Statistica Sinica **30** (2020), 783-807
doi:https://doi.org/10.5705/ss.202016.0531

A NEW SEMIPARAMETRIC APPROACH TO FINITE MIXTURE OF REGRESSIONS USING PENALIZED REGRESSION VIA FUSION

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

[1] *University of Colorado Denver and* [2] *University of Minnesota, Minneapolis*

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, in which 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 the proposed approach better identifies subpopulations and their corresponding models than a semiparametric FMR method does. Our new method fits models for each person via grouping pursuit, utilizing a new group-truncated L_1 penalty that shrinks the differences between estimated parameter vectors. The methodology causes the individuals' models 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 analyses. Simulations using FMR models and a real-data analysis show that the method performs promisingly well.

Key words and phrases: FMR, group LASSO, group TLP, grouping pursuit, penalized regression, semiparametric.

1. Introduction

A traditional way of assessing 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 do not 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 framework 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 based on the likelihood assumptions used when fitting the model. Wedel and Desarbo (1995) showed that the algorithm successfully estimates the regression parameters for mixtures of common distributions, such as normal and binomial distributions. An EM-like algorithm was developed by Benaglia, Chauveau and Hunter (2009) to allow for more generality in the error term. However, although it lowers the 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 developed by Levine, Hunter and Chauveau (2011) to remedy the Benaglia shortcomings. However, the algorithm's advantages 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 in which it is possible to (1) specify the mixture distribution and its corresponding number of components and (2) initialize the algorithm.

The EM algorithm has also served as the main statistical tool for another category of approaches to subpopulation estimation, namely, clustering subject-specific regression models. In an early work, Desarbo and Cron (1988) used the EM algorithm for a 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 component assumptions, their work

demonstrates the potential of the clustering of subject-specific models.

In settings in which 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. A penalized regression as part of an FMR model estimation has shown promise as one such improvement. Specific to the goal of variable selection, some researchers have incorporated a penalized regression within their models. An effective EM algorithm developed by Khalili and Chen (2007) for a penalized mixture model was applied to an FMR setting for the purpose of variable selection. However, the estimation was based on a parametric likelihood assumption. Khalili, Chen and Lin (2010) developed an EM approach using a penalized likelihood for variable selection. The approach was effective in simulations at selecting important covariates, but this was after applying a screening method. To the best of our knowledge, the most successful approaches to date for estimating FMR models depend on a methodology based on some form or approximation of the EM algorithm, and thus depend on making successful likelihood assumptions or successful density estimations.

We take a novel approach to 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 simultaneously fit *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 to include a grouping feature to penalize the 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 regressions have been researched specifically to assess their 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 least absolute shrinkage and selection operator (gLASSO) penalty to select groups of variables or factors results

in better performance than when using the standard LASSO. Another penalized regression approach, the fused LASSO of Tibshirani et al. (2005), adds an additional penalty to the 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 incorporates a penalization of the differences in the component regression coefficients. The method, called MR-LASSO, demonstrated the ability to use a 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, using a truncated L_1 penalty (TLP) for grouping pursuit. This approach makes it possible to successfully identify and estimate unknown homogenous groups of effects. The method uses a single linear regression model for a single response, but assumes that the full coefficient vector can be partitioned into subsets of homogeneous coefficients. The new method improves parameter estimation and group identification by penalizing the differences within these smaller vectors. In a related work, Pan, Shen and Liu (2013) developed a penalized regression-based clustering (PRclust) method, in which they apply the TLP penalty to differences in the centroids of the data points. PRclust performs well in situations such as nonconvex clusters, where other, more common methods do not. Pivotal to the current work, the success of PRclust demonstrated the potential of comparisons across subjects with a grouping penalty. Subsequently, Chi and Lange (2015) demonstrated how the alternating direction method of multipliers (ADMM) can be effective when solving convex clustering problems 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 using the centroid difference TLP penalization.

