BIOMETRICS 00, 1–26 DOI: 000
0000 0000

Pursuing Sources of Heterogeneity in Modeling Clustered Population

Yan Li[1], Chun Yu[2], Yize Zhao[3], Weixin Yao[4], Robert H. Aseltine[5], and Kun Chen[1,5,*]

[1]Department of Statistics, University of Connecticut, Storrs, CT, US.

[2]School of Statistics, Jiangxi University of Finance and Economics, Nanchang, China.

[3]Department of Biostatistics, Yale School of Public Health, New Haven, CT, US.

[4]Department of Statistics, University of California, Riverside, CA, US.

[5]Center for Population Health, University of Connecticut Health Center, Farmington, CT, US.

email: kun.chen@uconn.edu

SUMMARY: Researchers often have to deal with heterogeneous population with mixed regression relationships, increasingly so in the era of data explosion. In such problems, when there are many candidate predictors, it is not only of interest to identify the predictors that are associated with the outcome, but also to distinguish the true *sources of heterogeneity*, i.e., to identify the predictors that have different effects among the clusters and thus are the true contributors to the formation of the clusters. We clarify the concepts of the source of heterogeneity that account for potential scale differences of the clusters and propose a *regularized finite mixture effects regression* to achieve heterogeneity pursuit and feature selection simultaneously. We develop an efficient algorithm and show that our approach can achieve both estimation and selection consistency. Simulation studies further demonstrate the effectiveness of our method under various practical scenarios. Three applications are presented, namely, an imaging genetics study for linking genetic factors and brain neuroimaging traits in Alzheimer's disease, a public health study for exploring the association between suicide risk among adolescents and their school district characteristics, and a sport analytics study for understanding how the salary levels of baseball players are associated with their performance and contractual status.

KEY WORDS: Clustering; Finite mixture model; Generalized lasso; Population heterogeneity.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Regression is a fundamental statistical problem, of which a prototype is to model a response $y \in \mathbb{R}$ as a function of a p-dimensional predictor vector $\mathbf{x}$. In many applications, the classical assumption that the conditional association between y and $\mathbf{x}$ is homogeneous in the population does not hold. Rather, their conditional association may vary across several latent sub-populations or clusters. Such population heterogeneity can be modeled by a finite mixture regression (FMR), which is capable of identifying the clusters by learning multiple models together. Since first introduced by Goldfeld and Quandt (1973), FMR has been further developed in various directions and is widely used in various fields; see, e.g., Jiang and Tanner (1999), Bohning (1999), McLachlan and Peel (2004), and Chen et al. (2018).

In the era of data explosion, regression problems with a large sample size and/or a large number of variables become increasingly common, which makes the modeling of population heterogeneity even more relevant. However, while many high-dimensional methods have been developed for mixture regression (Khalili and Chen, 2007; Städler et al., 2010; Khalili, 2011), utilizing regularization has been mainly for the purpose of variable selection, i.e., to identify the predictors that are relevant to the modeling of the outcome.

In this paper, we tackle a challenging and interesting problem in the context of mixture model: to identify the predictors that are truly the sources of heterogeneity. That is, besides the selection of important predictors, we aim to further divide the selected predictors into two categories, the ones that only have common effects on the outcome and the ones that have different effects in different clusters. Being able to identify the sources of heterogeneity not only could reduce the complexity of the mixture model, but also could improve the model interpretability and enable us to gain deeper insights on the outcome-predictor association.

One important field that motivates our study is the imaging genetics with application to mental disorders such as Alzheimer's disease. As demonstrated by twin studies (Van Cauwen-

berghe et al., 2016), genetic factors play an import role in Alzheimer's disease and offers great promise for disease modeling and drug development. Compared with categorical diagnoses, neuroimaging trait has distinct advantages to capture disease etiology, and has been used in replacement of conventional clinical behavioral phenotypes in genome wide association studies (GWAS). Due to the availability of large-scale brain imaging and genetics data in landmark studies like the Alzheimer's Disease Neuroimaging Initiative (Weiner et al., 2013), a large body of literature in imaging genetics focuses on high-dimensional modeling to identify risk genetic variants (Vounou et al., 2012; Lu et al., 2015; Zhao et al., 2019). However, a major challenge in the field that has not been well investigated is how to link the imaging-associated genetic factors to actual disease diagnosis or progression and provide meaningful interpretations. Specifically, for progressive mental illness like Alzheimer's disease, it is critical to identify biomarkers that can predict the disease at early time. Therefore, we believe that not only there are genetic factors that impact overall disease risk, but also there are the ones that have differential impacts across some sub-groups which may be corresponding to different progressive periods/stages. While a few attempts have been made to bridge the pathological paths among genotype, imaging and clinical outcomes (Hao et al., 2017; Bi et al., 2017; Xu et al., 2017), to the best of our knowledge, none of the existing methods consider the heterogeneity within patient cohort or imaging endophenotype, nor are they capable to identify genetic factors that give arise disease sub-groups.

Indeed, the problem of heterogeneity pursuit is prevelent in various fields, ranging from genetics, population health, to even sports analytics. In a study on suicide risk among adolescents, we used data from the State of Connecticut to explore the association between suicide risk among 15-19 year old and the characteristics of their school districts. It is of great interest to learn whether different association patterns co-exist and whether they are due to the differences in demographic, social-economic, and/or academic factors of the

school districts. In a study on major league baseball players, the goal is to find out which performance measures and contract/free agent statues of the players contributed to the formation of distinct salary mechanisms or clusters.

In this work, we propose a *regularized finite mixture effects regression* model to perform feature selection and identify sources of heterogeneity simultaneously. The problem is formulated using the effects model parameterization (in analogous to the formulations used in analysis of variance), that is, the effect of each predictor on the outcome is decomposed to a common effect term and a set of cluster-specific terms that are constrained to sum up to zero. We consider adaptive ℓ_1 penalization on both the cluster-specific effect parameters and common effect parameters, which leads to the identification of the relevant variables and those with heterogeneous effects. The model estimation is conducted via an Expectation-Maximization (EM) algorithm, in which the M step results in a linearly constrained ℓ_1 penalized regression and is solved by a Bregman coordinate descent algorithm (Bregman, 1967; Goldstein and Osher, 2009). We show that the proposed approach can also be cast as a regularized finite mixture regression with a generalized lasso penalty; this connection facilitates our theoretical analysis in showing the estimation and selection consistency. Although we mainly focus on normal mixture model and ℓ_1 regularization, our approach can be readily generalized to other non-Gaussian models with broad class of penalties and constraints. A user-friendly R package is developed for practitioners to apply our approach.

2. Mixture Effects Model For Heterogeneity Pursuit

2.1 *An Overview of Finite Mixture Regression (FMR)*

We start with a description of the classical normal finite mixture regression (FMR). Let $y \in \mathbb{R}$ be a response/outcome variable and $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ be a p-dimensional predictor vector. In FMR with m components, it is assumed that a linear regression model holds for each of the m components, i.e., with probability π_j, a random sample $(y, \mathbf{x})$ belongs

to the jth mixture component ($j = 1, \ldots, m$), for which we have that $y = \mathbf{x}^{\mathrm{T}}\mathbf{b}_j + \epsilon_j$, where $\mathbf{b}_j \in \mathbb{R}^p$ is a fixed and unknown coefficient vector, and $\epsilon_j \sim N(0, \sigma_j^2)$ with $\sigma_j^2 > 0$. For the ease of notation, here the intercept term is included by setting the first element of $\mathbf{x}$ as one. Therefore, the conditional probability density function of y given $\mathbf{x}$ is

$$\sum_{j=1}^{m} \pi_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\{\frac{-(y - \mathbf{x}^{\mathrm{T}}\mathbf{b}_j)^2}{2\sigma_j^2}\} \tag{1}$$

where π_j's are the mixing probabilities satisfying $\pi_j > 0$, $\sum_{j=1}^{m} \pi_j = 1$. We write $(\mathbf{b}_1, \ldots, \mathbf{b}_m) = (\widetilde{\mathbf{b}}_1, \ldots, \widetilde{\mathbf{b}}_p)^{\mathrm{T}}$, where $\widetilde{\mathbf{b}}_k \in \mathbb{R}^m$ collects m component-specific coefficients for the predictor x_k. With finite samples, the maximum likelihood approach is often used for parameter estimation and inference in FMR, via the celebrated EM algorithm (Dempster et al., 1977) and its many variates (Meng and Rubin, 1991, 1993). Khalili and Chen (2007) was among the first to propose penalized likelihood approach for variable selection in FMR models; asymptotic properties were established in their work under the fixed p, large n paradigm. Städler et al. (2010) studied ℓ_1 penalized FMR and derived estimation errors bounds and selection consistency under general high-dimensional setups. Khalili and Lin (2013) further studied penalized FMR for a general family of penalty functions. Other relevant works include Wedel and DeSarbo (1995), Weruaga and Vía (2015), Bai et al. (2016), and Doğru and Arslan (2017). For a comprehensive review, see, e.g., Khalili (2011). The penalized FMR models have been widely applied in many real-world problems, such as gene expression analysis (Xie et al., 2008), disease progression subtyping (Gao et al., 2016), multi-species distribution modeling (Francis K. C. Hui and Foster, 2015), protein clustering (Chen et al., 2018), among others.

