Statistica Sinica **29** (2019), 1873-1889
doi:https://doi.org/10.5705/ss.202017.0179

ROBUST SUBGROUP IDENTIFICATION

Yingying Zhang[1], Huixia Judy Wang[2] and Zhongyi Zhu[1]

[1]*Fudan University and* [2]*The George Washington University*

Abstract: In many applications, subgroups with different parameters may exist even after accounting for the covariate effects, and it is important to identify the meaningful subgroups for better medical treatment or market segmentation. We propose a robust subgroup identification method based on median regression with concave fusion penalizations. The proposed method can simultaneously determine the number of subgroups, identify the group membership for each subject, and estimate the regression coefficients. Without requiring any parametric distributional assumptions, the proposed method is robust against outliers in the response and heteroscedasticity in the regression error. We develop a convenient algorithm based on local linear approximation, and establish the oracle property of the proposed penalized estimator and the model selection consistency for the modified Bayesian information criteria. The numerical performance of the proposed method is assessed through simulation and the analysis of a heart disease data.

Key words and phrases: Concave fusion penalization, heterogeneous parameters, median regression, model-based clustering, robust.

1. Introduction

Most statistical modeling relies on a common assumption that the same set of model parameters apply to all subjects. However, in some applications, there may exist different parameters in subgroups even after accounting for the covariate information. Mixture models have been widely used for identifying subgroups from a heterogeneous population; see for instance Banfield and Raftery (1993), Hastie and Tibshirani (1996), McNicholas (2010), Wei and Kosorok (2013), Shen and Huang (2010), Chaganty and Liang (2013) and Shen and He (2015). One advantage of mixture models is that they provide a formal and convenient model framework, and thus can easily incorporate the covariate effects of different forms. However, most mixture-model-based approaches require specifying the number of components, and the underlying distribution for each component, for which the most popular choice is normal distribution giving rise to normal mixture models. It's well known that testing for the number of components in mixture models is technically challenging due to the nonidentifiability of parameters under the null

hypothesis; see Zhu and Zhang (2004), Chen and Li (2009), Li and Chen (2010), Kasahara and Shimotsu (2015), Shen and He (2015) for some related discussions. On the other hand, the normality assumption for normal mixture models may be restrictive or susceptible to outliers.

In this paper, we develop a new robust approach that can automatically detect and identify subgroups through pairwise fused penalization. Let y_i be the scalar response variable and $\mathbf{x}_i$ the p-dimensional covariate associated with subject i, where $i = 1, \ldots, n$. We consider the following median regression model,

$$y_i = \alpha_{k_i} + \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, i = 1, \ldots, n, \tag{1.1}$$

where $\boldsymbol{\beta}$ is the unknown slope coefficient vector, ε_i are independent random errors such that $P(\varepsilon_i < 0|\mathbf{x}_i) = 1/2$, $k_i \in \{1, \ldots, K\}$ is the unknown group membership of subject i, and α_k is the group-specific intercept for $k = 1, \ldots, K$. Throughout the paper, the number of groups, K, is assumed to be unknown and bounded, and p is allowed to increase with n but we suppress its dependence on n for notational simplicity. In model (1.1), we assume that the population is heterogeneous in terms of location after accounting for the covariate effect, but the proposed method can also be easily extended to identify the heterogeneity in slopes with some modification.

We propose a median-based penalization approach through penalizing the pairwise differences of intercept coefficients across subjects. The proposed method can identify subgroups, determine K, and estimate the parameters $\boldsymbol{\beta}$ and α_k simultaneously. The idea of using penalization for clustering has been considered in Hocking et al. (2011), Pan, Shen and Liu (2013), Chi and Lange (2015), and Wu et al. (2016), to name a few. The closest literature with our model (1.1) is Ma and Huang (2017). Compared to Ma and Huang (2017), this paper has the following major differences and advances. First, by considering the median regression model (1.1), our method is based on L_1 loss and allows the errors to be heavy-tailed or dependent of $\mathbf{x}_i$, and thus provides more flexibility and robustness against outliers. The nice properties of L_1 distance for heavy-tailed distributions have been studied in the classification setup in Hall, Titterington and Xue (2009). Second, we develop a convenient algorithm through local linear approximation (LLA, Zou and Li (2008)) to deal with L_1 loss and concave penalties. At each step of the iteration, the optimization becomes a linear programming problem and can be solved easily with existing software. As a result, this method enables more efficient initial values and leads to much faster convergence than the alternating direction method of multipliers (ADMM) algorithm in

Ma and Huang (2017). We also suggest a divide-and-conquer algorithm to further reduce the computational burden for data with large n. Third, we establish the oracle property of the proposed estimator without the restrictive distributional or moment conditions on the random errors as required in Ma and Huang (2017). The oracle property shows that the oracle estimator (obtained with correct group membership) is a local solution of the proposed median regression with fused penalization, indicating that the method can identify the correct subgroups with high probability. Lastly, we propose a modified Bayesian information criterion (BIC) to choose the penalization parameter, and establish its consistency.

The proposed method can be regarded as an unsupervised learning or a model-based clustering method, which performs grouping pursuit of subject-specific intercepts through penalization. Unlike supervised learning, the method does not model the characteristics of subgroups, so the estimation cannot be directly used to predict the group membership of new subjects. On the other hand, the method is more data adaptive and assumption-lean since it does not require specifying a parametric model for the grouping probability as in e.g. Shen and He (2015). In practice, we can apply the proposed method as a first step to obtain some assumption-lean clustering, and then use the clustering results as responses to perform binary or multinomial regression to characterize the subgroups by a given set of variables. Such two-step analysis has also been considered in Dusseldorp, Conversano and Van Os (2010) for subgroup identification in clinical trials and Müllensiefen, Hennig and Howells (2017) for marketing segmentation.

The rest of this paper is organized as follows. In Section 2, we present the L_1-based penalization estimator, and describe the proposed algorithm. In Section 3, we state technical assumptions and establish the asymptotic property of the proposed estimator as well as the modified BIC for tuning parameter selection. We assess the finite sample performance of the proposed method through simulation in Section 4 and a real data analysis in Section 5. Technical proofs are given in the online supplementary file.

