constructed iteratively by replacing $S_2(\beta)$ at the iteration $h+1$ by its piecewise affine minimization

$$S_2(\beta)^{(h)} = S_2(\hat{\beta}^{(h)}) + (\lambda/\tau)\sum_{i<j}\Big[I\Big(\Big\|\begin{pmatrix}\hat{\beta}_{0i}\\ \hat{\beta}_i\end{pmatrix}^{(h)} - \begin{pmatrix}\hat{\beta}_{0j}\\ \hat{\beta}_j\end{pmatrix}^{(h)}\Big\|_2 \geq \tau\Big)\times$$

$$\Big(\Big\|\begin{pmatrix}\beta_{0i}\\ \beta_i\end{pmatrix} - \begin{pmatrix}\beta_{0j}\\ \beta_j\end{pmatrix}\Big\|_2 - \Big\|\begin{pmatrix}\hat{\beta}_{0i}\\ \hat{\beta}_i\end{pmatrix}^{(h)} - \begin{pmatrix}\hat{\beta}_{0j}\\ \hat{\beta}_j\end{pmatrix}^{(h)}\Big\|_2\Big)\Big]$$

at iteration h, yielding an upper convex approximation for $S(\beta)$ at iteration $h+1$:

$$S^{(h+1)}(\beta) = (1/2)\Big\|Y - X\beta\Big\|_2^2 + (\lambda/\tau)\sum_{i<j}\Big(\Big\|\begin{pmatrix}\beta_{0i}\\ \beta_i\end{pmatrix} - \begin{pmatrix}\beta_{0j}\\ \beta_j\end{pmatrix}\Big\|_2\Big) I\Big(\Big\|\begin{pmatrix}\hat{\beta}_{0i}\\ \hat{\beta}_i\end{pmatrix}^{(h)} - \begin{pmatrix}\hat{\beta}_{0j}\\ \hat{\beta}_j\end{pmatrix}^{(h)}\Big\|_2 < \tau\Big).$$

Because of this we can use ADMM for gTLP by replacing step two of the gLASSO algorithm with

$$z^{(h+1)} = \begin{bmatrix} S_{\lambda_h/\rho}\big(F_{1,2}\beta^{(h+1)} + u_{1,2}^{(h)}\big) \\ \vdots \\ S_{\lambda_h/\rho}\big(F_{n-1,n}\beta^{(h+1)} + u_{n-1,n}^{(h)}\big)\end{bmatrix}$$

where $\lambda_{h/\rho} = \lambda(\rho\tau)^{-1}\,\mathrm{I}\Big(\Big\|\begin{pmatrix}\hat{\beta}_{0i}\\ \hat{\beta}_i\end{pmatrix}^{(h)} - \begin{pmatrix}\hat{\beta}_{0j}\\ \hat{\beta}_j\end{pmatrix}^{(h)}\Big\|_2 < \tau\Big)$ is calculated for each comparison i < j.

Our method is distinct from the competing FMR estimation methods, which are intended to find estimates at the component level. In particular, it is semiparametric in form, as no specific parametric distribution for the errors is assumed. The choice to use the squared loss function was made to align with ordinary linear regression, essentially to use the common form of loss given the linear components of our model. A different choice for the loss function could influence the error structure when performing computation, presenting another opportunity for future work to improve gTLP by better pairing loss functions to problem structures. Therefore, for comparison we present results from application of the semiparametric FMR methodology of Hunter and Young (2012), which estimates β_{0k} and β_k for $k = 1, ..., K$ (refer to equation (2.1)); that is, an estimate of β_0 and β for each *component* k. The semiparametric models were fit with the default settings of the `spregmix` function in the `R` package `mixtools` from Benaglia et al. (2009b).

For both penalty types models were fit with a large decreasing sequence of λ in order to show a wide range of degree of selection. When fitting models for a data set we started with the largest value of penalty. The resulting parameter estimates were used to initialize the subsequent model's estimation for the same data set (the model fit using the next smallest candidate in the sequence). We repeated this process until the fitting of the model with

the smallest λ was initialized with the estimates found with the second smallest λ. For the TLP models we considered a range of small to large candidates for the tuning parameter τ's in order to show results from situations where nearly all differences exceeded the threshold to situations with performance similar to LASSO. Last, a necessary question to resolve is whether the method can identify true differences in regression models and by proxy identify true subgroups. Ma and Huang (2017) provided a detailed exploration of the theoretical properties of their method that also applies to ours; consequently, we can apply their findings to gTLP.