In the above mixture setup, the variance parameters σ_j^2 play important roles. Unlike in regular linear regression where its single variance parameter generally can be treated as nuisance in the estimation of the regression coefficients, the variance parameters in mixture models directly impact on the scaling (thus interpretation) and estimation of the regression coefficients of the multiple mixture components, and consequently, they also affect the

assessment and even the definition of "heterogeneous regression effects". To facilitate the further discussion, we present a re-scaled version of FMR (Städler et al., 2010),

$$\boldsymbol{\phi}_j = \frac{\mathbf{b}_j}{\sigma_j} = (\phi_{1j}, \ldots, \phi_{pj})^{\mathrm{T}}, \ \rho_j = \sigma_j^{-1} \quad (j = 1, \ldots, m),$$

and subsequently rewrite the conditional density in (1) as

$$f(y \mid \mathbf{x}, \boldsymbol{\vartheta}) = \sum_{j=1}^{m} \pi_j \frac{\rho_j}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(\rho_j y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\phi}_j)^2\}, \tag{2}$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_m; \pi_1, \ldots, \pi_m; \rho_1, \ldots, \rho_m)$ collects all the unknown parameters. We write

$$\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_m) = (\widetilde{\boldsymbol{\phi}}_1, \ldots, \widetilde{\boldsymbol{\phi}}_p)^{\mathrm{T}} \in \mathbb{R}^{p \times m}, \qquad \boldsymbol{\phi} = \mathrm{vec}(\boldsymbol{\Phi}^{\mathrm{T}}) \in \mathbb{R}^{pm},$$

where $\widetilde{\boldsymbol{\phi}}_k \in \mathbb{R}^m$ collects m component-specific regression coefficients for the predictor x_k for $k = 1, \ldots, p$ and $\mathrm{vec}(\cdot)$ is the columnwise vectorization operator.

2.2 *Sources of Heterogeneity under Finite Mixture Regression*

Now let's consider predictor selection and heterogeneity pursuit. A predictor x_k is said to be relevant or important, if $\widetilde{\mathbf{b}}_k \neq \mathbf{0}$, or equivalently, $\widetilde{\boldsymbol{\phi}}_k \neq \mathbf{0}$. Correspondingly, define

$$\mathcal{S}_R = \{k; 1 \leqslant k \leqslant p, \widetilde{\boldsymbol{\phi}}_k \neq \mathbf{0}\}$$

to be the index set of all the relevant predictors, and let $p_0 = |\mathcal{S}_R|$ denote its size. Estimating $\mathcal{S}_R$ is typically the main task of a variable selection method.

We aim higher. That is, besides identifying the relevant variables, we want to also find out among them which ones actually contribute to the population heterogeneity. However, the concept of "source of heterogeneity" is not as easily defined as it appears, since the different mixture components are possibly with different scales. We consider two definitions.

DEFINITION 1: A predictor x_k is said to be a source of heterogeneity, if $\widetilde{\mathbf{b}}_k \neq c\mathbf{1}$ for any $c \in \mathbb{R}$.

DEFINITION 2: A predictor x_k is said to be a scaled source of heterogeneity, if $\widetilde{\boldsymbol{\phi}}_k \neq c\mathbf{1}$ for any $c \in \mathbb{R}$.

Both definitions have their own merits. Definition 1 is in terms of the inequality of each raw coefficient vector $\widetilde{\mathbf{b}}_k$ from (1), which is simple and aims to draw a direct comparison of the raw effects of x_k in different mixture components regardless of their scales. Definition 2 is in terms of the scaled counterpart $\widetilde{\boldsymbol{\phi}}_k$ in (2), and the motivation is to distinguish the heterogeneity induced by the predictors and that caused by inherit scaling differences. In other words, under the second definition, we compare the standardized effects of x_k in different mixture components after putting them on the same scale. An analogy can be drawn from the familiar analysis of variance context: comparing the means of different groups is mostly appropriate when the groups are with the same variances. Notice that the two definitions become equivalent when the component variances are equal, e.g., $\sigma_1^2 = \cdots = \sigma_m^2$, which is a commonly adopted assumption in mixture regression analysis.

In this work, we shall mainly focus on Definition 2, although our methodologies can be readily modified to handle the alternative definition. Based on Definition 2, let $\mathcal{S}_H = \{k; 1 \leqslant k \leqslant p, \widetilde{\boldsymbol{\phi}}_k \neq c\mathbf{1}, \forall c \in \mathbb{R}\}$ and $p_{00} = |\mathcal{S}_H|$. Henceforth, our objective is to recover both $\mathcal{S}_R$ and $\mathcal{S}_H$. This can potentially lead to a much more parsimonious and interpretable model. To see this, consider as above that in a m-component mixture model with p predictors, there are p_0 relevant variables, and among those, only p_{00} variables are sources of heterogeneity. The classical FMR fits a model with mp free regression parameters, which can be infeasible when p is even moderately large comparing to the sample size. Meanwhile, the best model a sparse predictor selection method can possibly produce would have mp_0 free regression parameters. We can do better: since only p_{00} predictors are truly the source of heterogeneity, the best model would have only $p_0 + (m-1)p_{00}$ regression parameters. The saving can be substantial when $p_{00} \ll p_0 \ll p$ and/or m is large. As an example, consider one of the simulation settings to be presented in Section 5 with $m = 3$, $p = 30$, $p_0 = 10$, and $p_{00} = 3$. The classic FMR is with $mp = 90$ regression parameters, the sparse selection method can possibly reduce

the number to be $mp_0 = 30$, while our method can possibly further reduce the number to $p_0 + (m-1)p_{00} = 16$ through identifying the sources of heterogeneity.

2.3 *Regularized Mixture Effects Regression*

Motivated by the so-called effects-model formulation commonly used in analysis of variance models, we propose the following *constrained mixture effects model* formulation, to facilitate the pursuit of the sources of heterogeneity in mixture regression,

$$f(y \mid \mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{m} \pi_j \frac{\rho_j}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(\rho_j y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_j)^2\}, \text{ s.t.} \sum_{j=1}^{m} \beta_{jk} = 0, k = 1, \ldots, p, \quad (3)$$

where $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^{\mathrm{T}} \in \mathbb{R}^p$ collects the common effects, and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^{\mathrm{T}} \in \mathbb{R}^p$, $j = 1, \ldots, m$, are the coefficient vectors of cluster-specific effects. The equality constraints are necessary to ensure the identifiablility of the parameters. We write

$$\mathbf{B} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m) = (\widetilde{\boldsymbol{\beta}}_1, \ldots, \widetilde{\boldsymbol{\beta}}_p)^{\mathrm{T}} \in \mathbb{R}^{p\times(m+1)}, \qquad \boldsymbol{\beta} = \mathrm{vec}(\mathbf{B}^{\mathrm{T}}) \in \mathbb{R}^{p(m+1)},$$

where $\widetilde{\boldsymbol{\beta}}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{mk})^{\mathrm{T}} \in \mathbb{R}^{m+1}$ collects the common effect and the m cluster-specific effects for predictor x_k. The rest of the terms are similarly defined as in (2), except that we now write $\boldsymbol{\theta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m; \pi_1, \ldots, \pi_m; \rho_1, \ldots, \rho_m)$ to correct all the parameters under this alternative effects-model parameterization.