2. Proposed Method

2.1. Penalized estimator

Throughout the paper, we denote $\mu_i \doteq \alpha_{k_i}$ as the intercept for subject i, where k_i is the unknown group membership, and let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$. We define the proposed penalized estimator of $(\boldsymbol{\mu}, \boldsymbol{\beta})$ as

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\beta}}) = \arg\min_{\boldsymbol{\mu},\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} |y_i - \mu_i - \boldsymbol{x_i}^T\boldsymbol{\beta}| + \sum_{1\le i<j\le n} p_\lambda(|\mu_i - \mu_j|), \tag{2.1}$$

where the first term on the right is the L_1 (least absolute deviation) loss function, and $p_\lambda(\dots)$ is the penalty function associated with a positive tuning parameter λ.

With the L_1 loss, one natural choice of the penalty function is the L_1 penalty $p_\lambda(\beta) = \lambda|\beta|$, for which the optimization can be easily solved by using linear programming. However, it is known that the L_1 penalty does not lead to consistent variable selection without proper assumptions; see Zhao and Yu (2006) and Leng, Lin and Wahba (2006). Particularly, the L_1 penalty tends to overshrink nonzero pairwise differences $|\mu_i - \mu_j|$, leading to an overestimation of the number of subgroups. On the other hand, adaptive weights can be incorporated in the L_1 penalty to reduce the bias (Zou (2006)), but the weights are difficult to estimate in our setup. Therefore, we consider two commonly used concave penalties, the smoothly clipped absolute deviations penalty (SCAD) of Fan and Li (2001), and the minimax concave penalty (MCP) of Zhang (2010), which can produce unbiased estimates and thus are more suitable for identifying subgroups. For a given $\lambda > 0$, the SCAD penalty is defined as

$$p_\lambda(|\beta|) = \lambda|\beta| I(0 \le |\beta| \le \lambda) + \frac{a\lambda|\beta| - (\beta^2+\lambda^2)/2}{a-1} I(\lambda \le |\beta| \le a\lambda) + \frac{(a+1)\lambda^2}{2} I(|\beta| > a\lambda), \text{ for some fixed } a > 2,$$

and the MCP penalty is defined as

$$p_\lambda(|\beta|) = \lambda\left(|\beta| - \frac{\beta^2}{2a\lambda}\right) I(0 \le |\beta| \le a\lambda) + \frac{a\lambda^2}{2} I(|\beta| > a\lambda),$$

for some fixed $a > 1$.

The penalty in (2.1) will shrink some of the pairs $\mu_i - \mu_j$ to zero. We can then partition the sample into subgroups based on the penalized estimator $\widehat{\mu}_j$. Let $\{\widehat{\alpha}_1, \dots, \widehat{\alpha}_{\widehat{K}}\}$ be the distinct values of $\widehat{\mu}_i$'s, where $\widehat{K}$ is the number of unique $\widehat{\mu}_i$'s. Let $\widehat{G}_k = \{i : \widehat{\mu}_i = \widehat{\alpha}_k, 1 \le i \le n\}, 1 \le k \le \widehat{K}$. Then $\{\widehat{G}_1, \dots, \widehat{G}_{\widehat{K}}\}$ constitutes a partition of $\{1, \dots, n\}$.

At a first glance, the proposed method with fused penalty looks similar to the pursuit of homogeneous covariate effects in regression settings, such as the fused Lasso in Tibshirani et al. (2005), the octagonal shrinkage in Bondell and Reich (2008), and the grouping pursuit in Shen and Huang (2010), to name a few. However, the aims are essentially different; we focus on identifying subgroups of

subjects with homogeneous intercept μ_i while these works used penalization to group p coefficients $(\beta_1, \ldots, \beta_p)$ to identify predictors with common effects. Our aim is also different from those in Raftery and Dean (2006), Gupta and Ibrahim (2007), Khalili and Chen (2007), which focused on the selection of variables for model-based clustering.

2.2. Basic computing algorithm

Local linear approximation. We propose an algorithm based on local linear approximation (**LLA**, Zou and Li (2008)) to minimize the objective function involving both L_1 loss and concave penalty. Specifically, we regard $\mu_i - \mu_j$ as an indivisible whole and approximate $\sum_{i<j} p_\lambda(|\mu_i - \mu_j|)$ by local linearization. Let μ_i^{t-1} denote the estimates of μ_i obtained at the $(t-1)$-th iteration. At the t-th iteration, we update the coefficients by solving

$$\arg\min_{\boldsymbol{\mu},\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} |y_i - \mu_i - \boldsymbol{x_i}^T\boldsymbol{\beta}| + \sum_{i<j} \omega_{ij}^{(t-1)}|\mu_i - \mu_j|, \tag{2.2}$$

where $\omega_{ij}^{(t-1)} = p'_\lambda(|\mu_i^{t-1} - \mu_j^{t-1}|) \geq 0$ denote the weights, and $p'_\lambda(\ldots)$ is the derivative of $p_\lambda(\ldots)$ with $p'_\lambda(0+)$ set as λ. The LLA algorithm is claimed to converge when the weights $\omega_{ij}^{(t)}$ stabilize, namely when $\sum_{1\leq i<j\leq n}(w_{ij}^{(t-1)} - w_{ij}^{(t)})^2$ becomes sufficiently small.

With the LLA, at each iteration, the optimization in (2.2) is a standard linear programming problem and thus can be easily solved with any existing linear programming algorithm. In our implementation, we use data augmentation to reformulate (2.2) as a simple weighted median regression problem.