To provide the reader with an overview for completeness in the current setting, Ma and Huang (2017) developed identification theorems from three conditions that are commonly met (or reasonably assumed true) in their penalized framework. The theorems specify the probability of recovery of the true group coefficients for K groups within a quantifiably small distance. Our methodology and computational algorithm fit Ma and Huang's described framework; consequently, we can apply their conclusions to gTLP. The three conditions, using Ma and Huang's original notation, are as follows:

1. The minimum eigenvalue of $[(Z,X)^T(Z,X)] \geq C_1|G_{min}|$ and $||X||_\infty \leq C_2 p$ where $i \in G_k$ represents membership in group k for sample i, $Z = \{z_{ik}\}$ is the $n \times K$ matrix with $z_{ik} = 1$ for $i \in G_k$ and 0 otherwise, and C_1 and C_2 are positive finite constants,
2. $P(\beta)$ is a symmetric function that is non-decreasing and concave for non-negative β, and $\rho(\beta) = \lambda^{-1}P(\beta)$ is constant for all $\beta \geq a\lambda$ for some constant $a > 0$ with $\rho(0) = 0$, in addition $\rho^{'}(\beta)$ exists and is continuous except for a finite number of β and $\rho^{'}(0+) = 1$, and
3. The noise vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ has sub-Gaussian tails such that $P(|a^T\epsilon| > ||a||x) \leq 2exp(-c_1x^2)$ for any vector $a \in R^n$ and $x > 0$, where c_1 is a positive finite constant.

Condition 1 is weak and can be satisfied by a bounded X that is not nearly perfectly correlated with the intercept terms. Condition 2 is met as gTLP is akin to generalized LASSO. Discussed by Ma and Huang, condition 3 is a common working assumption in high-dimensional settings. Given these three conditions hold, in addition to $K = o(n)$, $p = o(n)$, and $|G_{min}| \gg \sqrt{(K+p)nlog(n)}$, Ma and Huang showed the coefficient estimates will be at most $c_1^{-1/2}C_1^{-1}\sqrt{K+p}|G_{min}|^{-1}\sqrt{nlog(n)}$ from the true values with probability at least $1 - 2(K+p)n^{-1}$. We would refer the reader to Ma and Huang's excellent work for more details; thus, enabling us to focus here on the question of how to best choose the penalization

parameters.

The threshold and penalty parameters used for the presented results were determined with generalized cross-validation (GCV). Golub, Heath, and Wahba (1979) showed GCV's viability in selecting the parameter in ridge regression, and Pan, Shen, and Liu (2013) used GCV successfully to choose the threshold parameter when applying their TLP based PRclust clustering algorithm. When calculating the GCV in our setting, first allow $\hat{\mu}_i = \hat{\beta}_{0i} + X_i\hat{\beta}_i$. Following Golub, Heath, and Wahba (1979) generalized cross-validation can be defined as

$$GCV(df) = \frac{RSS}{(n-df)^2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2}{(n-df)^2}.$$

Here, the notation shows how the GCV statistic is a function of df, equal to the degrees of freedom used when generating the μ_i. Pan, Shen, and Liu (2013) found estimates could be improved by using generalized degrees of freedom (GDF) instead of the usual $df = p$. Ye (1998) provided the calculation for GDF, which in our problem is

$$GDF = \sum_{i=1}^{n} \lim_{\delta \to 0} E_\mu \Big[\frac{\hat{\mu}_i(Y_i + \delta e_i) - \hat{\mu}_i(Y_i)}{\delta} \Big]$$

where e_i is the i^{th} column of the $n \times n$ identity matrix. Correspondingly, Ye (1998) provided the following Monte Carlo algorithm to estimate GDF (adapted to our setting) when applying one of our four penalties:

1. Repeat steps 2 and 3 for $b = 1, \ldots, B$. In the following we set $B = 100$
2. Generate $\Delta_b = (\delta_{b,1}, \ldots, \delta_{b,n})$ with $\delta_{b,i}$ iid $\mathcal{N}(0, \nu)$. For our problems $\nu \approx .5\sigma_Y$
3. Compute $\hat{\mu}(Y + \Delta_b)$ with the penalty-specific algorithm using data $Y + \Delta_b$
4. Calculate $\hat{h}_i$ as the regression slope from $\hat{\mu}_i(Y + \Delta_b) = \alpha + \hat{h}_i\delta_{b,i}$ for $b = 1, \ldots, B$
5. Use $GDF = \sum_{i=1}^{n} \hat{h}_i$ when calculating GCV for the $\hat{\beta}$ found with a specified λ and τ (TLP only).