Now a predictor x_k is deemed to be relevant whenever $\widetilde{\boldsymbol{\beta}}_k \neq \mathbf{0}$. Moreover, a relevant variable is deemed to be a source of heterogeneity only if there exists a $1 \leqslant j \leqslant m$ such that $\beta_{jk} \neq 0$. As such, variable selection and heterogeneity pursuit can be achieved together through a sparse estimation of $\mathbf{B}$. With n independent samples $\{(y_i, \mathbf{x}_i); i = 1, \ldots, n\}$, we propose to conduct model estimation by maximizing a constrained penalized log-likelihood criterion,

$$\max_{\boldsymbol{\theta}} \left\{ \ell_\lambda^\gamma(\boldsymbol{\theta}) \equiv \sum_{i=1}^{n} \log\{f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})\} - n\lambda \sum_{k=1}^{p} \mathcal{P}_\gamma(\widetilde{\boldsymbol{\beta}}_k) \right\}, \text{ s.t.} \sum_{j=1}^{m} \beta_{jk} = 0, k = 1, \ldots, p, \quad (4)$$

where $f(y \mid \mathbf{x}, \boldsymbol{\theta})$ is the conditional density function from (3), and $\mathcal{P}_\gamma(\cdot)$ is a penalty function with λ being its tuning parameter; we mainly focus on the ℓ_1 penalty (Tibshirani, 1996) and

its adaptive version (Zou, 2006; Huang et al., 2008), i.e.,

$$\mathcal{P}_\gamma(\widetilde{\boldsymbol{\beta}}_k) = \sum_{j=0}^{m} w_{jk}|\beta_{jk}|, \qquad w_{jk} = |\widehat{\beta}^0_{j,k}|^{-\gamma} \tag{5}$$

where $w_{jk}s$ are the adaptive weights constructed from some initial estimator $\widehat{\beta}^0_{j,k}$, with $\gamma = 0$ corresponding to the non-adaptive version and $\gamma > 0$ the adaptive version. Apparently there are many other reasonable choices of penalty functions (Fan and Li, 2001; Khalili and Lin, 2013), but our choice of ℓ_1 is simple, convex and yet fundamental for sparse estimation.

Interestingly, the proposed constrained sparse estimation approach can also be understood as a generalized lasso method (She, 2010; Tibshirani and Taylor, 2011) based on the unconstrained model formulation in (2). To see this, observe that each $\widetilde{\boldsymbol{\beta}}_k$ can be written as a function of $\widetilde{\boldsymbol{\phi}}_k$ as

$$\widetilde{\boldsymbol{\beta}}_k = \mathbf{A}\widetilde{\boldsymbol{\phi}}_k, \qquad \mathbf{A} = \begin{pmatrix} 1/m\mathbf{1}_m^{\mathrm{T}} \\ \mathbf{I}_m - 1/m\mathbf{J}_m \end{pmatrix} \in \mathbb{R}^{(m+1)\times m},$$

where $\mathbf{1}_m$ is the $m \times 1$ vector of all ones, $\mathbf{I}_m$ is the $m \times m$ identity matrix and $\mathbf{J}_m$ is the $m \times m$ matrix of ones. Therefore, the generalized lasso criterion is expressed as

$$\max_{\boldsymbol{\vartheta}} \left\{ l^\gamma_\lambda(\boldsymbol{\vartheta}) \equiv \sum_{i=1}^{n} \log\{f(y_i \mid \mathbf{x}_i, \boldsymbol{\vartheta})\} - n\lambda\|\mathbf{W}(\mathbf{I}_p \otimes \mathbf{A})\boldsymbol{\phi}\|_1 \right\}, \tag{6}$$

where $f(y \mid \mathbf{x}, \boldsymbol{\vartheta})$ is the conditional density function from (2), and $\mathbf{W} = \mathrm{diag}\{w_{jk}\} \in \mathbb{R}^{p(m+1)\times p(m+1)}$ is constructed from the adaptive weights in (5) accordingly.

PROPOSITION 1: The two problems in (4) and (6) are equivalent, in the sense that

- If $\widehat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\rho}})$ solves (6), then $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\rho}})$ solves (4) where $\widehat{\boldsymbol{\beta}} = (\mathbf{I}_p \otimes \mathbf{A})\widehat{\boldsymbol{\phi}}$.
- And conversely, if $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\rho}})$ solves (4), then $\widehat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\rho}})$ solves (6) where $\widehat{\boldsymbol{\phi}}$ is such that $\widehat{\boldsymbol{\phi}}_j = \widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_j$, $j = 1, \ldots, m$.

It turns out that (4) is more convenient to use in computation, while (6) is more useful in the theoretical investigation. We also show that these penalized estimation criteria avoids the unbounded likelihood problem (McLachlan and Peel, 2004) in Web Appendix B.

3. Asymptotic Properties

The generalized lasso formulation allows us to perform the asymptotic analysis under the unconstrained mixture regression model setup given in (2). The main issue is then in dealing with the special form of the generalized lasso penalty in (6). To make things clear, we use $\boldsymbol{\theta}^*$ or $\boldsymbol{\vartheta}^*$ to denote the true parameters. We have defined $\mathcal{S}_R$ and $\mathcal{S}_H$ as the sets of relevant predictors and the predictors of sources of heterogeneity, respectively. Correspondingly, define

$$\mathcal{S} = \{i; ((\mathbf{I}_p \otimes \mathbf{A})\boldsymbol{\phi}^*)_i \neq 0\}.$$

Recall that $\boldsymbol{\beta}^* = \text{vec}(\mathbf{B}^{*\text{T}}) = (\mathbf{I}_p \otimes \mathbf{A})\boldsymbol{\phi}^*$, which means that $\mathcal{S}$ encodes the sparsity pattern of all the regression coefficients $\boldsymbol{\beta}^*$ in the effects models. Then the recovery of $\mathcal{S}_R$ and $\mathcal{S}_H$ is immediate if $\mathcal{S}$ can be recovered.

We consider the setup that the design is random, and the number of predictors p and the number of components m are considered as fixed as the sample size n grows. Building upon the works by Fan and Li (2001), Städler et al. (2010) and She (2010), our main results are presented in the following two theorems.

THEOREM 1 (Non-adaptive Estimator): Consider model (2) with random design, fixed p and m. Choose $\lambda = O(n^{-1/2})$. Assume the regularity conditions (A)-(C) from Web Appendix A on the joint density of $(y, \mathbf{x})$ hold. Then for $\gamma = 0$, there exists a local maximizer $\widehat{\boldsymbol{\vartheta}}^{\gamma}_{\lambda}$ of (6) such that $\sqrt{n}(\widehat{\boldsymbol{\vartheta}}^{\gamma}_{\lambda} - \boldsymbol{\vartheta}^*) = O_p(1)$.

THEOREM 2 (Adaptive Estimator): Consider model (2) with random design, fixed p and m. Choose $\sqrt{n}\lambda \to 0$, $n^{(\gamma+1)/2}\lambda \to \infty$ as $n \to \infty$, and suppose the initial estimator in constructing the weights is $\sqrt{n}$-consistent, i.e., $\sqrt{n}(\widehat{\boldsymbol{\vartheta}}^{ini}_{\lambda} - \boldsymbol{\vartheta}^*) = O_p(1)$. Assume the regularity conditions (A)-(C) from Web Appendix A on the joint density of $(y, \mathbf{x})$ hold. Then for any $\gamma > 0$, there exists a local maximizer $\widehat{\boldsymbol{\vartheta}}^{\gamma}_{\lambda}$ of (6) such that it is $\sqrt{n}$-consistent and $P(\widehat{\mathcal{S}}^{\gamma}_{\lambda} = \mathcal{S}) \to 1$ as $n \to \infty$.

Theorem 1 shows that the non-adaptive estimator can achieve $\sqrt{n}$-consistency in model estimation, under typical regularity conditions on the joint density of $(y, \mathbf{x})$. Theorem 2 shows that the adaptive estimator, under the same conditions and with weights constructed from a consistent estimator such as the non-adaptive one in Theorem 1, can further achieve consistency in feature selection and heterogeneity pursuit.

4. Computation

We propose a generalized EM algorithm for optimizing the criterion in (4), which enjoys desirable convergence guarantee that the object function is monotone ascending along the iterations. The algorithmic structure is mostly straightforward based on the work of Städler et al. (2010), except that in the M-step we need to efficiently solve an ℓ_1 regularized weighted least squares problem with equality constraints. A Bregman coordinate descent algorithm (Goldstein and Osher, 2009) is proposed to solve it. For tuning the number of component m and the penalty parameter λ, we propose to minimize a Bayesian information criterion (BIC). To save space, the derivations of the algorithm and the details on tuning are provided in Web Appendix C.

5. Simulation

We compare the following methods via simulation,

- Normal mixture regression with variable selection via lasso (Mix-L, or M1) and via adaptive lasso (Mix-AL, or M2), proposed by Städler et al. (2010).
- The proposed normal mixture effects regression with variable selection and heterogeneity pursuit via lasso (Mix-HP-L, or M3) and via adaptive lasso (Mix-HP-AL, or M4).