Below we illustrate the idea of data augmentation to solve (2.2) with weights ω_{ij}. Denoting all the parameters as $\boldsymbol{\delta} = (\mu_1, \ldots, \mu_n, \beta_1, \ldots, \beta_p)^T$, we can rewrite the objective function in (2.2) equivalently as

$$\begin{aligned}
&\sum_{i=1}^{n} |y_i - \mu_i - \boldsymbol{x_i}^T\boldsymbol{\beta}| + \sum_{i<j} n\omega_{ij}|\mu_i - \mu_j| \\
&= \sum_{i=1}^{n} |y_i - (\boldsymbol{e_i}^T, \boldsymbol{x_i}^T)\boldsymbol{\delta}| + \sum_{i<j} |0 - (nw_{ij}(\boldsymbol{e_i} - \boldsymbol{e_j})^T, \mathbf{0}_p^T)\boldsymbol{\delta}|,
\end{aligned}$$

where $\boldsymbol{e}_i$ denotes a n-dimensional vector with the i-th element one and else zero. Let $\tilde{\mathbf{W}}$ denote a $n(n-1)/2 \times n$ matrix that consists of $(w_{ij}(\boldsymbol{e_i} - \boldsymbol{e_j})^T)$ for $1 \leq i < j \leq n$. That is, $\tilde{\mathbf{W}} = \big(w_{12}(\boldsymbol{e_1} - \boldsymbol{e_2}), \ldots, w_{1n}(\boldsymbol{e_1} - \boldsymbol{e_n}), w_{23}(\boldsymbol{e_2} - \boldsymbol{e_3}), \ldots, w_{n-1,n}(\boldsymbol{e_{n-1}} - \boldsymbol{e_n})\big)^T$. To minimize (2.2), we only need to fit median regression using the augmented dataset $\{(\tilde{y}_l, \tilde{\mathbf{x}}_l), l = 1, \ldots, n + n(n-1)/2\}$ with $(\tilde{y}_l, \tilde{\mathbf{x}}_l^T) =$

$(y_l, (\boldsymbol{e}_{\boldsymbol{l}}^T, \boldsymbol{x}_{\boldsymbol{l}}^T))$ for $l = 1, \ldots, n$, and $(\tilde{y}_l, \tilde{\mathbf{x}}_l^T) = (0, n\tilde{\mathbf{w}}_{l-n}^T, \mathbf{0}_p^T)$ for $l = n+1, \ldots, n+n(n-1)/2$, where $\tilde{\mathbf{w}}_j$ is the jth row of $\tilde{\mathbf{W}}$.

The augmented design matrix is sparse with many zeros even though the dimension may appear daunting. Therefore, we can solve the minimization problem by using the sparse Frisch-Newton interior algorithm, implemented by the "rq.fit.sfn" function in the R package *quantreg*, which reduces the computational time to be proportional to the number of nonzero elements in the design matrix. Our numerical investigation shows that the proposed LLA algorithm is much faster than the ADMM algorithm considered in Ma and Huang (2017) especially for models with heavy-tailed or covariate-dependent errors.

Choice of the tuning parameter. The tuning parameter λ controls the strength of penalization. In practice, we can choose λ by minimizing the following modified Bayesian information criterion,

$$BIC\{\hat{\boldsymbol{\delta}}(\lambda)\} = \log\left\{ n^{-1} \sum_{i=1}^{n} |y_i - \widehat{\mu}_i(\lambda) - \boldsymbol{x}_{\boldsymbol{i}}^T \widehat{\boldsymbol{\beta}}(\lambda)| \right\} + |\hat{S}_\lambda| \phi_n, \tag{2.3}$$

where $\hat{\boldsymbol{\delta}}(\lambda) = (\hat{\mu}_1(\lambda), \ldots, \hat{\mu}_n(\lambda), \hat{\boldsymbol{\beta}}(\lambda)^T)^T$ is the penalized estimator and $\hat{S}_\lambda$ is the resulting model associated with the tuning parameter λ, $|\hat{S}_\lambda| = \widehat{K}(\lambda) + p$ measures the size of the model with $\widehat{K}(\lambda)$ as the estimated number of subgroups, and ϕ_n is some positive sequence that goes to zero. Our numerical studies suggest that $\phi_n = c \log\log(n) \log(n+p)/n$ with $c \in [1, 5]$ provides a good choice. We shall establish the validity of this tuning parameter selector in Section 3.

Initial values. The optimization problem depends on the choice of initial values. We propose to use the following Lasso estimator as initial values,

$$\arg\min_{\boldsymbol{\mu}, \beta} \frac{1}{n} \sum_{i=1}^{n} |y_i - \mu_i - \boldsymbol{x}_{\boldsymbol{i}}^T \boldsymbol{\beta}| + \sum_{i<j} \lambda^* |\mu_i - \mu_j|, \tag{2.4}$$

where λ^* is a small tuning parameter to avoid over shrinkage. As discussed earlier, even though the Lasso estimator does not have nice properties of those based on concave penalties, its computation is much simpler since the estimator can be obtained directly by median regression through data augmentation without any iteration. Our numerical studies show that this Lasso estimator provides good initial values, which leads to quick convergence of the LLA algorithm.

Remark 1. This fused penalty term in the objective function (2.1) involves total $n(n-1)/2$ pairs of differences, so the basic algorithm can become computationally intensive for large samples. We suggest to modify it using a divide-and-conquer idea for massive data analysis. In the first stage, we divide the data into H

subsamples randomly. For each subsample, we apply the pairwise penalization algorithm with a small tuning parameter to cluster subjects into subgroups. In the second stage, we perform another pairwise penalization to further merge these subgroups identified in the first stage.

3. Asymptotic Properties

In this section, we will first establish the theoretical properties of the proposed pairwise penalized estimator. Under some regularity conditions, we show that the set of local minimizers of the proposed penalized objective function (2.1) covers the oracle estimator, obtained with known group membership as *a priori*, with probability approaching one.

Let $S_{\mathrm{o}} = \{G_k, k = 1, \ldots, K_0\}$ denote the true group structure, where G_k denotes the set of samples from group k and K_0 is the true number of groups. In the ideal case where S_{o} is known in advance, we can estimate $(\alpha_1, \ldots, \alpha_{K_0}, \boldsymbol{\beta})$ by the oracle estimator defined as

$$(\tilde{\alpha}_1(S_{\mathrm{o}}), \ldots, \tilde{\alpha}_{K_0}(S_{\mathrm{o}}), \tilde{\boldsymbol{\beta}}(S_{\mathrm{o}})) = \underset{\alpha_1, \ldots, \alpha_{K_0}, \boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n} \sum_{k=1}^{K_0} \sum_{i \in G_k} |y_i - \alpha_k - \boldsymbol{x_i}^T \boldsymbol{\beta}|, \tag{3.1}$$

where $\alpha_k = \mu_i$ for $i \in G_k$ is the common intercept for the kth group. Denote $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{K_0})^T$, $\mathbf{Z} = \{z_{ik}\}$ as a $n \times K_0$ matrix with $z_{ik} = 1$ for $i \in G_k$ and 0 otherwise, and $\mathbf{z}_i$ as the ith row of $\mathbf{Z}$. Then we can rewrite (3.1) as

$$(\tilde{\boldsymbol{\alpha}}(S_{\mathrm{o}}), \tilde{\boldsymbol{\beta}}(S_{\mathrm{o}})) = \underset{\boldsymbol{\alpha}, \boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} |y_i - \mathbf{z}_i^T \boldsymbol{\alpha} - \boldsymbol{x_i}^T \boldsymbol{\beta}|. \tag{3.2}$$