The parameter values for the following results are those with the smallest $GCV(GDF)$ statistic among the candidates considered. Once the candidate (λ, τ) pair with the smallest $GCV(GDF)$ statistic is found, K can be calculated as the number of unique regression coefficient vectors, (β_{0i}, β_i^T), among the i samples.

3. Simulations

We initially explored multiple settings with increasingly less separation in a single continuous response generated from a standard linear regression model with one continuous covariate ($p = 1$) and an intercept for $n = 100$ or 200 subjects. The responses were generated from a FMR model with $K = 2$ components; that is, the responses were generated using different regression models for $k = 1$ and $k = 2$. The settings were chosen to first verify the method's capabilities in a unambiguous scenario and second provide insight into data features where our new gTLP method would improve on classic semiparametric approaches. The choice to simulate with a single covariate was deliberate. In isolating a single covariate while varying its effect's size and direction by subgroup, the simulations can provide clearer evidence of scenarios ideal for gTLP simply because alternative sources of sample clustering have been minimized or eliminated. In this manner our first examinations established a foundation for gTLP, a foundation with an embedded flexibility that allows it to be extended naturally into more complex settings. $K = 2$ was chosen for the same reason. Following the initial phase of simulation, we built the single covariate $K = 2$ simulations into a single covariate three subgroup simulation that permitted us to test our conclusions about gTLP in a more challenging setting.

3.1 Simulation Design

The component when $K = 2$ for sample i was simulated from a Bernoulli distribution with mean equal to 0.5; that is, equal probability of either component generating the true response. Resulting from the use of the Bernoulli distribution to randomly assign group, The subjects' responses were created with each component. The simulated response was generated as

$$Y_i|X_i, k = \beta_{0k} + X_i\beta_{1k} + \epsilon_i, \tag{3.1}$$

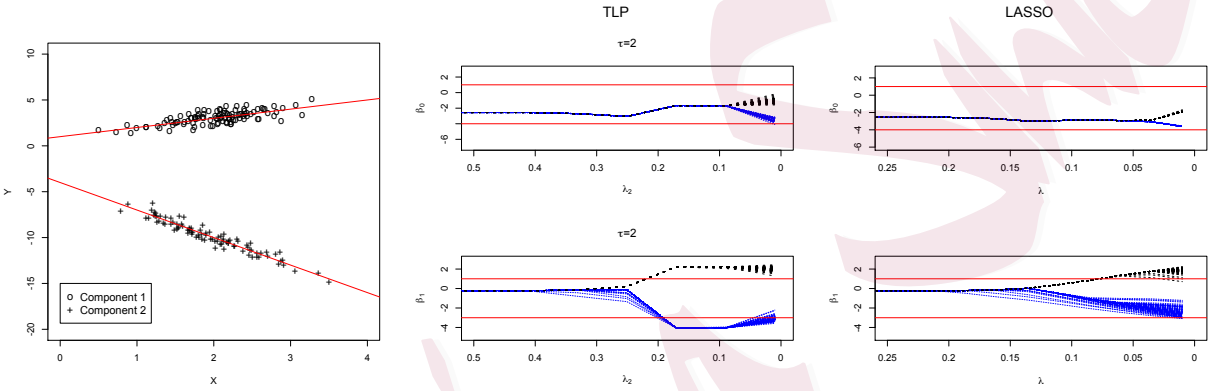
where $k \in \{1, 2\}$ indicates the component generating Y_i and $(\beta_{0k}, \beta_{1k})^T$ are the intercept and regression coefficient for the k^{th} regression component.

The first stage of the simulation is the generation of the covariate value. Let X_i represent a continuous covariate. Specifically, X_i is generated from a normal distribution with mean 2 and standard deviation 0.5. In the following we describe simulations where we considered two different $(\beta_{01}, \beta_{11})^T$ and $(\beta_{02}, \beta_{12})^T$ combinations and generated Y_i from the respective regression components using equation (3.1). A natural inquiry relates to how our method handles different error structures; consequently, we utilized randomly sampled ϵ_i from varied distributions.

3.2 Simulation Results

The first simulation evaluates a scenario with strong separation between responses gener-

ated from different components in order to demonstrate gTLP's efficacy on a simple problem that could be easily modified to add challenge. Set $\beta_{01} = 1$ and $\beta_{11} = 1$ for component one and $\beta_{02} = -4$ and $\beta_{12} = -3$ for component two. Errors were generated from the symmetric normal distribution ($\mathcal{N}(0, 0.5)$). The (X_i, Y_i) pairs are plotted in Figure 3.1(a). Subjects from the first component are plotted with circles and subjects from the second component are plotted with pluses. Additionally, the true regression lines for the two components are plotted with solid lines.