The sample size is set to $n = 200$ and the number of components is set to $m = 3$. The data on the predictors, $\mathbf{x}_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$, are generated independently from multivariate

normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. We consider two correlation structures, i.e., the uncorrelated case with $\mathbf{\Sigma} = \mathbf{I}_p$, and the correlated case with $\sigma_{ij} = 0.5^{|i-j|}$ where σ_{ij} denotes the (i, j)'s entry of $\mathbf{\Sigma}$. We consider three predictor dimensions: $p \in \{30, 60, 120\}$. As such, the number of free model parameters is 95, 185, and 365, respectively, being either comparable or much larger than the sample size.

In each setting, the first $p_0 = 10$ predictors are relevant, and among them only $p_{00} = 3$ predictors have scaled heterogeneous effects over different components according to Definition 2. Specifically, under the mixture effects model (3) with $m = 3$, the sub-vectors of the first 10 entries of the scaled coefficient vectors $\boldsymbol{\beta}_j$, denoted as $\boldsymbol{\beta}_{j0}$, $j = 0, 1, 2, 3$, are set as

$$\boldsymbol{\beta}_{00} = (1, 1, 1, 1, 1, 1, 0, 0, 0)^{\mathrm{T}}/\sqrt{\delta}, \quad \boldsymbol{\beta}_{10} = (0, 0, 0, 0, 0, 0, 0, 0, -3, 3)^{\mathrm{T}}/\sqrt{\delta},$$

$$\boldsymbol{\beta}_{20} = (0, 0, 0, 0, 0, 0, 0, -3, 3, 0)^{\mathrm{T}}/\sqrt{\delta}, \quad \boldsymbol{\beta}_{30} = (0, 0, 0, 0, 0, 0, 0, 3, 0, -3)^{\mathrm{T}}/\sqrt{\delta},$$

and the variance components are set as $(\sigma_1^2, \sigma_2^2, \sigma_3^2)^{\mathrm{T}} = \delta \times (0.1, 0.1, 0.4)^{\mathrm{T}}$, where δ controls the signal to noise ratio (SNR) defined as $\mathrm{SNR} = \sum_{j=1}^m \pi_j \mathbf{b}_j^{\mathrm{T}} \mathrm{cov}(\mathbf{X}) \mathbf{b}_j / \sum_{j=1}^m \pi_j \sigma_j^2$, with $\mathbf{b}_j = (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_j) \times \sigma_j$, $j = 1, \ldots, m$ being the corresponding unscaled coefficient vectors as in the mean model (1). We remark that $\mathbf{b}_j$s remain the same for different δ values, for facilitating the comparison among different SNRs. We choose $\delta = 1/8, 1/4, 1/2, 1, 2$, corresponding to $\mathrm{SNR} = 200, 100, 50, 25, 12.5$, respectively. We set $\pi_1 = \pi_2 = \pi_3 = 1/m$ and generate the response values by (3). We choose the tuning parameter λ and the number of components $m \in \{2, 3, 4\}$ by minimizing BIC. The experiment is repeated 500 times under each setting.

The following performance measures are computed. The estimation performance for the unscaled regression coefficients $(\mathbf{b}_1, \ldots, \mathbf{b}_m)$, the mixing probability $(\pi_1, \ldots, \pi_m)$ and the variances $(\sigma_1^2, \ldots, \sigma_m^2)$ is measured by their corresponding mean squared errors (MSE). The variable selection performance is measured by the false positive rate (FPR) and the true positive rate (TPR) for identifying relevant predictors, and the false heterogeneity rate (FHR) for identifying predictors with heterogeneous effects. Specifically, they are defined as below:

- FPR = #falsely selected variables with no effects / #variables with no effects;
- TPR = #correctly selected variables with effects / #variables with effects;
- FHR = #falsely selected variables with heterogeneous effects / #variables with common effects.

Figure 1 displays the boxplots of mean squared errors in various simulation settings, and Table 1 shows the detailed results for $p = 60$ with $\mathbf{\Sigma} = \mathbf{I}_p$. The results for $p \in \{30, 120\}$ and for the cases of correlated predictors convey similar messages, which are provided in Web Appendix D. The findings are summarized as follows.

- As expected, in general the larger the signal to noise ratio and the smaller the model dimensions, the better the performance of each method.
- Adaptive method in general leads to more accurate results in both model estimation and variable selection than its non-adaptive counterpart. The improvement can be substantial. Specifically, both Mix-HP-L and Mix-HP-AL rarely miss important variables, but the former tends to select a larger model with more irrelevant variables. Indeed, the over-selection property of ℓ_1 penalization is well known.
- The proposed methods Mix-HP-L and Mix-HP-AL outperform their counterparts without heterogeneity pursuit, Mix-L and Mix-AL, respectively, in most simulation setups, except that when SNR $= 12.5$ and $p = 120$, all methods suffer from very low signal to noise ratio and very high dimensionality. The Mix-HP-AL has the best performance among all the competing methods; its improvement over others can be substantial especially when the signal is weak or moderate and the model dimension is high; moreover, in those relatively difficult scenarios, even Mix-HP-L can outperform Mix-AL.
- We have examined settings where all relevant predictors have heterogeneous effects, for which the methods with or without heterogeneity pursuit perform similarly. We have also considered settings with unequal mixing probabilities, where the implications are similar; see

Web Appendix D. These results clearly demonstrate the benefit of heterogeneity pursuit, as it enables the potential of identifying the most parsimonious model.

We conclude that overall the proposed heterogeneity pursuit approach with adaptive lasso (Mix-HP-AL) is preferable to both the non-adaptive counterpart Mix-HP-L and the conventional methods like Mix-L and Mix-AL. The proposed method is particularly beneficial when it is believed that only very few predictors contribute to the regression heterogeneity.

[Figure 1 about here.]

[Table 1 about here.]

6. Applications

6.1 *Alzheimer's Disease Neuroimaging Initiative (ADNI)*

We performed imaging genetics analysis based on data from the ADNI database which is a public-private partnership to study the progression of mild cognitive impairment and early Alzheimer's disease based on different data sources including genetics, neuroimaging, clinical assessments, etc. (See ADNI for detailed study design and data collection information). Our goal here is to find out whether distinct clusters of disease-gene associations exist, possibly corresponding the disease stages, and to identify common genetic factors associated with overall disease risk, as well as cluster-specific ones.

Briefly, to control data quality and reduce population stratification effect, we only include 760 Caucasian subjects in this analysis. For each subject, single nucleotide polymorphisms (SNPs) genotyping data were acquired by Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA) according to the manufacturer's protocols; and raw MRI data were collected through 1.5 Tesla MRI scanners and then preprocessed by standard steps including anterior commissure and posterior commissure correction, skull-stripping, cerebellum removing, intensity inhomogeneity correction, segmentation, and registration (Shen and Davatzikos,

2004). The preprocessed brain images were further labelled regionally by existing template and then transferred following the deformable registration of subject images (Wang et al., 2011), which eventually led to 93 regions of interest over whole brain. After removing the ones with sex check failure, more than 10% missing single nucleotide polymorphisms (SNPs), and outliers, 741 subjects including 174 Alzheimer's disease, 362 mild cognitive impairment and 205 healthy controls remain in the analysis.

We consider the following imaging phenotypes: two global brain measurements, i.e. whole brain/white matter volumes, and two Alzheimer's disease related endophenotypes, i.e. left and right lateral ventricles volumes. For each imaging trait, we include both SNPs belonging to the top 10 Alzheimer's disease candidate genes provided by the AlzGene database and those identified from United Kingdom (UK) Biobank (Zhao et al., 2019) ($\sim$20,000 subjects) under the same imaging phenotype. The final lists of SNP names are provided in supplementary materials. We fit our proposed Mix-AL-HP model for each imaging trait and its corresponding genetic predictors to examine the cluster patterns and select risk factors that impact the whole cohort with common effects and those impact the sub-groups/clusters heterogeneously. Age, gender and the top five genetic principal components are always included in the models as controls with common effects and no regularization. We fit models with different component numbers ($m \in \{1, 2, \ldots, 5\}$) and with/without the assumption of equal variance; the best model is selected based by BIC.

We first examine the identified clusters for each imaging trait to see whether the cluster pattern is associated with disease progression. The numbers of clusters for the four imaging traits, left/right ventricles and whole brain/white matter volumes, are 2, 3, 3, 1 with the smallest BIC values regarding λ being 1873.12, 1871.58, 1794.04 and 2144.72, respectively. And among the three imaging traits with more than one identified cluster, the average values of imaging phenotype are shown to be clearly different over different clusters. See

Web Figure 2 in Web Appendix E, which shows the cluster-specific boxplots of each imaging trait. Intriguingly, given the fact that the size of brain increases along Alzheimer's disease progression, we are able to clearly align the identified clusters to different disease stages in light of the average volume of imaging traits. Note that for the white matter volume, which is a global brain phenotype, no cluster pattern is detected, which is biologically reasonable due to its weaker pathological bounding to disease etiology.

