Statistica Sinica Preprint No: SS-2021-0061	
Title	Hybrid Hard-Soft Screening for High-dimensional Latent Class Analysis
Manuscript ID	SS-2021-0061
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0061
Complete List of Authors	Wei Dong, Xingxiang Li, Chen Xu and Niansheng Tang
Corresponding Author	Niansheng Tang
E-mail	nstang@ynu.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

HYBRID HARD-SOFT SCREENING FOR HIGH DIMENSIONAL LATENT CLASS ANALYSIS

Wei Dong[1], Xingxiang Li[2], Chen Xu[3] and Niansheng Tang[1]

[1] *Yunnan University,* [2] *Xi'an Jiaotong University and* [3] *University of Ottawa*

Abstract: Latent class analysis (LCA) is a powerful tool for detecting unobservable subgroups within a population. When a large number of covariates (features) are considered, an LCA faces great challenges in terms of both classification accuracy and computational efficiency. In this paper, we propose a novel feature screening procedure that eliminates most irrelevant features before an LCA is conducted. The proposed method is built on an EM-based hybrid hard-soft thresholding update (HHS-EM) of the latent class parameters, which naturally accounts for the joint effects between features. We show that the HHS-EM enjoys the sure screening property and leads to a refined LCA that is effective and consistent for high-dimensional classification. The performance of the proposed method is illustrated by means of simulation studies and a real–data example.

Key words and phrases: Feature screening, High-dimensional classification, Latent class analysis, Misclassification error, Sure joint screening.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

1. Introduction

High-dimensional data sets containing unobservable subgroups are common in fields such as the biomedical, social, behavioral, and economic sciences. Here, researchers often wish to discover the latent subgroups (classes) within a population, and to classify new cases based on input data. In the literature, this scientific task is typically conducted using a latent class analysis (LCA), which attempts to reveal hidden data sub-groups using a finite mixture model with the group membership influenced by the input features (e.g., Khalili (2010); Ghosh et al. (2011); Weller et al. (2020)). To improve the classification accuracy, researchers may consider a large number of covariates (features) at the initial stage of modeling. For example, in community and crime studies, sociologists are often interested in identifying the latent profiles of a target population using criminal activities that may be linked to hundreds of community indices, such as household income, school engagement, insurance coverage, food insufficiency, and so forth. However, when the number of features p is large, an LCA faces simultaneous methodological and computational challenges, owing to the curse of dimensionality.

To cope with the large p situation, it is often reasonable to assume that only a handful of features are relevant to the analysis. Many feature selection and assessment methods have been developed for LCAs based on this sparsity

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

assumption. For example, Houseman et al. (2006) proposed a feature-specific penalized LCA for genomic data, and Wu (2013) proposed a sparse LCA for clustering large-scale discrete data, Zhang and Ip (2014) proposed a feature assessment method for LCAs based on their discriminative power. See Fop and Murphy (2018) for a comprehensive review of feature selection in LCAs and model-based clustering. However, although the aforementioned methods are helpful in choosing relevant features, most of them lack solid theoretical support. Moreover, they are mainly designed for situations in which p is moderate (fixed p design); thus, they might not be suitable for a high-dimensional LCA with a large p, owing to the computational burden and algorithm instability.

To ease the implementation difficulties for a high-dimensional LCA, a natural strategy is to remove most irrelevant features before conducting the LCA. Such a strategy is referred to as feature screening. Reducing the dimensionality drastically reduces the associated analytical difficulties. For example, Fan and Lv (2008) proposed the sure independent screening (SIS) procedure and its iterated version (ISIS) for linear models, and Wang (2009) used a forward regression to sequentially screen features. Fan and Song (2010) extended the SIS (ISIS) procedure to generalized linear models using the maximum marginal likelihood estimation (MMLE). Xu and Chen (2014) proposed

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

a sparsity-restricted screening procedure for generalized linear models. Cui, Li, and Zhong (2015) advocated a mean-variance (MV) screening procedure for ultrahigh-dimensional discriminant analyses, Li et al. (2020) proposed a distributed screening framework for the divide-and-conquer setup, and Xie et al. (2020) and Tang et al. (2021) developed category-adaptive and quantile correlation-based variable selection methods, respectively. Refer to Liu et al. (2015) for an overview of feature screening.