(a) (b) (c)

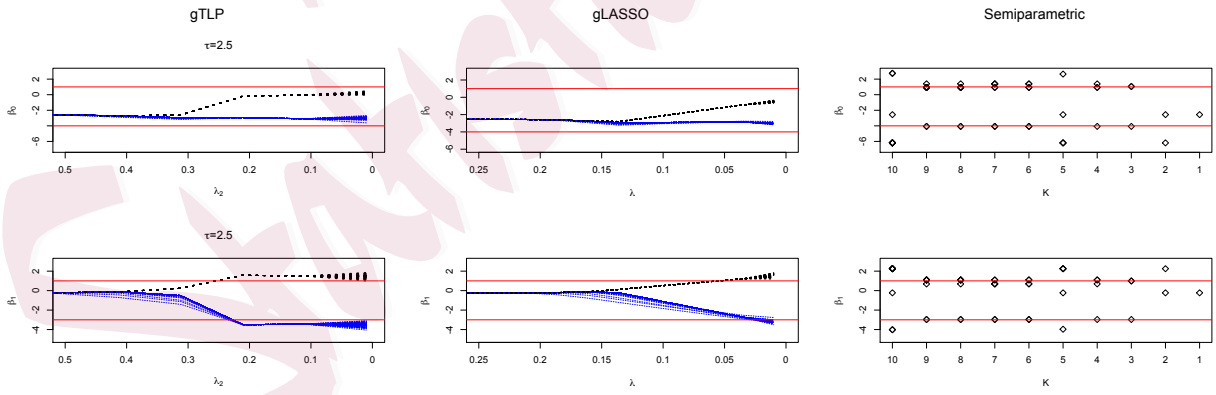
Figure 3.1: (a) Y_i and X_i scatterplot with true regression lines, and β_0 (row 1) and β_1 (row 2) estimates using (b) TLP and (c) LASSO

The results in Figure 3.1 examine the performance of penalized regression with the nongroup penalties: TLP (b) and LASSO (c). The individual λ regularization paths for each subject i are plotted for β_{0i} (top row) and β_{1i} (bottom row). In our usage, a regularization path is the curve connecting the estimates obtained for person i when using each value of λ (horizontal axis) in decreasing sequential order. From left to right the value of the penalty parameter is decreasing to allow any natural hierarchical structure to be exhibited. For TLP the plot is based on $\tau = 2$, the value with the lowest combined GCV statistics across the candidate penalty parameters. True coefficient values are given as horizontal lines, and the regularization paths for subjects from the first component are darker than those from the second.

Subjects from the two components can be distinguished for both TLP and LASSO given small enough λ. Not unexpectedly, the divergence in parameter estimates for subjects in the same component generally increases with both the TLP and LASSO methods as the

penalty decreases. This becomes significant because the λ at which the groups separate is different for the β_{0i}'s and the β_{1i}'s. TLP does outperform LASSO in terms of providing closer estimates of the true β_i as λ decreases, but there is still no λ range for either method at which both components' β_0 or β_1 estimates are simultaneously within even one unit for all n subjects (using a course metric for illustrative purposes). These two deficiencies prompted an investigation of the effect of a group penalty applied to the distance between the samples' coefficient vectors.

Figure 3.2(a) reveals the success of our group TLP (gTLP) method at overcoming these issues. The individual λ regularization path for each sample i are plotted for $\tau = 2.5$ (lowest total λ path GCV). As before the hierarchical structure can be seen in both the β_{0i} and β_{1i} plots, where the two distinct groups become more apparent as the penalty is reduced. The key observation is how the estimates themselves show increased β_0 and β_1 accuracy for both components simultaneously unlike in the TLP or LASSO versions (closer to the true values for small λ). gTLP definitely exhibits this property more than that of the group LASSO plots in Figure 3.2(b). We see the gLASSO is effective at identifying two distinct components, but shows less accuracy (distance between the true and estimated values) than the gTLP approach in at least one parameter. Comparing the group and nongroup approaches, the largest penalty parameter value which induces separation between components is the same for both the slope and coefficient.