All the selected SNPs and their types of effect (common or cluster-specific) are summarized in Figure 2. Most of the identified genetic risk variants (e.g. SNPs within genes CD2AP, MRVI1, GNA12) associated with two Alzheimer's disease imaging biomarkers are consistent and subtype-related, indicating the existence of varying genetic effects on brain structure over diesease progression. Meanwhile, the selected SNPs related to global brain phenotypes are generally with common effect; again, this is due to their weaker pathological bounding to disease etiology compared with the Alzheimer's disease related endophenotypes.

Figure 3 provides a visualization of the estimated coefficients of each selected SNP under different clusters. Based on Figure 3, we successfully detect a few SNPs showing a particular strong impact on early- to middle-stage Alzheimer's disease including rs2025935, rs677909 and rs798532 located in genes BIN1, MS4A4E and GNA12. Among them, BIN1 is the key molecular factor to modulate tau pathology and has recently been recognized as an important risk locus for late-onset Alzheimer's disease (Tan et al., 2013); MS4A4E has been detected by GWAS as a genetic risk factor for Alzheimer's disease based on Alzheimer's Disease Genetic Consortium (Hollingworth et al., 2011); and GNA12, though has not been extensively reported in existing experiments, is known to over-express in human brain. Due to a typical small/moderate effect of single genetic signal, some of these variants are highly likely to be buried under existing methods without clustering overall heterogeneity. Moveover, our results provide valuable insights to prioritize future early therapeutic strategies even among

all the Alzheimer's disease genotypes. In terms of other selected SNPs, most of them have been recognized as Alzheimer's disease risk factors in previous experiments or analyses, and they either show a common effect across all the clusters or a mixing one including both early and late stages in our results. Detailed estimation results are reported in Web Appendix E.

[Figure 2 about here.]

[Figure 3 about here.]

6.2 *Connecticut Adolescent Suicide Risk Study*

Suicide among youth is a serious public health problem in the United States. The Centers for Disease Control (CDC) reported that suicide is the third leading cause of death of youth aged 15–24 based on 2013 data, and more alarmingly there has been an increasing trend over time. Suicide prevention among youth is a very challenging task, which requires a systematic approach through developing reliable metrics for assessing suicide risk, locating areas of greater risk for effective resource allocation, identifying important risk factors, among others.

We use data from the State of Connecticut at the school district level to explore the association between suicide risk among 15-19 year olds and the characteristics of their school district. Specifically, the overall suicide risk of the 15-19 age group within each school district is proxied by its log-transformed 5-year average rate of inpatient hospitalizations due to suicide attempts from 2010 to 2014 (per year per 10,000 population). Several characteristics of the $n = 119$ school district characteristics were collected in the same period: (1) demographic measures, including percent of households that included an adult male, average household size, percent of the population that are under 18 years of age, percent of population who are White; (2) academic measures, including average score on the Connecticut Academic Performance Test (CAPT), graduation rate, dropout rate, and attendance rate of high schools in the district; (3) behavioral measures, including incidence rate, defined as the ratio between the number of disciplinary incidences and the total enrollment, and serious

incidence rate; (4) economic measures, including median income and free lunch rate. More details about the data can be found in Chen and Aseltine (2017).

In the previous study, a generalized mixed-effects model was used to estimate the common effects of the school district characteristics on the suicide risk (through fixed-effects terms) and identify the "overachievers" and "underachievers" (through district-level random effects) among school districts. (It was also shown that there was no significant spatial effect.) Indeed, the existence of these anomalous school districts suggests that the regression association may not be homogeneous, and thus it is interesting to see whether additional insights can be gained by a mixture regression analysis, to reveal the potentially heterogeneous association structure, cluster the school districts, and identify the district characteristics that drive the heterogeneity. We thus apply our proposed Mix-HP-AL method to analyze the data. For dealing with the highly-correlated school district measurements, we perform group-wise principal component analysis and use each leading factor to summarize the information of each category, which results in $p = 4$ district factors; the details of the principal component analysis results are provided in Web Appendix F.

Table 2 reports the estimation results, and Figure 4 shows the corresponding cluster pattern of the school districts using the naive Bayes classification rule with the estimated component probabilities $\widehat{p}_{ij}$, i.e., $\widehat{z}_{ik} = 1$ if $k = \arg\max_j \widehat{p}_{ij}$. A three-component model is selected based on the BIC, in which the three factors differentiate school districts not in terms of their overall suicide risk as we did in our prior analysis, but in terms of the association between the risk factors and suicide risk. In Table 2 one can see that only the demographic and academic factors are selected; when conditioning on the selected factors, the economic factor and the behavior factor are no longer related to the suicide risk, which may be partly due to the fact that the four factors are still moderately correlated. The major difference among the 3 clusters of communities involves the direction of effects of the demographic factor, which indicates

418 a great deal of heterogeneity in how this factor impact suicide risk across communities.
419 The majority of the school districts are in cluster 3, in which the suicide risk is negatively
420 associated with the demographic factor; that is, in general, the larger the household size, the
421 greater percentage of households with an adult male, the greater the proportion of population
422 under age 18, and higher the proportion of White residents are associated with *lower* suicide
423 risk, after adjusting for the effect of academic performance. In contrast, in cluster 1 the
424 association between the suicide risk and the demographic factor is positive, such that higher
425 rates of male householders, larger household size, higher proportions of children under 18, and
426 a higher proportion of White residents is associated with *higher* suicide risk. Further analysis
427 reveals that the 12 school districts in cluster 1 have significantly lower mean suicide risk than
428 those in cluster 3; this suggests that the impact of the demographic factors on suicide risk
429 may change and even flip sign with the mean suicide risk level itself. It is possible that this
430 is caused by some "unmeasured" factors confounded with the demographic factor. Cluster
431 2 is the smallest in size among the three, consisting of "Regional 19" (near the University
432 of Connecticut), "New London" and "Monroe"; these are anomalous districts with very low
433 suicide risk. The academic factor, in contrast, is identified to have only common effects after
434 scaling by the variances according to Definition 2, which makes the estimated model even
435 more parsimonious. That the effect of the academic factor is always positive indicates that
436 suicide risk tends to be higher in those school districts with better academic performance;
437 as discussed in Chen and Aseltine (2017), students in school districts of better academic
438 performance could be under higher pressure, which in turn may induce more psychological
439 distress. In general, our results agree well with previous studies, and we gain additional
440 insight on the changing impact of the school district characteristics on the suicide risk.

441 [Table 2 about here.]

442 We have also compared Mix-HP-AL to Mix-AL by performing a random-splitting pro-

cedure to evaluate their out-of-sample predictive performance. Each time the data is split to 80% training for model fitting and 20% testing for out-of-sample evaluation, and the procedure is repeated 500 times. The average predictive log-likelihood (with standard error in the parenthesis) is -24.6 (2.32) and -23.2 (2.16) for Mix-AL and Mix-HP-AL, respectively, indicating that the proposed method indeed performs better for this dataset.

[Figure 4 about here.]

The proposed method has also been applied to another application in sports analytics for understanding how the salaries of baseball players are associated with their performance and contractual status. Due to space limit, the application is detailed in Web Appendix G.

7. Discussion

In this paper, we propose a mixture regression method to thoroughly explore the heterogeneity in a population of interest, which is increasingly encountered in the era of big data. Our approach goes beyond the conventional variable selection methods, by not only identifying the relevant predictors, but also distinguishing from them the true sources of heterogeneity. As such, the proposed approach can potentially lead to a much more parsimonious and interpretable model to facilitate scientific discovery.

There are a number of future research directions. It is pressing to extend the proposed method to handle non-Gaussian outcomes, such as binomial mixture and Poisson mixture. This extension can help us to improve the analysis for the suicide risk study, as the raw counts of the suicide-related hospital admissions may be better modeled by Poisson distribution. Another possible direction is to consider other forms of penalty functions. For example, when the predictors are highly correlated, it could be beneficial to use the elastic-net penalty (Zou and Hastie, 2005) to ensure stable coefficient estimation. Non-convex penalties could also be considered to improve variable selection. A related task is to extend the theoretical

analysis to high-dimensional settings where the number of variables may grow with or exceed the sample size. A potential byproduct of the proposed approach is that it can lead to automatic reduction of the number of pre-specified clusters when the effects of some clusters are estimated to be exactly the same; it is hopeful that this interesting feature can allow us to build a more general mixture learning framework where relevant variables, sources of heterogeneity and the number of clusters are simultaneously learned. It would also be interesting to consider heterogeneity pursuit in multivariate mixture regression, but it is not straightforward. The mixture components may have different covariance matrices which complicate the definition of the sources of heterogeneity, and the set of predictors with heterogeneous effects may differ across different responses.