However, despite the rich body of literature on feature screening, existing methods are based mainly on a direct measurement of the correlation between the response and the data features. Thus, they may not be directly applicable to an LCA, where the response is implicitly linked to the features by unobserved class labels. We propose a new feature screening approach for high-dimensional LCAs. The proposed method is built on a hybrid hard-soft EM thresholding procedure (HHS-EM) that attempts to find an approximate estimate of the LCA coefficients with a user-specified sparsity κ. By setting $\kappa \ll p$, the obtained estimate readily serves as an LCA feature screener. The HHS-EM procedure efficiently improves the mixture likelihood, naturally accounting for the joint effects between features. Using the hybrid hard-soft thresholding, we are able to reach a good balance between sparsity and convexity in the numerical updates, such that the procedure is less sensitive to

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

the choice of initial values and the overall algorithmic stability is enhanced. We show the proposed procedure enjoys the sure screening property and leads to a refined LCA within the given sparsity constraint. The refined LCA preforms an effective classification for both training and test data, as if the true model were known in advance. We use extensive numerical examples to demonstrate the promising performance of the HHS-EM procedure.

The proposed HHS-EM screening procedure is inspired from solving a hybrid L_0-L_1 constrained optimization, the idea of which shares some similarity with the trimmed Lasso discussed in (Bertsimas et al., 2017). Although both methods are built on a hybrid penalization technique, they are designed under different model setups, serve different research purposes, and are studied from different angles. Specifically, the trimmed Lasso was proposed as a robust estimation method for the least squares problem. The work of (Bertsimas et al., 2017) on the trimmed Lasso focuses on finding its robustness interpretation and showing its connection to other penalized methods. Their work was conducted mainly from an optimization perspective, and did not study any statistical properties. Here, we propose the HHS-EM procedure for feature screening in a high-dimensional LCA, where accurate selection by penalized methods is often difficult to achieve (Fan and Lv (2008)). Although the proposed procedure is inspired by an optimization problem, our primary

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

interest lies in the screening accuracy and misclassification errors, rather than the optimization itself.

The rest of this paper is organized as follows. Section 2 reviews the LCA and introduces the HHS-EM procedure. We investigate the theoretical properties of the proposed procedure in Section 3. To justify the HHS-EM numerically, we present our simulation results in Section 4 and a real–data example in Section 5. Section 6 concludes the paper. All figures, tables, and proofs are presented in the online Supplementary Material.

2. HHS-EM feature screening

2.1 Model setup and notation

We consider the following latent class model with Q classes:

$$\boldsymbol{y}_i|\boldsymbol{\omega}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma} \overset{\text{ind}}{\sim} \sum_{m=1}^{Q} \omega_{im}\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \quad \text{for} \quad i = 1, \ldots, n,$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ie})^\top$ is an $e \times 1$ vector of continuous responses for the ith subject, and $\boldsymbol{\mu}_m = (\mu_{m1}, \ldots, \mu_{me})^\top$ and $\boldsymbol{\Sigma}_m$ represent the mean vector and covariance matrix, respectively, corresponding to the mth class. Here, $\boldsymbol{\omega}_i = (\omega_{i1}, \ldots, \omega_{iQ})^\top$ is a $Q \times 1$ vector of binary latent allocation variables, where $\omega_{im} = 1$ if the ith subject belongs to class m, and zero otherwise, for $m = 1, \ldots, Q$. We require $\sum_{m=1}^{Q} \omega_{im} = 1$. Furthermore, we use $\mathcal{N}(\cdot, \cdot)$ to denote

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

the normal distribution, and define $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_Q\}$, $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_Q\}$.