(a) (b) (c)

Figure 3.2: β_0 (row 1) and β_1 (row 2) estimates using (a) gTLP, (b) gLASSO, and (c) SP

Semiparametric (abbreviated SP) FMR models were fit with $K = 1, ..., 10$ specified

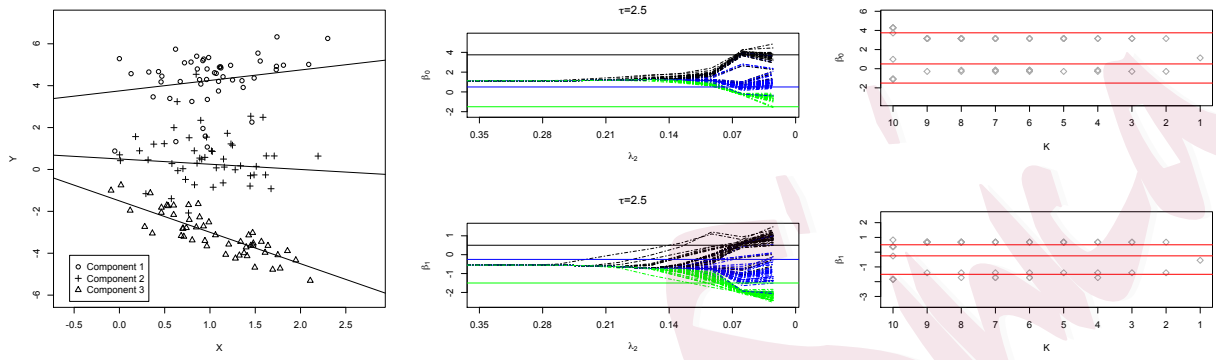
components (in descending order on the x-axis), and the parameter estimates are plotted in the third panel of the figure. Figure 3.2(c) reports β_{0k} (top row) and β_{1k} (bottom row) for the $k = 1, ..., K$ components. The figures reveal that for $K = 2$, the true component number, SP estimation is overall not successful, seeming to provide estimates centered around one of the two true component parameter values for both β_0 and β_1. Please note that as the default `spregmix` function incorporates a random initialization, we confirmed the results of 3.2(c) by repeating the process 25 times, and each showed the same result.

The first simulation yielded evidence that gTLP can outperform the other methods and modeling approaches on an admittedly simple problems, but it also provided better results with fewer assumptions. Our second simulation considered the next logical complication, that of partially overlapping responses for both components. The second simulation built in two additional complications: (1) we added weight to the tails of our previous symmetric normal distribution, and (2) we added a skewed distribution for one of the component's errors. As the structure and conclusions of the second scenario were built into a third even more challenging simulation scenario, the full discussion of the second simulation is presented as a supplement to this manuscript.

A possible insight from the early simulations is that gTLP and to a lesser degree gLASSO are best when responses do not display a large degree of overlap. The additional thresholding parameter when using TLP may be advantageous when the distance between component coefficient vectors is dominated by one parameter. Similarly, it may be valuable to truncate penalization in order to reduce the effect of penalizing samples that are truly in different subpopulations.

Last, we believed it important to further test our conclusions about gTLP's strengths, especially in a more challenging problem; thus, we created a simulation scenario with $K = 3$ subgroups using elements of each of our previous scenarios. To be consistent with earlier scenarios, we again let X_i represent a continuous covariate, generating it from a normal distribution with mean 1 and standard deviation 0.5 and fixed $\beta_{01} = 3.75$ and $\beta_{11} = \frac{1}{2}$ for component one, $\beta_{02} = \frac{1}{2}$ and $\beta_{12} = \frac{-1}{4}$ for component two, and $\beta_{03} = -1.5$ and $\beta_{13} = -1.5$ for component three. Next, we let $\epsilon_{i1} \sim \ln\mathcal{N}(0, 1.25)$, $\epsilon_{i2} \sim \mathrm{t}_{10}$, and $\epsilon_{i3} \sim \mathcal{N}(0, .5)$. We used a sample size $n = 150$ with probability equal to $\frac{1}{3}$ for each of the three components. Please observe that our three group simulation is (1) utilizing a more complicated mixture of all the distribution families from out first two simulations, (2) is continuing to use regression coefficients that vary is magnitude and direction from 0, and (3) has two largely separated components (key feature of simulation 1) that are essentially connected by a third component that overlaps both to either a small or moderate degree (key feature of simulation 2). The

effect is best viewed in Figures 3.3(a).



(a) (b) (c)

Figure 3.3: (a) Y_i and X_i scatterplot with true regression lines, and β_0 (row 1) and β_1 (row 2) estimates using (b) gTLP and (c) SP. Note: the samples from components 1 to 3 are in increasingly lighter colors in panel (b)

Comparing Figures 3.3(b) gTLP and (c) SP to their counterparts from earlier simulations and considering their findings, we would expect components one and three to be distinct when viewing gTLP results. The current simulation does go further than simulation 1 in that component one is generated from a log-normal distribution. As λ_2 decreases in value in Figure 3.3(b), we see that gTLP is not hindered by the skewed distribution as their is clear divergence in both regression coefficients for components one and three, observed in the divide of the top and bottom most sets of estimates. Noteworthy, gTLP appears to handle the second component, which overlapped the other two components, better in the $K = 3$ setting than in the previous $K = 2$ scenario. While the estimated regression coefficients show more variance within component two than the other components (middle cluster of estimates), they as a group have noticeable separation from the other two components. Although not unexpected, a trade-off for the improved component identification with gTLP in the $K = 3$ scenario is increased bias in the actual coefficient estimates. This was interestingly only for the intercept (β_{03}) of the normally distributed component, but not the other two components' β_{0k}. For the β_{1k} it was similarly comment-worthy how the least bias occurred in estimates of the log-normal component.