The best-performing penalized regression-based strategies have forgone the explicit use of FMR. Instead, they compare either subject-level differences in vectors of numerical variables to aggregate into clusters, or 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 a 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 because their methodology is approximately one case within our more general framework, thus demonstrating the potential value of our approach. Ma and Huang emphasized the estimation of the intercept term in the subgroup analysis. In contrast, we emphasize a slope parameter in the mixture regression. That is, our work incorporates a grouping pursuit framework to shrink the differences between the full subject-specific models for problems similar to FMR. Our approach to the penalized regression uses grouping pursuit when simultaneously fitting *separate* models for *each* subject. Specifically, 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 of Shen, Pan and Zhu (2012), 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 the differences in the subjects' models parameter by parameter, and does not explicitly shrink the differences between the full models. Therefore, we next apply two group penalties, based on the LASSO and TLP, respectively, to the differences in the estimated parameter vectors for each pair of sample regression models. Our work extends the research introduced above and, in particular, that of Ma and Huang in a few critical ways. First, we allow the subgroups to be defined by differences in observable factors. We feel this is important because it accounts for the possibility that any set of variables might only affect a subset of the population. Second, we incorporate the nonconvex TLP penalty in our clustering approach. Previous works, such as that of Pan, Shen and Liu (2013), have demonstrated the advantages of the TLP over L_1 penalties for subject-level penalizations.

When applied, it is our hypothesis that we will see a hierarchical clustering of individual models that depends on the magnitude of the penalty and the 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 to which our method can be applied. The following discussion uses simulated FMR models to permit comparisons with previous methods, and is followed by application to

a relevant true genetics data setting. The intent of the following is to show that the proposed 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 single-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 describe the FMR model. Then, we present our penalized regression approach and its computation.

2.1. FMR model

To motivate and contrast with our new method, we briefly review the FMR 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 $i = 1, \ldots, n$. Let X_{ji} equal subject i's value for covariate $j = 1, \ldots, p$; therefore, $X_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$ 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 an 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 $K < \infty$ because it is a mixture of K densities (known as component densities). In this equation, the unknown parameters are $\Psi = (\beta_1, \beta_2, \ldots, \beta_K, \phi, \pi)$, where $\beta_k = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{pk})^T$, $\phi = (\phi_1, \phi_2, \ldots, \phi_K)^T$, and $\pi = (\pi_1, \pi_2, \ldots, \pi_{K-1})^T$, such that both $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$.

Parametric approaches that specify a parametric form of $f(\theta, \phi)$ and estimate $f(\hat{\theta}, \hat{\phi})$ 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 using 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. As described in the next section, our method starts with n overspecified sets of regression coefficients. Then, it 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 a 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 a penalized regression with a group feature to reveal subpopulations via the clustering of 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. For each subject i, assume 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 $\mathrm{E}(\epsilon_i) = 0$. Note how we initially allow for a sample-dependent (β_{0i}, β_i^T) for each subject, and at no time specify or estimate a density function for ϵ_i. Our method is semiparametric, because 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. However, we require asymptotic sub-Gaussian tails, as we explain shortly in a discussion on the conditions for identifiability.

Observe from our model how the covariates associated with an outcome would have nonzero values in β_i, but we do not assume the set of nonzero 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 when the same set of covariates affect multiple subpopulations, the magnitude and/or direction of the 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 (β_{0i}, β_i^T) 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).

Our method provides estimates for β_{0i} and β_i by minimizing

$$\left(\frac{1}{2}\right) \|Y - X\beta\|_2^2 + \lambda P(\beta),$$

$$\text{with } Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_1 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & 1 & X_2 & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & & \mathbf{0} & \dots & \mathbf{0} & 1 \; X_n \end{bmatrix} \text{ with } \mathbf{0}^T = \left.\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right\} p+1,$$

$$\text{and } \beta = \begin{bmatrix} \beta_{01} \\ \beta_1 \\ \beta_{02} \\ \beta_2 \\ \vdots \\ \beta_{0n} \\ \beta_n \end{bmatrix}.$$

The form of our objective function follows those of Wu et al. (2016) and Ma and Huang (2017), 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 nonconvex TLP (Shen, Pan and Zhu (2012)). 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 its selection on the between-sample differences in individual coefficient estimates. Depending on the size of λ, the nongroup version chooses the nonzero differences between the final estimated sample models by comparing corresponding parameters separately. In contrast, the group version, $P_{gL}(\beta)$, shrinks the differences between the full estimated parameter sets, and is more likely to have $(\beta_{0i}, \beta_i^T) = (\beta_{0j}, \beta_j^T)$.