In this work, we mainly focus on the framework of mixture regression to pursue the sources of heterogeneity at the "global" level. An interesting direction is to extend our work to utilize the frameworks of individualized modeling and sub-group analysis which mainly pursue the sources of heterogeneity at the "individual" level (Tang et al., 2020). To lessen the assumptions of mixture regression, several recent works formulate the problem as a penalized regression with a fusion-type penalty. Ma and Huang (2017) proposed a concave pairwise fusion approach to identify sub-groups with pairwise penalization on subject-specific intercepts. Austin et al. (2020) proposed a grouping fusion approach to identify unknown sub-groups and their corresponding regression models. Tang et al. (2020) proposed a method to simultaneously achieve individualized variable selection and sub-grouping. Comparing to the mixture model framework, an individualized penalized regression approach may not fully utilize the potential global mixture structure and fails to consider the potential heterogeneity in variances. Therefore, we will explore the idea of combining mixture model and individualized fusion, to simultaneously perform global and individualized heterogeneity pursuits.

Acknowledgments

Yu is supported by National Natural Science Foundation of China grant 11661038 and Jiangxi Provincial Natural Science Foundation grant 20202BABL201013. Yao is supported by U.S. National Science Foundation grant DMS-1461677 and the Department of Energy award No. 10006272. Aseltine is supported by the U.S. National Institutes of Health grant R01-MH112148, R01-MH112148-03S1, and R01-MH124740. Chen is supported by the U.S. National Science Foundation grants DMS-1613295 and IIS-1718798, and the U.S. National Institutes of Health grants R01-MH112148, R01-MH112148-03S1, and R01-MH124740. Li and Yu contributed equally to this work.

ADNI data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this paper.

Data Availability Statement

The data that support the findings of this paper are available upon request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

ADNI. http://adni.loni.usc.edu (accessed March 2020).

AlzGene. http://www.alzgene.org (accessed March 2020).

Austin, E., Pan, W., and Shen, X. (2020). A new semiparametric approach to finite mixture of regressions using penalized regression via fusion. *Statistica Sinica* .

Bai, X., Chen, K., and Yao, W. (2016). Mixture of linear mixed models using multivariate t distribution. *Journal of Statistical Computation and Simulation* **86,** 771–787.

Bi, X., Yang, L., Li, T., Wang, B., Zhu, H., and Zhang, H. (2017). Genome-wide mediation

analysis of psychiatric and cognitive traits through imaging phenotypes. *Human Brain Mapping* **38,** 4088–4097.

Bohning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications*. Chapman and Hall/CRC, Boca Raton, FL.

Bregman, L. (1967). The relaxation method of finding the common point of convex sets and its application of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics* **7,** 200–217.

Chen, K. and Aseltine, R. H. (2017). Using hospitalization and mortality data to identify areas at risk for adolescent suicide. *Journal of Adolescent Health* **61,** 192–197.

Chen, K., Mishra, N., Smyth, J., Bar, H., Schifano, E., Kuo, L., and Chen, M.-H. (2018). A tailored multivariate mixture model for detecting proteins of concordant change in the pathogenesis of *Necrotic Enteritis*. *Journal of the American Statistical Association* **113,** 546–559.

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)* pages 1–38.

Doğru, F. Z. and Arslan, O. (2017). Parameter estimation for mixtures of skew Laplace normal distributions and application in mixture regression modeling. *Communications in Statistics-Theory and Methods* **46,** 10879–10896.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348–1360.

Francis K. C. Hui, D. I. W. and Foster, S. D. (2015). Multi-species distribution modeling using penalized mixture of regressions. *Annals of Applied Statistics* **9,** 866–882.

Gao, C., Zhu, Y., Shen, X., and Pan, W. (2016). Estimation of multiple networks in gaussian mixture models. *Electronic Journal of Statistics* **10,** 1133–1154.

Goldfeld, S. M. and Quandt, R. E. (1973). A markov model for switching regression. *Journal of Econometrics* **1,** 3–15.

Goldstein, T. and Osher, S. (2009). The split bregman method for l1-regularized problems. *SIAM Journal on Imaging Sciences* **2,** 323–343.

Hao, X., Li, C., Du, L., Yao, X., Yan, J., Risacher, S. L., Saykin, A. J., Shen, L., Zhang, D., Weiner, M. W., et al. (2017). Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease. *Scientific Reports* **7,** 44272.

Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.-C., Carrasquillo, M. M., Abraham, R., Hamshere, M. L., Pahwa, J. S., Moskvina, V., et al. (2011). Common variants at abca7, ms4a6a/ms4a4e, epha1, cd33 and cd2ap are associated with alzheimer's disease. *Nature Genetics* **43,** 429.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for high-dimensional regression models. *Statistica Sinica* **18,** 1603–1618.

Jiang, W. and Tanner, M. A. (1999). Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *The Annals of Statistics* **27,** 987–1011.

Khalili, A. (2011). An overview of the new feature selection methods in finite mixture of regression models. *Journal of Iranian Statistical Society* **10,** 201–235.

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. and Lin, S. (2013). Regularization in finite mixture of regression models with diverging number of parameters. *Biometrics* **69,** 436–446.

Lu, Z., Zhu, H., Knickmeyer, R. C., Sullivan, P. F., Williams, S. N., Zou, F., and Alzheimer's Disease Neuroimaging Initiative (2015). Multiple snp set analysis for

genome-wide association studies through Bayesian latent variable selection. *Genet Epidemiol* **39,** 664–77.

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. (2004). *Finite mixture models*. John Wiley & Sons.

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86,** 899–909.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80,** 267–278.

She, Y. (2010). Sparse regression with exact clustering. *Electron. J. Statist.* **4,** 1055–1096.

Shen, D. and Davatzikos, C. (2004). Measuring temporal morphological changes robustly in brain mr images via 4-dimensional template warping. *NeuroImage* **21,** 1508–1517.

Städler, N., Bühlmann, P., and Van De Geer, S. (2010). ℓ_1-penalization for mixture regression models. *Test* **19,** 209–256.

Tan, M.-S., Yu, J.-T., and Tan, L. (2013). Bridging integrator 1 (bin1): form, function, and Alzheimer's disease. *Trends in Molecular Medicine* **19,** 594–603.

Tang, X., Xue, F., and Qu, A. (2020). Individualized multidirectional variable selection. *Journal of the American Statistical Association* .

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

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics* **39,** 1335–1371.

Van Cauwenberghe, C., Van Broeckhoven, C., and Sleegers, K. (2016). The genetic landscape of Alzheimer disease: Clinical implications and perspectives. *Genet Med* **18,** 421–30.

Vounou, M., Janousova, E., Wolz, R., Stein, J. L., Thompson, P. M., Rueckert, D., Montana, G., and Alzheimer's Disease Neuroimaging Initiative (2012). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *Neuroimage* **60,** 700–16.

Wang, Y., Nie, J., Yap, P.-T., Shi, F., Guo, L., and Shen, D. (2011). Robust deformable-surface-based skull-stripping for large-scale studies. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 635–642. Springer.

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

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E., Morris, J. C., Petersen, R. C., Saykin, A. J., Schmidt, M. E., Shaw, L., Shen, L., Siuciak, J. A., Soares, H., Toga, A. W., and Trojanowski, J. Q. (2013). The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. *Alzheimers Dement* **9,** e111–94.

Weruaga, L. and Vía, J. (2015). Sparse multivariate gaussian mixture regression. *IEEE Transactions on Neural Networks and Learning Systems* **26,** 1098–1108.

Xie, B., Pan, W., and Shen, X. (2008). Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics* **64,** 921–930.

Xu, Z., Wu, C., Pan, W., and Alzheimer's Disease Neuroimaging Initiative (2017). Imaging-wide association study: Integrating imaging endophenotypes in GWAS. *Neuroimage* **159,** 159–169.

Zhao, B., Luo, T., Li, T., Li, Y., Zhang, J., Shan, Y., Wang, X., Yang, L., Zhou, F., Zhu, Z., et al. (2019). GWAS of 19,629 individuals identifies novel genetic variants for regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *bioRxiv* page 586339.

Zhao, Y., Zhu, H., Lu, Z., Knickmeyer, R. C., and Zou, F. (2019). Structured genome-wide association studies with bayesian hierarchical variable selection. *Genetics* **212,** 397–415.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429.