Extending our gained insights about gTLP, the $K = 3$ simulation at a minimum confirms gTLP's value in distinguishing samples from mostly non-overlapping components. In

addition, the $K = 3$ simulation demonstrates that gTLP can do this for a mixture of symmetric and skewed distributions. The $K = 3$ simulation likewise reaffirms, at a minimum, how gTLP can potentially extract a clearer partition of components, via their estimated regression coefficients, when outcomes overlap enough to be too problematic for classic approaches like the semiparametric one shown in Figure 3.3(c). Shown in the figure, the SP method essentially identifies two groups even when explicitly given $K = 3$. Overall, gTLP's performance in the $K = 3$ scenario exceeded its performance in the individual $K = 2$ scenarios. gTLP's success with increased numbers of components in a complicated scenario where several of the components had responses that were not clearly differentiated or were overlapping to some degree is a promising finding to be leveraged in continuing work.

4. Real Data

The final applied section shows a real data example to test the conclusions we drew from the simulated data sets. Namely, we sought to test our conclusion that gTLP would be promising in a scenario where differences in a continuous factor's effect were consistent enough by group to cause some degree of clustering of responses. Gene expression data is a natural setting as one goal of research in this area is to find creative ways to quantify the impact of differentially expressed genes. The expression levels of single genes provide logical factors for exploration with gTLP that parallel our simulation set-ups.

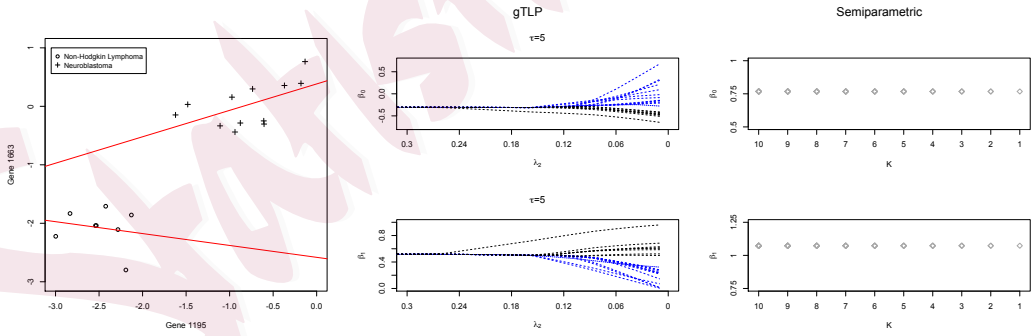
4.1 Small, Round Blue Cell Tumor Data

Khan et al. (2001) explored the ability to train artificial neural networks to use gene expression data from cDNA microarrays to classify types of small, round blue cell cancerous tumors (SRBCTs) in children (Khan, et al., 2001). The data was made available in the `CMA` R package (Slawski, Boulesteix, and Bernau, 2009). The resource provided expression data for 2308 genes meeting the authors' quality control standards. The expression data was from 63 subjects with one of four specified classes of SRBCTs: neuroblastoma ($n = 12$), rhabdomyosarcoma ($n = 20$), non-Hodgkin lymphoma ($n = 8$), and Ewing family ($n = 23$). Optimal treatment differs by type, but diagnosis by traditional clinical methods is difficult per the authors. Please observe how the Khan, et al. data allows us to explore a likely expectation of researchers when using semi- or nonparametric statistical tools; that of applicability to small sample sizes.

Khan et al. used artificial neural network (ANN) models incorporating 96 of the genes to classify the cancer set. The 96 genes represented the parsimonious subset of the 2308 genes that minimized their classification error rates. In order to agnostically limit our candidate set, we restricted our analysis to genes with a statistically significant ($\alpha = 0.05$) difference in

expression level between the four types of tumors as determined by a global F-test on means. Three genes met this threshold, labeled 154, 1195, and 1663 in the data. To this point our intent was to identify an even smaller subset of the candidates genes in an objective and commonplace way, and then we would assess gTLP's performance with respect to the same goal of Khan et al. To select the final genes and cancer subtypes for our assessment, we chose the two genes and two cancer types with a relationship between the expression data that best matched our simulation-derived conclusions about where gTLP would be effective. Of note, our final two subtypes (non-Hodgkin lymphoma and neuroblastoma) had the highest sensitivity for both the original and test cases using the ANN approach (although all cancer subtypes were classified correctly to a very high degree). We thought this an important supporting detail as the true type of cancer was only assessed to a high degree of confidence with classic approaches; thus, the credibility of our conclusions about correct classification is further enhanced by ANN-based confirmation of the true cancer subtype.