The LASSO penalty shrinks all coefficient differences. However, if there are in fact multiple groups, then the gLASSO encourages 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 the 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. $P_{gTLP}(\beta) := gTLP(\beta) := \sum_{i<j} \min(\|\binom{\beta_{0i}}{\beta_i} - \binom{\beta_{0j}}{\beta_j}\|_2/\tau, 1)$.

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

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. (2012).

We develop an 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 the ADMM, the problem with respect to the gLASSO is stated as:

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

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

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

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

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

Boyd et al. (2011) showed that 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)" denotes the hth iteration. S is the vector of the *soft thresholding operator*: $S_\kappa(a) = \big(1 - \kappa/\|a\|_2\big)_+ a$, and a_+ is equal to the positive part of a. Note that $S_\kappa(a)$ can shrink a whole vector to zero if the coefficient vectors being compared are the same, 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 with $F_{i,j}$. For our estimation we set ρ, the augmented Lagrangian parameter, equal to one.

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

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

Similarly 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

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

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

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

$$S_2(\beta)^{(h)} = S_2(\hat{\beta}^{(h)}) + \left(\frac{\lambda}{\tau}\right)\sum_{i<j}\Big[I\Big(\left\|\binom{\hat{\beta}_{0i}}{\hat{\beta}_i}^{(h)} - \binom{\hat{\beta}_{0j}}{\hat{\beta}_j}^{(h)}\right\|_2 \geq \tau\Big)$$
$$\times\Big(\left\|\binom{\beta_{0i}}{\beta_i} - \binom{\beta_{0j}}{\beta_j}\right\|_2 - \left\|\binom{\hat{\beta}_{0i}}{\hat{\beta}_i}^{(h)} - \binom{\hat{\beta}_{0j}}{\hat{\beta}_j}^{(h)}\right\|_2\Big)\Big],$$

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

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

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

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

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

Our method is distinct from competing FMR estimation methods, which are intended to find estimates at the component level. In particular, our method is semiparametric in form, because 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 the loss, given the linear components of our model. A different choice for the loss function could influence the error structure when performing the computation, presenting another opportunity for future work to improve the gTLP by better pairing loss functions with problem structures. Therefore, for comparison, we present the results from applying the semiparametric FMR methodology of Hunter and Young (2012), which estimates β_{0k} and β_k for $k = 1, \ldots, K$ (refer to equation (2.1)), that is, an estimate of β_0 and β for each *component* k. The semiparametric models were fitted using the default settings of the `spregmix` function in the `R` package `mixtools` of Benaglia et al. (2009).

For both penalty types, models were fitted using 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 the penalty. The resulting parameter estimates were used to initialize the subsequent model's estimation for the same data set (the model fitted using the next smallest candidate in the sequence). We repeated this process until the model with the smallest λ was initialized using the estimates from the second smallest λ. For the TLP models, we considered a range of small to large candidates for the tuning parameter τ. As such, we show results ranging from situations where nearly all differences

exceed the threshold to situations with performance similar to that of the LASSO. Lastly, 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, which also applies to our method. Therefore, we can apply their findings to the gTLP.

To provide 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 the 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 nondecreasing and concave for nonnegative 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 $\rho'(0+) = 1$.

3. The noise vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ has sub-Gaussian tails, such that $P(|a^T\epsilon| > \|a\|x) \leq 2\exp(-c_1x^2)$, 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 because the gTLP is similar to the generalized LASSO. As discussed by Ma and Huang, condition 3 is a common working assumption in high-dimensional settings. Given that these three conditions hold, in addition to $K = o(n)$, $p = o(n)$, and $|G_{\min}| \gg \sqrt{(K+p)n\log(n)}$, Ma and Huang showed that the coefficient estimates will be at most $c_1^{-1/2}C_1^{-1}\sqrt{K+p}|G_{\min}|^{-1}\sqrt{n\log(n)}$ from the true values, with probability at least $1 - 2(K+p)n^{-1}$. Refer to Ma and Huang's excellent work for further details. Here, we focus on the question of how to best choose the penalization parameters.