Consequently we can define the oracle estimator of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ as $\tilde{\boldsymbol{\mu}}(S_{\mathrm{o}}) = (\tilde{\mu}_1(S_{\mathrm{o}}), \ldots, \tilde{\mu}_n(S_{\mathrm{o}}))^T$ with $\tilde{\mu}_i(S_{\mathrm{o}}) = \tilde{\alpha}_k(S_{\mathrm{o}})$ for $i \in G_k$. The oracle estimator is denoted as $\tilde{\boldsymbol{\delta}}(S_{\mathrm{o}}) = (\tilde{\mu}_1(S_{\mathrm{o}}), \ldots, \tilde{\mu}_n(S_{\mathrm{o}}), \tilde{\boldsymbol{\beta}}(S_{\mathrm{o}})^T)^T$. In addition, let $G_{\min} = \min_{1 \le k \le K_0} |G_k|$ and $G_{\max} = \max_{1 \le k \le K_0} |G_k|$, where $|G_k|$ denotes the number of elements in group G_k, and $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$. We assume the following regularity conditions.

C1. (i) There exists some positive constants M_1, M_2 such that $|x_{ij}| \le M_1$, and $E(x_{ij}^4) \le M_2$, $\forall 1 \le i \le n, 1 \le j \le p$. (ii) There exists some positive constants C_1 and C_2 such that $C_1 \le \lambda_{\min}[n^{-1}(\mathbf{Z}, \mathbf{X})^T(\mathbf{Z}, \mathbf{X})] \le \lambda_{\max}[n^{-1}(\mathbf{Z}, \mathbf{X})^T(\mathbf{Z}, \mathbf{X})] \le C_2$, where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalues of a matrix, respectively.

C2. The conditional distribution of ε_i given $(\mathbf{z}_i, \mathbf{x}_i)$, denoted by $F_i(\cdot|\mathbf{z}_i, \mathbf{x}_i)$, has a continuous density $f_i(\cdot|\mathbf{z}_i, \mathbf{x}_i)$, which is uniformly bounded away from zero and

infinity in the neighborhood of zero across i.

C3. The number of parameters satisfies $p = O(n^{c_1})$ for some $0 \le c_1 < 1/3$.

C4. Let $b_n = \min_{i \in G_{k'}, j \in G_k, k' \neq k} |\mu_{0i} - \mu_{0j}|$ be the minimal difference of the common intercepts between two groups, where μ_{0i} is the true value of μ_i. There exist some positive constants c_2 and M_3 such that $2c_1 < c_2 \le 1$ and $n^{(1-c_2)/2} b_n \ge M_3$.

Assumption C1 poses some boundedness condition on the design. In our model setup, $\mathbf{Z}^T\mathbf{Z} = \text{diag}(|G_1|, \ldots, |G_{K_0}|)$. Note that $\lambda_{\min}[(\mathbf{Z}, \mathbf{X})^T(\mathbf{Z}, \mathbf{X})] \le \min\{\lambda_{\min}(\mathbf{Z}^T\mathbf{Z}), \lambda_{\min}(\mathbf{X}^T\mathbf{X})\}$ and $\lambda_{\max}[(\mathbf{Z}, \mathbf{X})^T(\mathbf{Z}, \mathbf{X})] \ge \max\{\lambda_{\max}(\mathbf{Z}^T\mathbf{Z}), \lambda_{\max}(\mathbf{X}^T\mathbf{X})\}$. Therefore, C1 (ii) indicates that $G_{\min}/n \ge C_1$ and $G_{\max}/n \le C_2$, which together with the fact that $K_0 G_{\min} \le n \le K_0 G_{\max}$ implies that $1/C_2 \le K_0 \le 1/C_1$. Condition C2 is standard in median regression, and it is more relaxed than the Gaussian and sub-Gaussian error condition assumed in Ma and Huang (2017). In C3, we assume that $p = O(n^{c_1})$, allowing p to increase with the sample size. Condition C4 requires the smallest signal not decay too fast, and similar conditions are commonly assumed in high dimensional sparse regression.

The following Theorem shows that the oracle estimator is a local minimizer of the proposed penalized objective function with probability approaching one. To account for the nonsmooth loss function and the nonsmooth and nonconvex penalty function involved in (2.1), we apply Lemma 2.1 in Wang, Wu and Li (2012), which gives a sufficient local optimization condition for the difference convex program based on the subdifferential calculus.

Theorem 1. *Let $\mathcal{B}_n(\lambda)$ be the set of local minimizers of* (2.1) *with either the MCP or SCAD penalty with tuning parameter λ. Suppose that conditions* C1-C4 *hold, $\lambda = o(n^{-(1-c_2)/2})$ and $n\lambda|G_{\min}| \to \infty$, then the oracle estimator $\widetilde{\boldsymbol{\delta}}(S_o)$ satisfies $P\{\widetilde{\boldsymbol{\delta}}(S_o) \in \mathcal{B}_n(\lambda)\} \to 1$ as $n \to \infty$.*

We next study the properties of the modified BIC for tuning parameter selection by establishing its consistency for model selection. For any candidate model S with K groups, we define the modified BIC as

$$BIC\{\tilde{\boldsymbol{\delta}}(S)\} = \log\left\{ n^{-1} \sum_{i=1}^{n} |y_i - \tilde{\mu}_i(S) - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}(S)| \right\} + |S|\phi_n, \tag{3.3}$$

where $\tilde{\boldsymbol{\delta}}(S) = (\tilde{\mu}_1(S), \ldots, \tilde{\mu}_n(S), \tilde{\boldsymbol{\beta}}(S)^T)^T$ is the unpenalized estimator obtained under model S. Here $BIC\{\tilde{\boldsymbol{\delta}}(S)\}$ is based on the unpenalized estimator obtained by minimizing the L_1 loss function under the candidate model S, while $BIC\{\hat{\boldsymbol{\delta}}(\lambda)\}$ in (2.3) is based on the penalized estimators obtained by minimiz-

ing (2.1) with the tuning parameter λ. Under conditions C1-C4, and additional conditions C2+ and C5 in the supplement, we obtain the following theorem.