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

Supporting Information

Web Appendices and Figures referenced in Sections 3–6, and the final lists of SNP names in ADNI analysis are available at the *Biometrics* website on Wiley Online Library. R package to implement the proposed approach is provided along with instructions to run the analyses.

Received March 2020.

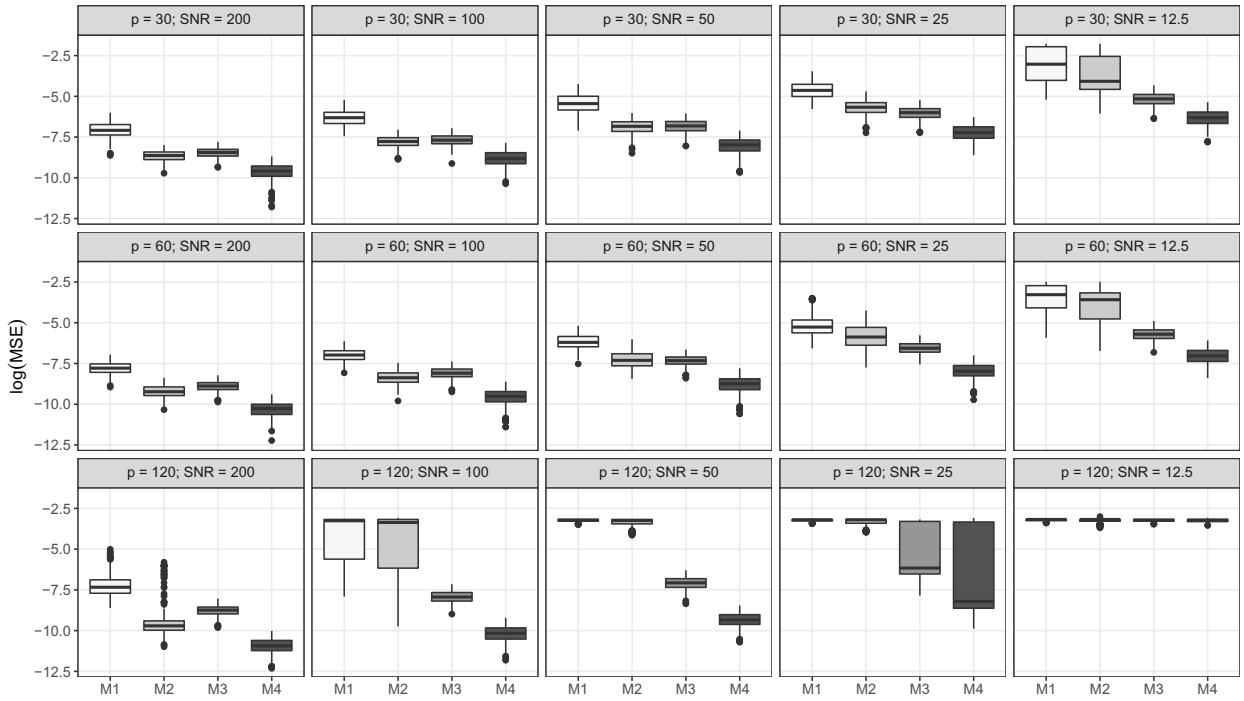

Figure 1. Boxplots of mean squared errors (in log scale) for estimating the unscaled coefficient vectors, for simulation settings with $n = 200$, $p \in \{30, 60, 120\}$, and $\mathbf{\Sigma} = \mathbf{I}_p$, and SNR $\in \{200, 100, 50, 25, 12.5\}$. Four methods are compared: Mix-L (M1), Mix-AL (M2), Mix-HP-L (M3) and Mix-HP-AL (M4). The log(MSE): logarithm of mean squared errors.

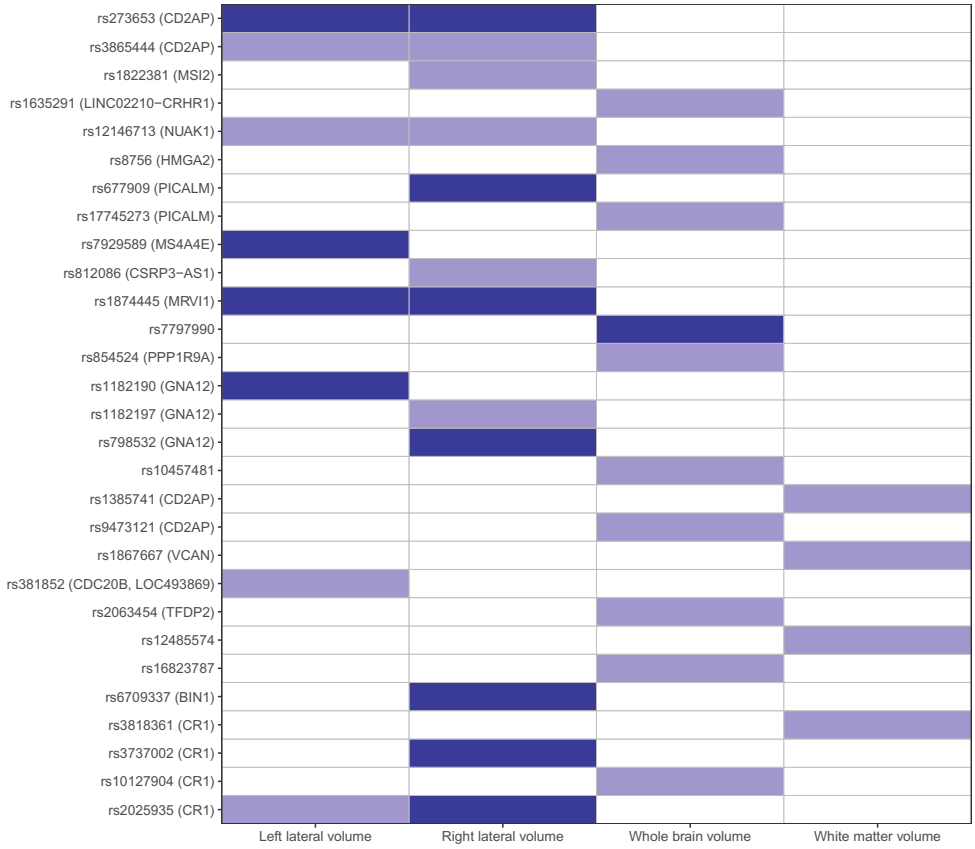

Figure 2. ADNI study: effects of selected SNPs and their associated genes for the four imaging phenotypes. Light color means a SNP has only common effect across clusters; Dark color means a SNP has different effects across clusters and thus is considered as a source of heterogeneity. The SNPs are ordered based on the their positions on chromosomes. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