The threshold and penalty parameters used for the results presented here were determined using generalized cross-validation (GCV). Golub, Heath and Wahba (1979) showed GCV's viability in selecting the parameter in a 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 μ_i. Pan, Shen and Liu (2013) found that their estimates could be improved by using the generalized degrees of freedom (GDF) instead of the usual $df = p$. Ye (1998) provided the calculation for the GDF, which in our problem is

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

where e_i is the ith column of the $n \times n$ identity matrix. Correspondingly, Ye (1998) provided the following Monte Carlo algorithm to estimate the 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 $\delta_{b,i}$ independent and identically distributed (i.i.d.) $\mathcal{N}(0, \nu)$. For our problems, $\nu \approx 0.5\sigma_Y$.
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 the GCV for $\hat{\beta}$ found using 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 an 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 an unambiguous scenario and, second, to provide insight into the data features, where our new gTLP method improves on the classic semiparametric approaches. The choice to simulate using 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 the gTLP, simply because alternative sources of sample clustering have been minimized or eliminated. In this manner, our first examinations established a foundation for the gTLP, with an embedded flexibility that allows it to be extended naturally to more complex settings. Here, $K = 2$ was chosen for the same reason. Following the initial phase of the simulation, we built the single-covariate $K = 2$ simulations into a single-covariate three-subgroup simulation, allowing us to test our conclusions about the 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, an equal probability of either component generating the true response. As a result of using the Bernoulli distribution to randomly assign groups, the subjects' responses were created using 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 that generates Y_i, and $(\beta_{0k}, \beta_{1k})^T$ denotes the intercept and regression coefficient for the kth regression component.

In the first stage of the simulation, we generate 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 in which 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, ϵ_i are randomly sampled from various distributions.

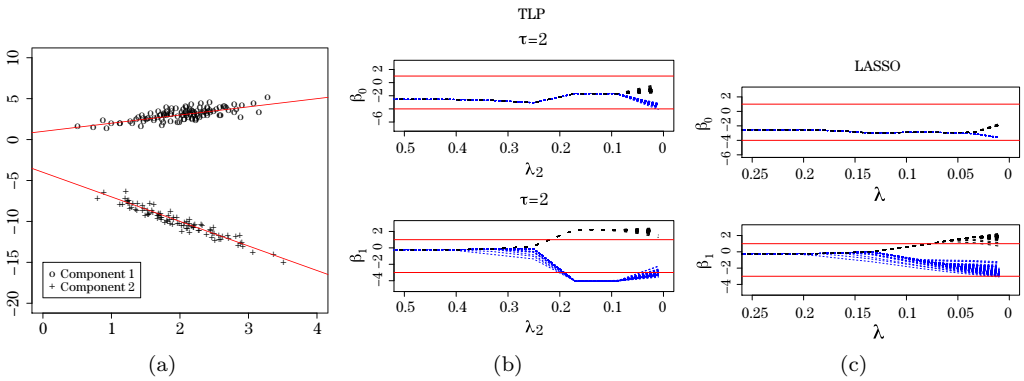

Figure 1. (a) Y_i and X_i scatterplot with true regression lines, and β_0 (row 1) and β_1 (row 2) estimates using (b) TLP and (c) LASSO.

3.2. Simulation results

The first simulation evaluates a scenario with strong separation between responses generated from different components; as such, it demonstrates the gTLP's efficacy on a simple problem that can easily be made more challenging. 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 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.

The results in Figure 1 show the performance of the penalized regression with 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. Note, the figures use the notation λ_2 on the horizontal axis for TLP-based plots, instead of λ, to make expressly clear that the TLP and gTLP methods are distinct from the LASSO-based methods. From left to right, the value of the penalty parameter is decreasing to allow any natural hierarchical structure to be exhibited. For the TLP, the plot is based on $\tau = 2$, the value with the lowest combined GCV statistics across the candidate penalty parameters. The 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.

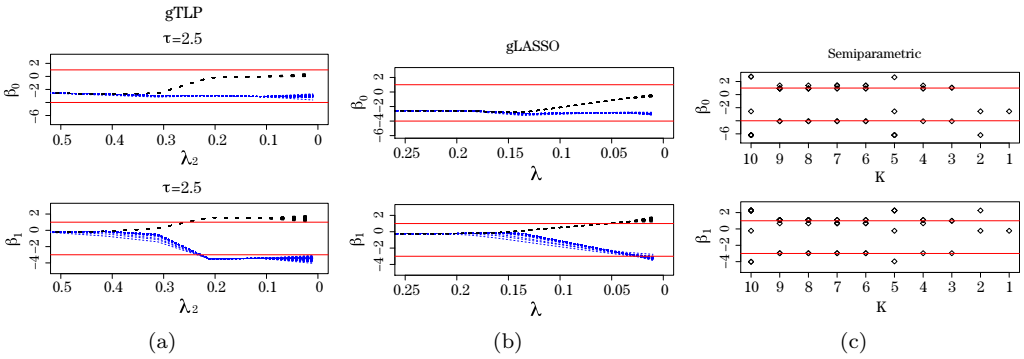

Figure 2. β_0 (row 1) and β_1 (row 2) estimates using (a) gTLP, (b) gLASSO, and (c) SP.