Theorem 2. *Assume that* C1-C4*, and* C2+ *and* C5 *spelled in the online supplement hold. For any sequence* $\phi_n \to 0$ *satisfying* $\log(n+p)/n = o(\phi_n)$*, we have*

$$P\left(\inf_{S \neq S_o, |S| < K_U + p} BIC\{\tilde{\boldsymbol{\delta}}(S)\} > BIC\{\tilde{\boldsymbol{\delta}}(S_o)\}\right) \to 1,$$

where $K_U \in (K_0, \infty)$ *is the upper bound for the number of groups.*

Remark 2. At a given λ, let $\hat{S}_\lambda$ denote the model corresponding to the penalized estimator $\hat{\boldsymbol{\delta}}(\lambda)$. By definitions, $BIC\{\hat{\boldsymbol{\delta}}(\lambda)\} \geq BIC\{\tilde{\boldsymbol{\delta}}(\hat{S}_\lambda)\}$ since the penalized and unpenalized estimators correspond to the same model but the latter minimizes the L_1 loss function. In addition, Theorem 1 implies that, with high probability, the oracle estimator $\tilde{\boldsymbol{\delta}}(S_{\mathrm{o}})$ can be produced by some λ_{o} on the solution path, so $BIC\{\hat{\boldsymbol{\delta}}(\lambda_o)\} = BIC\{\tilde{\boldsymbol{\delta}}(S_{\mathrm{o}})\}$. Therefore, by Theorem 2, for any λ not inducing the oracle model, we have $BIC\{\hat{\boldsymbol{\delta}}(\lambda)\} \geq BIC\{\tilde{\boldsymbol{\delta}}(\hat{S}_\lambda)\} > BIC\{\tilde{\boldsymbol{\delta}}(S_{\mathrm{o}})\} = BIC\{\hat{\boldsymbol{\delta}}(\lambda_o)\}$. This suggests that the modified BIC in (2.3) is consistent for tuning parameter selection.

4. Simulation

In this section, we use three examples to assess the finite-sample performance of the proposed method based on the SCAD penalty with $a = 3.7$. The method with MCP gives similar results and thus is omitted. For comparison, we also include the mean-based penalization method from Ma and Huang (2017) based on the SCAD penalty and ADMM algorithm. We consider four different metrics: (1) MAE_μ: the mean absolute error for the intercept estimate, defined by $\mathrm{MAE}_\mu = n^{-1}\sum_{i=1}^n |\widehat{\mu}_i - \mu_i|$; (2) MAE_β: the mean absolute error for the slope estimate, defined by $\mathrm{MAE}_\beta = \sum_{j=1}^p |\widehat{\beta}_j - \beta_j|/p$; (3) $\bar{\hat{K}}$ and $\tilde{\hat{K}}$: the average and median number of identified subgroups across simulation, respectively; and (4) RI: the rand index. The rand index is commonly used in clustering analysis to measure the percentage of correct decisions of a clustering algorithm, and is defined as

$$RI = \frac{TP + TN}{TP + FP + FN + TN},$$

where TP (true positive) means the number of pairs of subjects in different subgroups that are assigned to different clusters, TN (true negative) denotes the number of pairs from the same subgroup that are assigned to the same cluster, FN (false negative) denotes the number of pairs from the same subgroup that

are assigned to different clusters, and FP (false positive) is the number of pairs from different subgroups that are assigned to the same cluster. Higher values of the rand index indicate better agreement of the identified clusters with the true group allocation. For all examples, the simulation is repeated 100 times.

Example 1. The data are generated from $y_i = \mu_i + \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i$, $i = 1, \ldots, n = 100$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{i5})^T$, $\mu_i = 1$ with probability $\pi_i = \exp(-0.5x_{i1} - 0.5x_{i6})$, $\mu_i = -1$ with probability $1 - \pi_i$, and $\beta_j = 1$ for $j = 1, \ldots, 5$. The covariates x_{ij} are generated independently from the standard normal distribution as well as x_{i6}. We consider three cases for generating ε_i. Case 1 (homoscedastic normal): $\varepsilon_i = 0.5\epsilon_i$ with $\epsilon_i \overset{i.i.d.}{\sim} N(0, 1)$. Case 2 (heavy-tailed): $\varepsilon_i \overset{i.i.d.}{\sim} 0.5t(3)$. Case 3 (heteroscedastic normal): $\varepsilon_i = \Phi(x_{i1})\epsilon_i$ with $\epsilon_i \overset{i.i.d.}{\sim} N(0, 1)$, where $\Phi(\cdot)$ is the distribution function of $N(0, 1)$.

Table 1 summarizes the simulation results of the proposed median method and the mean method in Ma and Huang (2017) in Cases 1–3 from Example 1. For the median regression, we use $\lambda^* = O(n^{-3/2})$ to obtain the initial values, and choose the penalization parameter λ by minimizing the BIC in (2.3) with $\phi_n = c \log\log(n) \log(n + p)/n$. We report the results for the median method based on $c = 1, 5$ and 10. The last two columns of Table 1 give the average computing time (in seconds) of different methods with the chosen λ, using R (version 3.3.2) on a 2.70GHz laptop, and the average number of iterations needed for convergence. Quantities in parentheses denote the standard errors and those in square brackets denote the ranges. For Case 1 with homoscedastic normal errors, the mean method performs slightly better than the median method in terms of RI and MAE. However, for models with heavy-tailed errors (Case 2) and heteroscedastic errors (Case 3), the median-based method shows clear advantages; it gives competitive RI and $\hat{K}$ closer to the truth, while the mean method often leads to larger models. In addition, the proposed algorithm is computationally much more efficient than the ADMM algorithm in Ma and Huang (2017). Our numerical study shows that in general the median method with $c \in [1, 5]$ gives quite consistent results, while $c = 10$ tends to underestimate K leading to lower RI. Therefore, we focus on $c = 5$ in the following analyses.