Prior works (e.g., Ghosh et al. (2011)) often assume that the latent class indicator variable vector $\boldsymbol{\omega}_i$ follows a multinomial distribution $\mathcal{MN}(1, \boldsymbol{\pi}_i)$, where $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iQ})^\top$, and the class probabilities $\pi_{im} = \Pr(\omega_{im} = 1)$ satisfy $\pi_{im} > 0$ and $\sum_{m=1}^{Q} \pi_{im} = 1$. In an LCA, the response $\boldsymbol{y}_i$ is implicitly linked to a set of p covariates (features) $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, which affects the corresponding class probability π_{im} with the following presumed form:

$$\pi_{im} = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_m)}{\sum_{k=1}^{Q} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}, \tag{2.1}$$

where $\boldsymbol{\beta}_m$ is a $p \times 1$ vector of LCA coefficients corresponding to class m, for $m = 1, \ldots, Q$. When all $\boldsymbol{\beta}_m$ are zero, none of the features are relevant to the analysis, and the above latent class model reduces to a finite Gaussian mixture model (Khalili and Chen (2007)). Following common practice, we set $\boldsymbol{\beta}_Q = \mathbf{0}$ and assume Q is known. In applications, Q is often specified using prior information and a pre-analysis of the data. Throughout this paper, we assume that the number of features p grows exponentially with the sample size n, that is, $\log(p) = O(n^\tau)$, for some constant $0 \leq \tau < 1$, but that only a small number of features have important effects on the latent classes. This amounts to assuming that the coefficient $\boldsymbol{\beta}_m$ has a sparse structure with many zero entries. Note that in model (2.1), $\boldsymbol{\beta}_m$ may vary across classes. Consequently, the probability of an object belonging to different latent classes may be related

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

to different sets of relevant features.

Let $s_m^* = \{j : \beta_{mj} \neq 0 \text{ for } j = 1, \ldots, p\}$ be the index set of the relevant features associated with class m, where β_{mj} is the jth component of β_m, and let $s_m^{c*} = \{1, \ldots, p\} \backslash s_m^*$ be the index set of the irrelevant features related to class m, for $m = 1, \ldots, Q-1$. Denote $s^* = \{s_1^*, \ldots, s_{Q-1}^*\}$, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{Q-1}\}$, $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, and $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$. Suppose that the cardinality of the true model s^* is $\|s^*\|_0 = \sum_{m=1}^{Q-1} \|s_m^*\|_0 = q < \kappa$, for some known κ.

The log-likelihood function of $\boldsymbol{\Theta}$ for the observed data $\boldsymbol{Y}$ is given by

$$\ell_n(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log \left\{ \sum_{m=1}^{Q} \pi_{im} \phi(\boldsymbol{y}_i; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \right\}, \tag{2.2}$$

where $\phi(\cdot; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ denotes the Gaussian density with mean vector $\boldsymbol{\mu}_m$ and covariance matrix $\boldsymbol{\Sigma}_m$. The aforementioned latent class model commonly suffers from the identifiability problem. That is, without further restrictions on the class parameters $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$, the above latent class model with different class parameters may lead to the same model. To address this problem, we adopt the widely used approach of imposing restrictions on the class means. Thus, we assume $\mu_{11} < \mu_{21} < \cdots < \mu_{Q1}$. For further information on the identifiability problem, see Frühwirth-Schnatter (2006).

The goal of this study is to screen out most of the irrelevant features associated with the zero-effect LCA coefficients from model (2.1) by analyzing

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

the observed data ($\boldsymbol{Y}$, $\boldsymbol{X}$).

2.2 Feature screening in an LCA

Our idea stems from the sparsity constraint optimization, which has attracted considerable attention in recent years (Xu and Chen (2014), Yang et al. (2016), Qu et al. (2021)). With an L_0 penalty specifying the number of features allowed in the model, this method attempts to roughly estimate a few of the most significant coefficients from the full model, while setting all other coefficients to zero. Because the estimation is carried out on the full model, the resulting sparse estimator readily serves as a feature screener, naturally taking the joint effects among features into account.

Despite the success of the L_0-based approaches, few studies have examined LCAs with a focus on clustering. In contrast to a regular regression model, the response in model (2.1) is implicitly linked to features via unobservable class labels, where relevant features may vary across different classes. This complication makes the L_0-based approaches less effective in practice, owing to the algorithm instability. One natural strategy is to use a modified L_0 penalty that helps to improve the stability, while maintaining high screening accuracy.