Figure 4.1(a) shows that our data has a relationship with elements of our first two simulations (but with much smaller samples sizes). However, after adding individual subtype fitted lines to (a) we see the separation does not result from components defined strongly by a linear association. The Neuronblastoma subtype exhibits a relationship closer to those used in simulation one, but the non-Hodgkin lymphoma samples show noticeably less of this. Both subtypes exhibit skew in the gene 1663 distributions, and it is possible that this is driven by an outlier for the non-Hodgkin lymphoma samples.



(a) (b) (c)

Figure 4.1: Khan et al. (2001) expression data presented as (a) scatterplot of gene 1195 and gene 1663, and β_0 (row 1) and β_1 (row 2) estimates using (b) gTLP and (c) SP

The gTLP performance is presented in Figure 4.1(b), and the semiparmetric perfor-

mance is shown in panel (c). SP is not able to distinguish between cancer subtypes. However, the gTLP appears to be able to successfully partition the cancer types. The dendrogram-like clustering is not present, but simply partitioning by positive or negative trend in the coefficient for gene 1195 as λ_2 decreases provides perfect classification. We want to fully acknowledge that we have a user driven setting within this real data example. However, it is perhaps even more remark-worthy that we have replicated a classification based on 96 genes by applying gTLP to the relationship between two genes found among 1000's using simple ANOVA. Last, Figure 4.1(b) shows an unexpected product of gTLP. Observe that the darker lines in part (b) of this figure correspond to the non-Hodgkin lymphoma samples. The estimated regression coefficients for this cancer subtype more closely match a model fit without the one likely outlier visible in 4.1(a). Consequently, gTLP might additionally provide a robustness in its estimation process.

5. Discussion

The article has provided evidence using real data supported by simulation that our new grouping pursuit gTLP method, and to a lesser extent a grouping pursuit gLASSO, handles certain types of problems for which previous methods such as Hunter and Young's semiparametric approach were not successful. Our novel gTLP approach was successful in scenarios using FMR when responses generated by different component regression models were at least generally clustered, but not necessarily distinct, in one-dimension. The gTLP method, which applies group penalization to differences between coefficient vectors, was able to correctly classify subpopulations in our applied gene expression data example. The gTLP method also returned reasonable to very good estimates of the known regression coefficients in simulations without strong overlap in subgroup responses. While warranting further investigation, the truncation threshold parameter (τ) used by the gTLP improved on gLASSO methods, likely because of its weighting of the penalty more towards within component differences. If the responses from different component regression models are well separated or do not exhibit a large degree of overlap, the gTLP may be better than gLASSO at maintaining between component/subpopulation separation in the coefficients while reducing within component differences. In addition, this work confirms that group penalties, such as gTLP and gLASSO, can improve component identification and regression model estimation over their corresponding coefficient specific penalties, TLP and LASSO.

Importantly, our new method focuses on the estimation (and then clustering) of individual regression models. This holds great promise for application to personalized medicine. In the present work we have only begun to show how a different grouping approach to penalized regression may be able to overcome some of the limitations of current approaches.