Subjects from the two components can be distinguished for both the TLP and the LASSO, given a sufficiently small λ. Not unexpectedly, the divergence in the parameter estimates for subjects in the same component increases, in general, with both the TLP and the LASSO methods as the penalty decreases. This becomes significant because the λ at which the groups separate is different for β_{0i} and β_{1i}. The TLP does outperform the LASSO in terms of providing closer estimates of the true β_i as λ decreases. However, there is still no range of λ 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 2(a) reveals the success of our gTLP method at overcoming these issues. The individual λ regularization path for each sample i is 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). The gTLP definitely exhibits this property more than the gLASSO plots in Figure 2(b) do. 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 that induces separation between components is the same for both the slope and the coefficient.

Semiparametric (abbreviated SP) FMR models were fitted using $K = 1, \ldots,$

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

The first simulation yielded evidence that the gTLP can outperform the other methods and modeling approaches on, an admittedly, simple problem, but it also provided better results with fewer assumptions. Our second simulation considers the next logical complication, that of partially overlapping responses for both components. This simulation includes 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. Because the structure and conclusions of the second scenario were incorporated into a third, even more challenging simulation scenario, a full discussion of the second simulation is presented as a supplement to this manuscript.

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

Lastly, we believed it important to further test our conclusions about the 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 $\beta_{11} = 1/2$ for component one, $\beta_{02} = 1/2$ and $\beta_{12} = -1/4$ 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 $\epsilon_{i3} \sim \mathcal{N}(0, 0.5)$. We used a sample size of $n = 150$, with probability equal to 1/3 for each of the three components. Note that our three-group simulation is (1) utilizing a more complicated mixture of the distribution families from the first two simulations, (2) is continuing to use regression coefficients that vary in

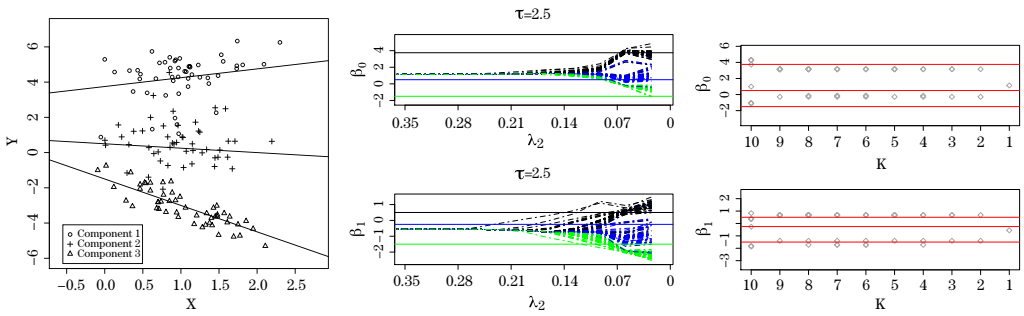

Figure 3. (a) Y_i and X_i scatterplot with true regression lines, and β_0 (row 1) and β_1 (row 2) estimates using (b) gTLP and (c) SP. Note: the samples from components 1 to 3 are shown in increasingly lighter colors in panel (b).

magnitude and direction from zero, 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 Figure 3(a).

Comparing Figures 3(b) gTLP and (c) SP to their counterparts from the earlier simulations and considering the findings, we would expect components one and three to be distinct when viewing the gTLP results. The current simulation does go further than simulation 1 in that component one is generated from a log-normal distribution. As λ decreases in value in Figure 3(b), we see that the gTLP is not hindered by the skewed distribution, because there is clear divergence in both regression coefficients for components one and three, observed in the divides of the top-most and bottom-most sets of estimates. Note that the 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 in the other components (middle cluster of estimates), as a group, they have noticeable separation from the other two components. Although not unexpected, a trade-off for the improved component identification with the gTLP in the $K = 3$ scenario is increased bias in the actual coefficient estimates. Interestingly, this occurred only for the intercept (β_{03}) of the normally distributed component, but not for the other two components' β_{0k}. For β_{1k}, the least bias occurred in estimates of the log-normal component.