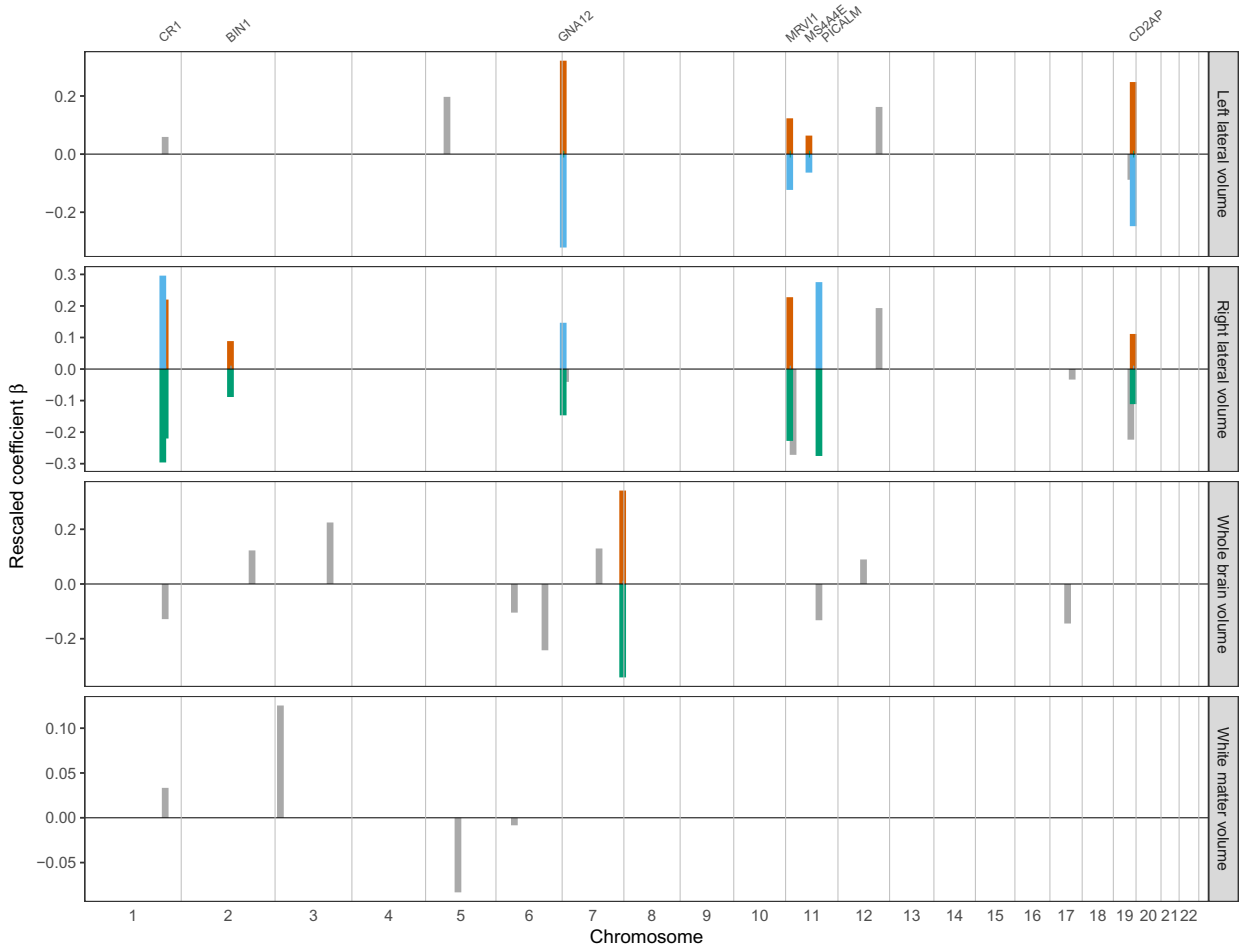

Figure 3. ADNI study: estimated scaled coefficients ($\widehat{\phi}_{kj}$) of selected SNPs for the four imaging phenotypes, showing along their positions on chromosomes. The numbers of clusters are 2,3,3,1 for the four imaging phenotypes, showing from top to bottom. For each imaging phenotype, its cluster labels are aligned with decreasing average values of the phenotype (thus correspond to different disease stages). Grey color means a SNP has only common effect across clusters; red color indicates cluster 1; blue color indicates cluster 2; and green color indicates cluster 3. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

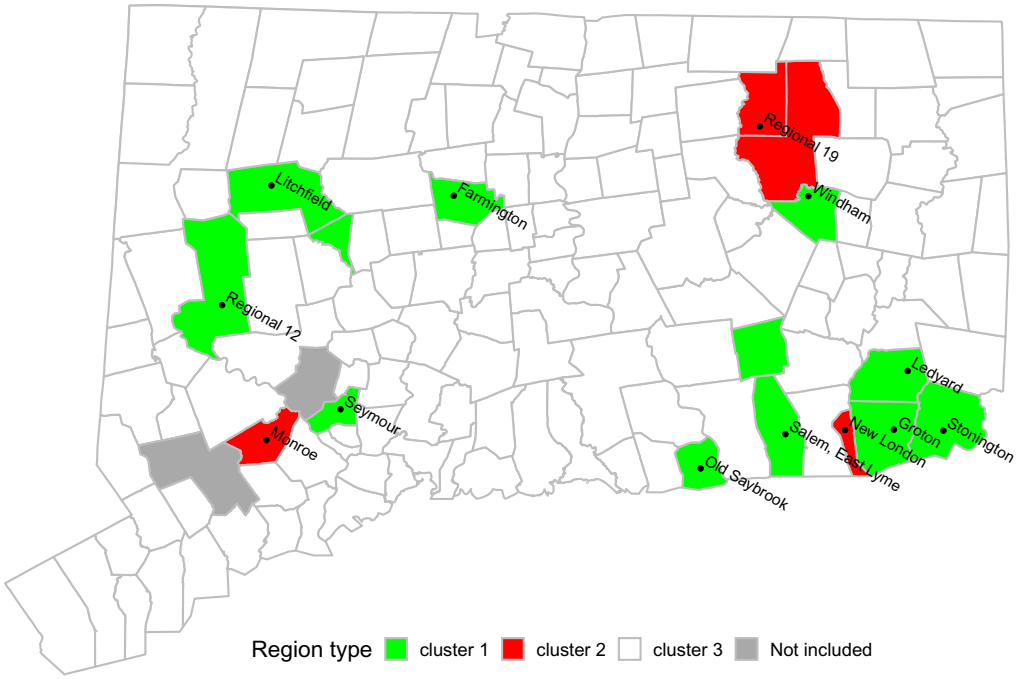

Figure 4. Suicide risk study: district clustering using Mix-HP-AL. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Table 1

Comparison of mean squared error of estimation, variable selection and heterogeneity pursuit performance of four methods, Mix-L, Mix-AL, Mix-HP-L and Mix-HP-L, under settings with $n = 200$, $p = 60$, and $\mathbf{\Sigma} = \mathbf{I}_p$. The mean squared errors (MSE) are reported along with their standard errors in the parenthesis. The simulation is based on 500 replications. The MSE values are scaled by multiplying 100, and the FPR, FHR, TPR values are reported in percentage.

		MSE			RATE		
SNR	Method	$\mathbf{b}$	$\boldsymbol{\sigma}^2$	$\boldsymbol{\pi}$	FPR	FHR	TPR
200	Mix-L	0.04 (0.02)	7.32 (4.04)	0.18 (0.15)	53.0	100.0	100.0
	Mix-AL	0.01 (0.00)	0.09 (0.08)	0.12 (0.10)	10.2	100.0	100.0
	Mix-HP-L	0.01 (0.00)	4.95 (1.88)	0.14 (0.11)	39.1	4.4	100.0
	Mix-HP-AL	0.00 (0.00)	0.07 (0.05)	0.11 (0.10)	4.0	0.3	100.0
100	Mix-L	0.10 (0.04)	10.58 (5.57)	0.22 (0.18)	48.3	100.0	100.0
	Mix-AL	0.03 (0.01)	0.10 (0.10)	0.12 (0.11)	13.3	100.0	100.0
	Mix-HP-L	0.03 (0.01)	7.04 (2.83)	0.15 (0.12)	36.8	3.8	100.0
	Mix-HP-AL	0.01 (0.00)	0.07 (0.06)	0.12 (0.10)	3.8	0.1	100.0
50	Mix-L	0.23 (0.11)	14.66 (8.24)	0.28 (0.23)	43.9	100.0	100.0
	Mix-AL	0.08 (0.05)	0.22 (0.22)	0.15 (0.13)	16.6	100.0	100.0
	Mix-HP-L	0.07 (0.02)	9.91 (3.85)	0.17 (0.14)	33.4	2.8	100.0
	Mix-HP-AL	0.02 (0.01)	0.09 (0.10)	0.13 (0.11)	3.7	0.1	100.0
25	Mix-L	0.72 (0.60)	29.07 (25.26)	0.54 (0.61)	33.9	100.0	100.0
	Mix-AL	0.36 (0.25)	1.34 (1.70)	0.28 (0.32)	19.0	100.0	100.0
	Mix-HP-L	0.15 (0.06)	12.57 (5.30)	0.19 (0.15)	33.2	1.9	100.0
	Mix-HP-AL	0.04 (0.02)	0.15 (0.16)	0.13 (0.11)	4.0	0.2	100.0
12.5	Mix-L	4.11 (2.58)	101.20 (76.46)	7.01 (5.90)	16.0	100.0	90.0
	Mix-AL	3.24 (2.56)	33.17 (37.18)	4.66 (4.71)	12.3	100.0	90.0
	Mix-HP-L	0.36 (0.13)	16.50 (8.27)	0.22 (0.18)	35.0	1.9	100.0
	Mix-HP-AL	0.10 (0.04)	0.38 (0.40)	0.15 (0.13)	5.5	0.1	100.0

Table 2
Suicide risk study: the coefficient estimates using Mix-HP-AL. The zero values are shown as blanks.

Factors	$\widehat{\boldsymbol{\phi}}_1$	$\widehat{\boldsymbol{\phi}}_2$	$\widehat{\boldsymbol{\phi}}_3$
Intercept	6.23	6.23	6.23
Demographic factor	0.27		-0.27
Academic factor	0.13	0.13	0.13
Behavioral factor			
Economical factor			
$\widehat{\sigma}$	0.33	0.20	0.46
$\widehat{\pi}$	0.13	0.02	0.85