Example 2. We consider a setting with three subgroups. The data is generated from $y_i = \mu_i + \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i, i = 1, \ldots, n = 150$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{i5})^T$ and $\boldsymbol{\beta}$ are the same as in Example 1. Let $\mu_i = -2$ with probability $\pi_{i1} = \exp(-0.5x_{i1} - x_{i6})$, $\mu_i = 0$ with probability $\pi_{i2} = \exp(-0.5x_{i2} - x_{i7})$ and $\mu_i = 2$ with probability $\pi_{i3} = 1 - \pi_{i1} - \pi_{i2}$, where x_{i6} and x_{i7} are independent standard normal random

Table 1. Estimation results of the median and mean methods for Cases 1–3 in Example 1 with different choices of c used in BIC for determining the penalization parameter.

Case	Method	RI	$\bar{\hat{K}}$	$\tilde{\hat{K}}$	MAE_μ	MAE_β	Time (s)	Iteration
1	Median	0.83	2.09	2	0.29	0.10	1.38	3.64
	($c = 1$)	(0.07)	(0.32)		(0.11)	(0.04)	[0.98,3.00]	[3,8]
	Median	0.83	2.00	2	0.29	0.10	1.36	3.53
	($c = 5$)	(0.07)	(0.14)		(0.11)	(0.04)	[1.03,1.92]	[3,4]
	Median	0.75	1.77	2	0.47	0.12	1.42	3.67
	($c = 10$)	(0.17)	(0.48)		(0.34)	(0.06)	[1.02,2.01]	[3,5]
	Mean	0.87	2.49	2	0.25	0.09	13.67	60.59
	($c = 5$)	(0.07)	(0.75)		(0.13)	(0.04)	[7.81,22.84]	[36,100]
2	Median	0.78	2.12	2	0.35	0.11	1.42	3.79
	($c = 1$)	(0.07)	(0.43)		(0.12)	(0.05)	[1.01,2.53]	[3,7]
	Median	0.78	2.00	2	0.35	0.11	1.36	3.66
	($c = 5$)	(0.07)	(0.14)		(0.12)	(0.05)	[1.02,2.35]	[3,6]
	Median	0.66	1.51	1.50	0.64	0.14	1.49	3.88
	($c = 10$)	(0.16)	(0.52)		(0.36)	(0.07)	[1.03,2.25]	[3,6]
	Mean	0.77	4.79	5	0.43	0.10	18.02	80.19
	($c = 5$)	(0.07)	(1.35)		(0.15)	(0.04)	[8.41,23.33]	[38,100]
3	Median	0.83	2.08	2	0.28	0.07	1.40	3.62
	($c = 1$)	(0.07)	(0.39)		(0.11)	(0.04)	[1.04,2.92]	[3,8]
	Median	0.83	2.02	2	0.28	0.07	1.39	3.59
	($c = 5$)	(0.07)	(0.20)		(0.11)	(0.04)	[1.03,2.98]	[3,8]
	Median	0.81	1.91	2	0.34	0.08	1.39	3.62
	($c = 10$)	(0.12)	(0.38)		(0.25)	(0.06)	[1.04,3.06]	[3,8]
	Mean	0.85	3.21	3	0.30	0.09	14.07	62.76
	($c = 5$)	(0.07)	(1.11)		(0.11)	(0.03)	[8.43,23.11]	[37,100]

RI: rand index; $\hat{K}$ and $\tilde{\hat{K}}$: the average and median number of identified subgroups; MAE_μ and MAE_β: the mean absolute error for the intercept and slope estimates; Time: the average computing time in seconds; Iteration: the average of number iterations needed for convergence. Quantities in parentheses denote the standard errors and those in square brackets denote the ranges.

variables. We consider three cases as in Example 1. Results in Table 2 also suggest that the median method outperforms the mean method for both heavy-tailed and heteroscedastic cases.

5. Empirical Study

In this section, we compare the performance of the proposed median method and the mean method in Ma and Huang (2017) by analyzing the Cleveland Heart Disease Dataset from the UCI machine learning repository. The dataset contains

Table 2. Estimation results of the mean and median methods for three cases in Example 2.

Case	Method	RI	$\bar{\hat{K}}$	$\tilde{\hat{K}}$	MAE_μ	MAE_β	Time (s)	Iteration
1	Median	0.87 (0.06)	3.14 (0.51)	3	0.33 (0.16)	0.10 (0.05)	7.12 [4.61,11.03]	4.45 [3,7]
	Mean	0.88 (0.06)	3.71 (1.10)	3	0.33 (0.16)	0.10 (0.05)	29.48 [20.87,36.46]	82.52 [59,100]
2	Median	0.81 (0.06)	3.19 (0.63)	3	0.49 (0.18)	0.13 (0.06)	7.16 [4.50,12.57]	4.64 [3,7]
	Mean	0.80 (0.06)	6.26 (1.94)	6	0.53 (0.18)	0.13 (0.06)	32.64 [22.74,36.32]	92.51 [64,100]
3	Median	0.86 (0.06)	3.17 (0.47)	3	0.34 (0.16)	0.08 (0.06)	6.76 [4.24,10.99]	4.59 [3,7]
	Mean	0.86 (0.05)	4.57 (1.44)	4	0.37 (0.14)	0.10 (0.05)	30.56 [20.72,36.29]	86.43 [58,100]

The notations follow Table 1.

13 clinical measurements on 297 individuals. The outcome of interest is thalach, the maximum heart rate achieved. As in Ma and Huang (2017), we use the fitted value of thalach as the response variable y, obtained by projecting thalach onto the linear space spanned by the variables including Chestpt (chest pain type), Exeriai (exercise induced angina indicator), STd (ST depression induced by exercise relative to rest), SlopeST (slope of the peak exercise ST segment), Numvess (the number of major vessels colored by fluoroscopy) and Hrtstat (the heart status). We aim to identify subgroups in the response distribution after adjusting for the effect of the remaining six covariates $\mathbf{x}$: Sex (0 for female), age in years, Restbps (resting blood pressure), Chol (serum cholesterol), Fbs (fasting blood sugar indicator) and Restecg (resting electrocardiographic results with 0 for normal). Prior to the data analysis, we centralize the four continuous covariates to have mean zero so that the intercept in model (1.1) corresponds to the median of a female with normal Restecg and average age, Restbps, Chol and Fbs.