Extending our insights about the gTLP, the $K = 3$ simulation, at a mini-

mum, confirms the gTLP's value in distinguishing samples from mostly nonoverlapping components. In addition, the $K = 3$ simulation demonstrates that the gTLP can do this for a mixture of symmetric and skewed distributions. Similarly, the $K = 3$ simulation reaffirms, at a minimum, how the gTLP can potentially extract a clearer partition of components, via their estimated regression coefficients, when the outcomes overlap enough to be too problematic for classic approaches, such as the semiparametric approach shown in Figure 3(c). The SP method essentially identifies two groups, even when explicitly given $K = 3$. Overall, the gTLP's performance in the $K = 3$ scenario exceeded its performance in the individual $K = 2$ scenarios. The 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 drawn from the simulated data sets. That is, we test our conclusion that the gTLP would be promising in a scenario in which the differences in a continuous factor's effect are consistent enough by group to cause some degree of clustering of responses. Gene expression data provide a natural setting, because such research tries to find creative ways to quantify the impact of differentially expressed genes. The expression levels of single genes provide logical factors with to explore the gTLP that parallel those of our simulations.

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 were made available in the `CMA` R package (Slawski, Boulesteix and Bernau (2009)), providing expression data for 2,308 genes that met the authors' quality control standards. The expression data were 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 ($n = 23$). The optimal treatment differs by type, but diagnoses using traditional clinical methods are difficult, per the authors. Note how the Khan et al. (2001) data allow us to explore a likely expectation of researchers when using semiparametric or

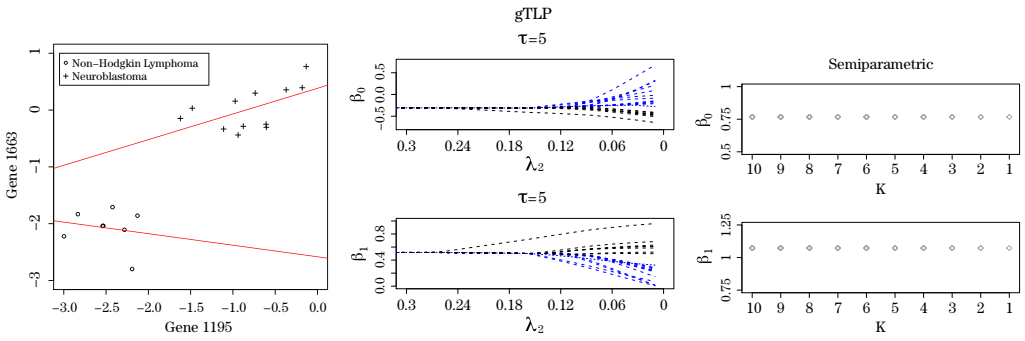

Figure 4. Khan et al. (2001) expression data presented as (a) a scatterplot of gene 1195 and gene 1663, and β_0 (row 1) and β_1 (row 2) estimates using (b) gTLP and (c) SP.

nonparametric statistical tools, namely, their applicability to small sample sizes.

Khan et al. (2001) used artificial neural network (ANN) models incorporating 96 of the genes to classify the cancer set. The 96 genes represented a parsimonious subset of the 2,308 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 the 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 candidate genes in an objective and commonplace way, and to then assess the gTLP's performance with respect to the same goal of Khan et al. (2001) 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 the gTLP would be effective. Our final two subtypes (non-Hodgkin lymphoma and neuroblastoma) had the highest sensitivity for both the original and the 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, because the true type of cancer was only assessed to a high degree of confidence using traditional diagnostic approaches; thus, the credibility of our conclusions about correct classification is further enhanced by the ANN-based confirmation of the true cancer subtype.

Figure 4(a) shows that our data are related to elements of the first two simulations (but with much smaller samples sizes). However, after adding in-

dividual subtype fitted lines to (a), we find that the separation does not result from the components defined strongly by a linear association. The neuroblastoma 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 skewness in the gene 1663 distributions, although it is possible that this is driven by an outlier for the non-Hodgkin lymphoma samples.