As such, we consider the following penalized likelihood problem for model

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

(2.1):

$$\begin{aligned} \ell_{n;\lambda}(\boldsymbol{\Theta}) &= \ell_n(\boldsymbol{\Theta}) - n\lambda \sum_{m=1}^{Q-1} \|\boldsymbol{\beta}_m\|_1 \\ &\text{subject to } \|\boldsymbol{\beta}\|_0 \leq \kappa, \end{aligned} \tag{2.3}$$

where $\lambda > 0$ is a regularization parameter and κ is the user-specified model sparsity. Problem (2.3) can be viewed as a hybrid L_0-L_1 regularized estimation. With $\lambda = 0$, it reduces to the L_0-sparse estimation, and with $\kappa \geq (Q-1)p$, it reduces to a Lasso-type estimation (Tibshirani (1996)). The benefit of using this hybrid penalty is clear: it attempts to combine the strength of L_0 and L_1 penalties to achieve more effective screening in an LCA. With the L_0 penalty, users can precisely control the model sparsity using κ. When κ is properly chosen, the resulting sparse estimator suggests no more than κ coefficients that receive the most support from the likelihood and, thus, it readily serves for feature screening. With the L_1 penalty, we enhance the problem convexity, and thus improve the overall algorithmic stability in the updating process.

From a computational perspective, the optimization problem (2.3) can be represented as the following proposition.

Proposition 1. *The optimization problem (2.3) and the optimization problem*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

$$\max_{\boldsymbol{\Theta},\boldsymbol{z}} \left\{ \ell_n(\boldsymbol{\Theta}) - n\lambda \sum_{m=1}^{Q-1} \|\boldsymbol{\beta}_m\|_1 - n\eta \sum_{m=1}^{Q-1} \langle \boldsymbol{z}_m, |\boldsymbol{\beta}_m| \rangle \right\}$$
$$s.t. \sum_{m=1}^{Q-1} \sum_{j=1}^{p} z_{mj} = p(Q-1) - \kappa, \quad \boldsymbol{z}_m = (z_{m1}, \ldots, z_{mp})^\top \in \{0,1\}^p \tag{2.4}$$

have the same optimal objective value and the same set of optimal solutions $\hat{\boldsymbol{\Theta}}$ *when* $\eta > (\max_j \|\boldsymbol{x}_j\|_2)/n$, *where* $\boldsymbol{z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{Q-1}\}$.

Proposition 1 indicates that a solution to the optimization problem (2.3) can be obtained by solving (2.4) as a surrogation. Thus, the L_0-L_1 hybrid estimator may give fewer penalties on the LCA coefficients associated with important features for the desired level of sparsity.

Following common practice for LCAs, an EM algorithm can be developed to evaluate the maximum likelihood estimates of the model parameters. To this end, it follows from Khalili and Chen (2007) that the complete data log-likelihood function of $\boldsymbol{\Theta}$ for $\boldsymbol{Y}$ and $\boldsymbol{\omega} = \{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n\}$ can be written as

$$\ell_n^c(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \sum_{m=1}^{Q} \omega_{im} [\log \pi_{im} + \log\{\phi(\boldsymbol{y}_i; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)\}].$$

Thus, the corresponding penalized complete data log-likelihood function has the form

$$\tilde{\ell}_{n;\lambda}(\boldsymbol{\Theta}) = \ell_n^c(\boldsymbol{\Theta}) - n\lambda \sum_{m=1}^{Q-1} \|\boldsymbol{\beta}_m\|_1,$$

subject to $\|\boldsymbol{\beta}\|_0 \leq \kappa$. Similarly to Khalili and Chen (2007), the EM algorithm

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

for evaluating the MLE of $\boldsymbol{\Theta}$ is implemented as follows. At the tth iteration with a current value $\boldsymbol{\Theta}^{(t)}$ of $\boldsymbol{\Theta}$, we iteratively update the following two steps:

E-step. Compute the conditional expectation $Q_{n;\lambda}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(t)}) = \mathrm{E}\{\tilde{\ell}_{n;\lambda}(\boldsymbol{\Theta})|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Theta}^{(t)}\}$, which is given by

$$Q_{n;\lambda}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(t)}) = \sum_{i=1}^{n}\sum_{m=1}^{Q}\delta_{im}^{(t)}\log\pi_{im} + \sum_{i=1}^{n}\sum_{m=1}^{Q}\delta_{im}^{(t)}\log\phi(\boldsymbol{y}_i;\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m) - n\lambda\sum_{m=1}^{Q-1}\|\boldsymbol{\beta}_m\|_1,$$

subject to $\|\boldsymbol{\beta}\|_0 \leq \kappa$, where the expectation $\mathrm{E}\{\cdot\}$ is taken with respect to the conditional distribution of $\boldsymbol{\omega}$ given $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\boldsymbol{\Theta}^{(t)}$, and the weights

$$\delta_{im}^{(t)} = \frac{\pi_{im}^{(t)}\phi(\boldsymbol{y}_i;\boldsymbol{\mu}_m^{(t)},\boldsymbol{\Sigma}_m^{(t)})}{\sum_{d=1}^{Q}\pi_{id}^{(t)}\phi(\boldsymbol{y}_i;\boldsymbol{\mu}_d^{(t)},\boldsymbol{\Sigma}_d^{(t)})}$$

are the conditional expectations of the latent variables ω_{im}.

M-step. Determine $\boldsymbol{\Theta}^{(t+1)}$ by maximizing $Q_{n;\lambda}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(t)})$ with respect to $\boldsymbol{\Theta}$, subject to $\|\boldsymbol{\beta}\|_0 \leq \kappa$, where $Q_{n;\lambda}(\cdot)$ can be re-expressed as

$$Q_{n;\lambda}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(t)}) = \ell_1(\boldsymbol{\beta};\boldsymbol{\Theta}^{(t)}) + \ell_2(\boldsymbol{\psi};\boldsymbol{\Theta}^{(t)}) - n\lambda\sum_{m=1}^{Q-1}\|\boldsymbol{\beta}_m\|_1, \tag{2.5}$$

where

$$\ell_1(\boldsymbol{\beta};\boldsymbol{\Theta}^{(t)}) = \sum_{i=1}^{n}\left[\sum_{m=1}^{Q-1}\delta_{im}^{(t)}\boldsymbol{x}_i^\top\boldsymbol{\beta}_m - \log\left\{1+\sum_{m=1}^{Q-1}\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}_m)\right\}\right],$$

$$\ell_2(\boldsymbol{\psi};\boldsymbol{\Theta}^{(t)}) = \sum_{i=1}^{n}\sum_{m=1}^{Q}\delta_{im}^{(t)}\log\phi(\boldsymbol{y}_i;\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m),$$

where $\boldsymbol{\psi} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. It is almost impossible to directly implement the above presented M-step by maximizing $Q_{n;\lambda}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(t)})$ with respect to $\boldsymbol{\Theta}$ under the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

constrained condition $\|\boldsymbol{\beta}\|_0 \leq \kappa$, owing to the nonconvex optimization problem involved. As an alternative to this EM algorithm, we develop a hybrid method for evaluating the MLEs of $\boldsymbol{\Theta}$ by combining the EM iterative procedure and the alternating direction method of multipliers (ADMM) algorithm to optimize (2.3). The ADMM algorithm can be regarded as a variant of the augmented Lagrangian method. The latter is similar to the regularity method in that it transforms a constrained optimization problem into a series of unconstrained optimization problems, and adds a penalty term to the objective function considered. It has been applied to solve various nonconvex optimization problems with objective functions that are potentially nonsmooth and have some hard constraints. For example, see Boyd et al. (2011), Wang and Yuan (2012), and Wang, Yin, and Zeng (2019).