Figure 1 shows the grouping results of the median and mean methods with varying penalization parameter λ for 297 subjects. Different color represents different subgroup membership. The mean regression identifies five subgroups with $\lambda = 0.04$. When λ increases, the mean method leads to subgroups with one dominating subgroup and other subgroups consisting of a few individuals, which are likely to be superficial and make the results hard to interpret. In contrast, the median regression identifies four subgroups with $\lambda = 0.1$. When λ is increased

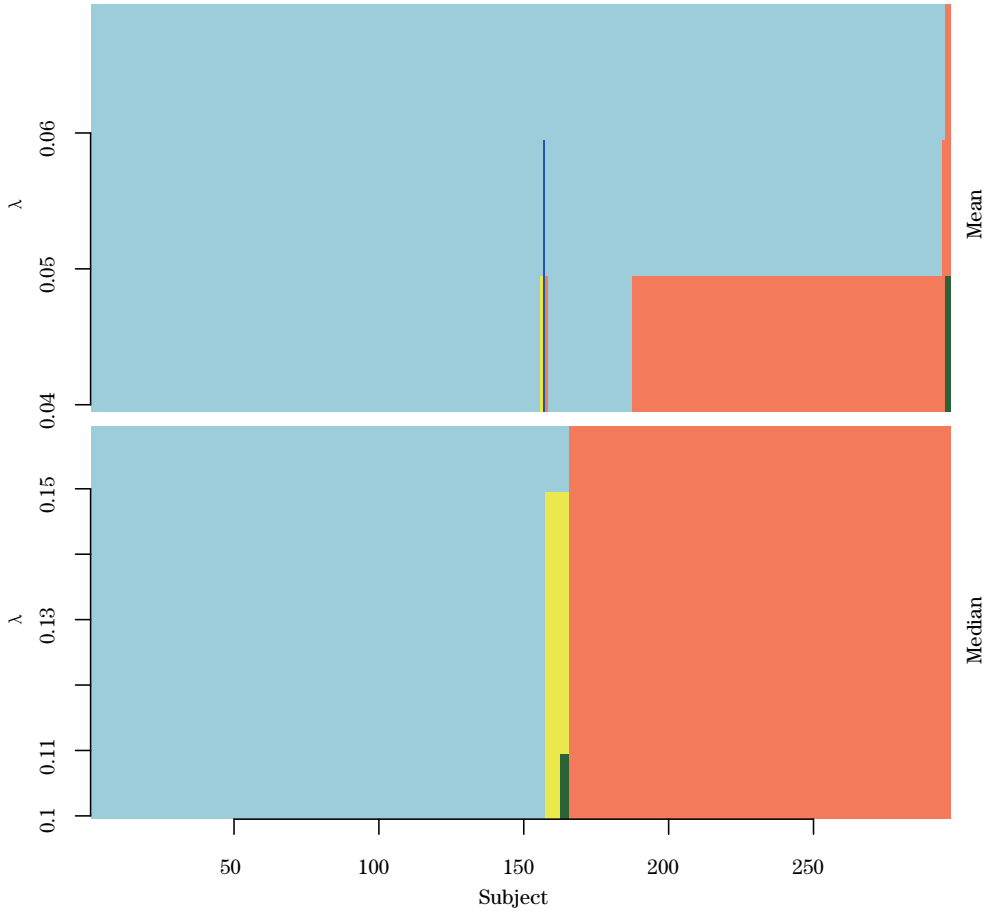


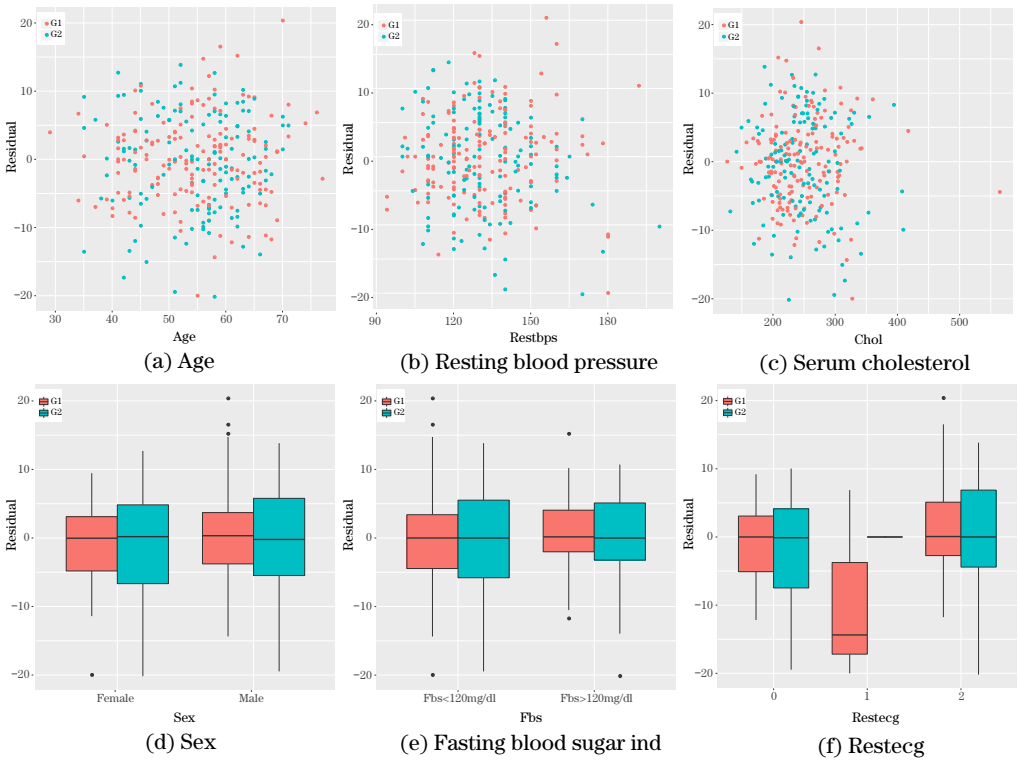
Figure 1. Grouping results of the 297 subjects in the Cleveland heart disease study obtained by the median and the mean methods with varying penalization parameter λ. Different color represents different subgroup membership.

to 0.15, the median regression leads to two subgroups of sizes 165 and 132.