The simulations were basic and do not cover a large range of possible combinations of component models, but they do provide minimally ambiguous support for gTLP's value in the essential setting (single variable) needed for analysis of more complicated scenarios. Future work will need to apply the method to more scenarios to further refine the class of problems for which gTLP shows strong promise. The work herein included one capstone simulation with three subgroups, offering further and in some ways stronger evidence for our conclusions about gTLP's advantages. Even more advantageous to future work, it built off of the basic scenarios, meaning that our final simulation likewise revealed how the foundation built in this paper can be effectively leveraged and adapted to future researcher needs. For example, a particular problem of interest occurs when a variant has a true effect for only one of several subsets of the population. Also, future work must include scenarios with more covariates in order to apply to the very likely scenario of health or diseases outcomes resulting from complex functions of multiple variables. Increasing the number of considered covariates will also enable exploration of variable selection features in addition to grouping features. Similarly, establishing gTLP's effectiveness with large samples will be paramount considering the growth in availability of data with multiple thousands of samples. However, these extensions represent an involved second phase of research that requires that one first establish a comprehensive and credible foundation for gTLP, which was the motivation for and aim of the current work. The authors thought it valuable to show how the penalty magnitude could uncover a hierarchal structure; thereby, showing the potential for different partitions of the population. The work to date employed the squared loss function only, but the method could be modified to accommodate different loss functions that might better serve a problem. For example, it could be interesting to look at an L_1 function in data with outliers, especially considering the robustness of gTLP found in the gene expression data example. Finally, we showed that GCV could be used to choose a single set of coefficient estimates among those generated by different threshold and penalty values, but it will be beneficial to revisit this issue and potentially develop a better criterion for selecting optimal tuning parameters and by extension discover the number of components (if indeed they exist). This GCV-based standard in the present article is possibly the most conservative standard for assigning samples to the same group; consequently, there is great promise in advancing our method to consider a more probabilistic-based approach. Our main goal here is to demonstrate the feasibility and promise of our proposed penalized regression approach as a proof of concept; however, the results herein go further and document the early successes that gTLP has had as a hierarchical clustering tool that can uncover a subpopulation structure in a data set.

6. Supplementary Materials

The full detail for the second simulation example referenced in Section 3.2 is available in the supplementary materials.

Acknowledgments

This research was supported by NIH grants R01HL65462, R01HL105397 and R01GM081535.

References

Benaglia, T., Chauveau, D., and Hunter, D.R. (2009). An EM-Like algorithm for semi- and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics* **18(2),** 505–526.

Benaglia, T., Chauveau, D., Hunter, D., and Young D. (2009b). `mixtools`: An R package for analyzing finite mixture models. *Journal of Statistical Software* **32(6),** 1–29.

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

Chi, E.C. and Lange, K. (2015). Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics* **24(4),** 994–1013.

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

DeSarbo, W.S. and Cron, W.L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification* **5(2),** 249–282.

Gaffney, S., and Smyth, P. (2003). Curve clustering with random effects regression mixtures. In *Proceedings of the ninth international workshop on artificial intelligence and statistics (AISTATS)*. Key West, FL.

Golub, G.H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21(2),** 215–223.

Hunter, D.R. and Young, D.S. (2012). Semiparametric mixtures of regressions. *Journal of Nonparametric Statistics* **24(1),** 19–38.

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* **102(479),** 1025–1038.

Khalili, A., Chen, J., and Lin, S. (2010). Feature selection in finite mixture of sparse normal linear models in high-dimensional feature space. *Biostatistics* **12(1),** 156–172.

Khan, J., Wei, J.S., Ringnér, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., and Meltzer, P.S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* **7(6),** 673–679.

Levine, M., Hunter, D.R., and Chauveau, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika* **98(2),** 403–416.

Luo, R., Wang, H. and Tsai, C.L. (2008). On mixture regression shrinkage and selection via the MR-LASSO. *International Journal of Pure and Applied Mathematics* **46,** 403–414.

Ma, S. and Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association* **112(517),** 410–423.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. New York: John Wiley & Sons, Inc.

Pan, W., Shen, X., and Liu, B. (2013). Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty. *The Journal of Machine Learning Research* **14(1),** 1865–1889.

Shen, X., Huang, H.C., and Pan, W. (2012). Simultaneous supervised clustering and feature selection over a graph. *Biometrika* **99(4),** 899–914.

Shen, X., Pan, W., and Zhu, Y. (2012b). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107(497),** 223–232.

Slawski, M., Boulesteix, A-L., and Bernau, C. (2009). `CMA`: Synthesis of microarray-based classification.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9(12),** 3273–3297.

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

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

Wedel, M. and DeSarbo, W.S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification* **12(1),** 21–55.

Wu, C., Kwon, S., Shen, X., and Pan, W. (2016). A new algorithm and theory for penalized regression-based clustering. *Journal of Machine Learning Research* **17(188),** 1–25.

Yang, S., Yuan, L., Lai, Y.C., Shen, X., Wonka, P., and Ye, J. (2013). Feature grouping and selection over an undirected graph. In *Graph Embedding for Pattern Analysis*, 27–43. Springer, New York.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* **93(441),** 120–131.

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

Erin Austin (Corresponding Author); Department of Mathematical and Statistical Sciences, University of Colorado Denver, 80204

E-mail: erin.e.austin@ucdenver.edu

Telephone: (303) 315-1744

Wei Pan; Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455

E-mail: weip@biostat.umn.edu

Xiaotong Shen; School of Statistics, University of Minnesota, Minneapolis, MN 55455

E-mail: xshen@umn.edu