To use the ADMM to solve problem (2.4), we consider the following corresponding augmented Lagrangian optimization problem:

$$\begin{aligned} \max_{\boldsymbol{\varphi}} \mathcal{L}(\boldsymbol{\varphi}) &= \max_{\boldsymbol{\varphi}} \left\{ \ell_n(\boldsymbol{\Theta}) - n\lambda\|\boldsymbol{\beta}\|_1 - \boldsymbol{\xi}^\top(\boldsymbol{\beta}-\boldsymbol{\theta}) - (\rho/2)\|\boldsymbol{\beta}-\boldsymbol{\theta}\|_2^2 \right\} \\ &\quad \text{subject to } \|\boldsymbol{\theta}\|_0 \leq \kappa, \end{aligned} \tag{2.6}$$

where $\boldsymbol{\varphi} = \{\boldsymbol{\Theta}, \boldsymbol{\theta}, \boldsymbol{\xi}\} = \{\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta}, \boldsymbol{\xi}\}$, $\boldsymbol{\xi}$ is the Lagrange multiplier, and ρ is some positive scaling parameter. If we set $\boldsymbol{\xi} = \rho\boldsymbol{u}$, the ADMM for solving

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

(2.6) is implemented using the following gradient descent iteration procedure:

$$\begin{cases} (\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\psi}^{(t+1)}) = \arg\max_{\boldsymbol{\beta},\boldsymbol{\psi}} \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\theta}^{(t)}, \boldsymbol{u}^{(t)}), \\ \boldsymbol{\theta}^{(t+1)} = \boldsymbol{H}(\boldsymbol{\beta}^{(t+1)} + \boldsymbol{u}^{(t)}; \kappa), \\ \boldsymbol{u}^{(t+1)} = \boldsymbol{u}^{(t)} + \boldsymbol{\beta}^{(t+1)} - \boldsymbol{\theta}^{(t+1)}, \end{cases} \tag{2.7}$$

where $\boldsymbol{\psi} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ and $\boldsymbol{H}(\boldsymbol{\gamma}; \kappa) = \boldsymbol{\gamma} I(|\boldsymbol{\gamma}| > r_\kappa)$, in which r_κ is the κth largest component of $|\boldsymbol{\gamma}|$ and $I(\cdot)$ is an indictor function. In general, an iterative solution to the optimization problem (2.6) may depend on a "hot" starting value, such as a Lasso-type initial value. However, the Lasso-type initial value may be unstable in many settings, such as complicated models and distributed learning. The ADMM algorithm is less sensitive to the initial value and more stable than the gradient-based hard iteration method in the sense that the initial value conditions for the former are weaker than those for the latter (see the remarks after Theorem 1). Thus, the aforementioned ADMM algorithm for solving the optimization problem (2.6) is an appealing method in big data analysis. In what follows, we discuss the implementation of the above gradient descent iteration procedure.

To implement the first iteration in (2.7), we rewrite the optimization problem (2.6) associated with $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ as

$$(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\psi}^{(t+1)}) = \arg\max_{\boldsymbol{\beta},\boldsymbol{\psi}} \left\{ Q_{n;\lambda}(\boldsymbol{\beta}, \boldsymbol{\psi}; \boldsymbol{\beta}^{(t)}, \boldsymbol{\psi}^{(t)}) - \frac{\rho}{2} \|\boldsymbol{\beta} - \boldsymbol{\theta}^{(t)} + \boldsymbol{u}^{(t)}\|_2^2 \right\}, \tag{2.8}$$

which indicates that there are no closed-form solutions for $\boldsymbol{\beta}^{(t+1)}$ and $\boldsymbol{\psi}^{(t+1)}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

In this case, the gradient descent iteration method can be adopted to solve the optimization problem (2.8). That is, at the tth iteration, given the current value $\boldsymbol{\beta}^{(t)}$ of $\boldsymbol{\beta}$, we denote $\boldsymbol{\beta}^{(t,0)} = \boldsymbol{\beta}^{(t)}$, and define $\boldsymbol{\beta}^{(t,h)}$ as the hth updating value of $\boldsymbol{\beta}$ to maximize $\mathcal{F}(\boldsymbol{\beta}) = \ell_1(\boldsymbol{\beta}; \boldsymbol{\Theta}^{(t,h)}) - n\lambda\|\boldsymbol{\beta}\|_1 - (\rho/2)\|\boldsymbol{\beta} - \boldsymbol{\theta}^{(t)} + \boldsymbol{u}^{(t)}\|_2^2$, where $\boldsymbol{\Theta}^{(t,h)} = \{\boldsymbol{\beta}^{(t,h)}, \boldsymbol{\psi}^{(t,h)}\}$, in which $\boldsymbol{\psi}^{(t,h)}$ is the hth updating value of $\boldsymbol{\psi}$ at the tth iteration step. Considering the Taylor expansion of $\ell_1(\boldsymbol{\beta}; \boldsymbol{\Theta}^{(t,h)})$ at $\boldsymbol{\beta}^{(t,h)}$ leads to