We further assess the heteroscedasticity based on the subgroup identification results from the median regression with $\lambda = 0.15$ that leads to two subgroups. Letting the group indicator $d_i = 0$ for group 1 and $d_i = 1$ for group 2, we fit the following regression model,

$$y_i = \alpha_1 + \theta d_i + \mathbf{x}_i^T \boldsymbol{\beta} + e_i, \tag{5.1}$$

and assess the error heteroscedasticity by applying the Breusch-Pagan test. The Breusch-Pagan test is a chi-squared-type test based on regressing the squared



(a) Age (b) Resting blood pressure (c) Serum cholesterol

(d) Sex (e) Fasting blood sugar ind (f) Restecg

Figure 2. Plots of the estimated residuals from model (5.1) against covariates.

residuals from fitting model (5.1) against covariates $(d_i, \mathbf{x}_i)$. The resulting Breusch-Pagan test statistic is 19.29 with a p-value of 0.01, suggesting a strong evidence of error heteroscedasticity. More specifically, Figure 2 shows that the error variance tends to depend on the rest blood pressure, and it varies between two identified subgroups. As shown in our simulation study, the mean penalization method often overestimates the number of subgroups for heteroscedastic models, and this agrees with the observation in this empirical study.

By fitting model (5.1) at median, we obtain the estimated subgroup effect as $\hat{\theta} = 20.72$, and the 95% score-type confidence interval as $(19.40, 23.08)$, suggesting that the first subgroup has a significantly smaller median than the second subgroup after accounting for the covariate effects.

Finally, to characterize the two identified subgroups, we fit a logistic regression by regressing d_i against the 12 available variables. We apply the SCAD penalized logistic regression method from the R package "ncvreg" with tuning parameter selected by cross-validation, and report the coefficient estimations and

Table 3. Characterization of the two clusters identified by the median method: the estimated coefficients of selected variables and standard errors in the logistic regression.

Variable	Slope STup	Chestpt3	Chestpt4	Restecg1	Restecg2	Sex	Exeriai	Numvess	Age
Coef	20.60	−4.34	−13.37	26.21	2.24	4.61	−11.64	−3.95	6.44
SE	4.91	1.54	3.38	6.54	0.96	1.40	2.88	0.99	1.62

SlopeSTup: the slope of the peak exercise ST segment is upsloping; Chestpt3: nonanginal chest pain type; Chestpt4: asymptomatic chest pain type; Restecg1: normal resting electrocardiographic results; Restecg2: having ST-T wave abnormality; Sex: 1 for male; Exeriai: exercise induced angina indicator; Numvess: the number of major vessels colored by fluoroscopy.

standard errors for the selected variables in Table 3. Results suggest that subjects with nonanginal and asymptomatic chest pain, exercise-induced angina and more major vessels colored by fluoroscopy are more likely assigned to group 1 (with lower thalach), while older males with normal or ST-T wave abnormality in the resting electrocardiographic results and up-slope of the peak exercise ST segment are more likely assigned to group 2.

Supplementary Materials

Proofs for Theorems 1 and 2 are provided in the online supplementary file.

Acknowledgment

The authors would like to thank the Editor, an associate editor, and two anonymous reviewers for their constructive comments that have significantly improved the paper. The research was partly supported by National Science Foundation grants DMS-1149355 and DMS-1712760, the OSR-2015-CRG4-2582 grant from KAUST, the National Natural Science Foundation of China grants 11671096, 11731011 and 11690013, and a fellowship from CSC (China Scholarship Council).

References

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64**, 115–123.

Chaganty, A. T. and Liang, P. (2013). Spectral experts for estimating mixtures of linear regressions. arXiv preprint arXiv:1306.3729.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics* **37**, 2523–2542.

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

Dusseldorp, E., Conversano, C. and Van Os, B. J. (2010). Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical Statistics* **19**, 514–530.

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.

Gupta, M. and Ibrahim, J. G. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Journal of the American Statistical Association* **102**, 867–880.

Hall, P., Titterington, D. M. and Xue, J. H. (2009). Median-based classifiers for high-dimensional data. *Journal of the American Statistical Association* **104**, 1597–1608.

Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **58**, 155–176.

Hocking, T. D., Joulin, A., Bach, F. and Vert, J. P. (2011). Clusterpath an algorithm for clustering using convex fusion penalties. *The 28th International Conference on Machine Learning*.

Kasahara, H. and Shimotsu, K. (2015). Testing the number of components in normal mixture regression models. *Journal of the American Statistical Association* **110**, 1632–1645.

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

Leng, C., Lin, Y. and Wahba, G. (2006). A note on the Lasso and related procedures in model selection. *Statistica Sinica* **16**, 1273–1284.

Li, P. and Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association* **105**, 1084–1092.

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

McNicholas, P. D. (2010). Model-based classification using latent Gaussian mixture models. *Journal of Statistical Planning and Inference* **140**, 1175–1181.

Müllensiefen, D., Hennig, C. and Howells, H. (2017). Using clustering of rankings to explain brand preferences with personality and socio-demographic variables. *Journal of Applied Statistics* **17**, 1–21.

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

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association* **101**, 168–178.

Shen, X. and Huang, H. C. (2010) Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association* **105**, 727–739.

Shen, J. and He, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association* **110**, 303–312.

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

Methodology) **67**, 91–108.

Wang, L., Wu, Y. and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association* **107**, 214–222.

Wei, S. and Kosorok, M. R. (2013). Latent supervised learning. *Journal of the American Statistical Association* **108**, 957–970.

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

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7**, 2541–2563.

Zhu, H. T. and Zhang, H. (2004). Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 3–16.

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

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36**, 1509–1533.

Department of Statistics, School of Management, Fudan University, Shanghai 200433, China.

E-mail: 13210690005@fudan.edu.cn

Department of Statistics, George Washington University, Washington, District of Columbia 20052, USA.

E-mail: judywang@email.gwu.edu

Department of Statistics, School of Management, Fudan University, Shanghai 200433, China.

E-mail: zhuzy@fudan.edu.cn

(Received April 2017; accepted January 2018)