The gTLP performance is presented in Figure 4(b), and the semiparametric performance is shown in panel (c). Here, 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 a positive or a negative trend in the coefficient for gene 1195 as λ decreases provides a perfect classification. Note that we have a user-driven setting within this real data-example. However, it is perhaps even more noteworthy that we have replicated a classification based on 96 genes by applying the gTLP to the relationship between two genes found among 1,000s using a simple ANOVA. Lastly, Figure 4(b) shows an unexpected result for the 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(a). Consequently, the gTLP might additionally provide a degree of robustness in its estimation process.

5. Discussion

Using real data, and supported by simulation, we have shown 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 a group penalization to the 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 overlaps in the subgroup responses. While warranting further investigation, the truncation threshold parameter (τ) used by the gTLP improved on the gLASSO methods, likely because

of its weighting of the penalty toward 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 the gLASSO at maintaining between-component/subpopulation separation in the coefficients, while reducing within-component differences. In addition, we have confirmed that group penalties, such as the gTLP and gLASSO, can improve component identification and regression model estimation over their corresponding coefficient-specific penalties, the 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 a 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 the 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 additional scenarios to further refine the class of problems for which the gTLP shows strong promise. This work included one capstone simulation with three subgroups, offering further, and in some ways, stronger evidence for our conclusions about the gTLP's advantages. Even more advantageous to future work, it is constructed from basic scenarios; thus, our final simulation shows how the foundation built in this study can be leveraged effectively and adapted to future research. For example, a particular problem of interest occurs when a variant has a true effect for only one of several subsets of the population. In addition, future work must include scenarios with more covariates in order to make it applicable to the very likely scenario of health or disease outcomes resulting from complex functions of multiple variables. Increasing the number of covariates will also enable an exploration of the variable selection features, in addition to the grouping features. Similarly, establishing the 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 we first establish a comprehensive and credible foundation for the gTLP, which was the motivation for and aim of the current work. We thought it valuable to show how the penalty magnitude can uncover a hierarchal structure, thus showing the potential for different partitions of the population. The work to date has employed the squared loss function only, but the method can be modified to ac-

commodate other loss functions that might better serve a problem. For example, it could be interesting to consider an L_1 function in data with outliers, especially considering the robustness of the gTLP found in the gene expression data example. Finally, we showed that GCV can be used to choose a single set of coefficient estimates from among those generated by different threshold and penalty values. However, 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 study 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, our results go further, documenting the early successes of the gTLP as a hierarchical clustering tool that can uncover a subpopulation structure in a data set.

Supplementary Material

The online Supplementary Material provides full details of the second simulation example referenced in Section 3.2.

Acknowledgments

This research was supported by NIH grants R01HL65462, R01HL105397, R01GM081535, 1R01GM126002, and 2R01HL105397.

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**, 505–526.

Benaglia, T., Chauveau, D., Hunter, D. R. and Young, D. S. (2009). mixtools: An R package for analyzing mixture models. *Journal of Statistical Software* **32**, 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**, 1–122.

Chi, E. C. and Lange, K. (2015). Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics* **24**, 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. (Statistical Methodology)* **39**, 1–22.

Desarbo, W. S. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification* **5**, 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. T. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–223.

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

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* **102**, 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**, 156–172.

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

Levine, M., Hunter, D. R. and Chauveau, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika* **98**, 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**, 410–423.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, Inc, New York.

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

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

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

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

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

Tibshirani, R., Saunders, M. A., 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**, 91–108.

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

Wu, C., Kwon, S., Shen, X. T. and Pan, W. (2016). A new algorithm and theory for penalized regression-based clustering. *Journal of Machine Learning Research* **17**, 6479–6503.

Yang, S., Yuan, L., Lai, Y., Shen, X. T., Wonka, P. and Ye, J. (2012). Feature grouping and selection over an undirected graph. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 922–930.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* **93**, 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**, 49–67.

Department of Mathematical and Statistical Sciences, University of Colorado Denver, 80204

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

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA.

E-mail: weip@biostat.umn.edu

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.

E-mail: xshen@umn.edu

(Received March 2014; accepted June 2018)