$$\begin{aligned}
\boldsymbol{\beta}^{(t,h+1)} &= \arg\max_{\boldsymbol{\beta}} \{\ell_1(\boldsymbol{\beta}^{(t,h)}; \boldsymbol{\Theta}^{(t,h)}) + \dot{\ell}_1(\boldsymbol{\beta}^{(t,h)}; \boldsymbol{\Theta}^{(t,h)})^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t,h)}) \\
&\quad -(\nu_1/2)\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t,h)}\|_2^2 - n\lambda\|\boldsymbol{\beta}\|_1 - (\rho/2)\|\boldsymbol{\beta} - \boldsymbol{\theta}^{(t)} + \boldsymbol{u}^{(t)}\|_2^2\} \\
&= (1 + \rho\nu_1^{-1})^{-1} \text{Soft}\Big(\boldsymbol{\beta}^{(t,h)} + \nu_1^{-1}\{\dot{\ell}_1(\boldsymbol{\beta}^{(t,h)}; \boldsymbol{\Theta}^{(t,h)}) + \rho(\boldsymbol{\theta}^{(t)} - \boldsymbol{u}^{(t)})\}, \\
&\quad \nu_1^{-1} n\lambda\Big),
\end{aligned} \tag{2.9}$$

where ν_1 is a stepsize guaranteeing $\mathcal{F}(\boldsymbol{\beta}^{(t,h+1)}) \geq \mathcal{F}(\boldsymbol{\beta}^{(t,h)})$, and $\dot{\ell}_1(\boldsymbol{\beta}; \boldsymbol{\Theta})$ is the gradient of $\ell_1(\boldsymbol{\beta}; \boldsymbol{\Theta})$ with respect to $\boldsymbol{\beta}$. Soft$(\cdot)$ is the soft-thresholding operator, that is, Soft$(\boldsymbol{\beta}, \lambda) = \text{sign}(\boldsymbol{\beta})(|\boldsymbol{\beta}| - \lambda)_+$. When the above updating procedure converges with the rule $\|\boldsymbol{\beta}^{(t,h_0+1)} - \boldsymbol{\beta}^{(t,h_0)}\|_2 \leq c$ (e.g., $c = 0.001$), for some positive integer h_0, we take $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t,h_0+1)}$.

Similarly, at the tth iteration, given the current value $\boldsymbol{\psi}^{(t)} = \boldsymbol{\psi}^{(t,0)}$ of $\boldsymbol{\psi}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

the $(h+1)$th updating value $\boldsymbol{\psi}^{(t,h+1)}$ of $\boldsymbol{\psi}$ can be obtained as

$$\boldsymbol{\psi}^{(t,h+1)} = \arg\max_{\boldsymbol{\psi}} \ell_2(\boldsymbol{\psi}; \boldsymbol{\Theta}^{(t,h)}), \tag{2.10}$$

which leads to $\boldsymbol{\mu}_m^{(t,h+1)} = (n_m^{(t,h)})^{-1} \sum_{i=1}^n \delta_{im}^{(t,h)} \boldsymbol{y}_i$ and $\boldsymbol{\Sigma}_m^{(t,h+1)} = (n_m^{(t,h)})^{-1} \sum_{i=1}^n \delta_{im}^{(t,h)} (\boldsymbol{y}_i - \boldsymbol{\mu}_m^{(t,h+1)})(\boldsymbol{y}_i - \boldsymbol{\mu}_m^{(t,h+1)})^\top$, where $n_m^{(t,h)} = \sum_{i=1}^n \delta_{im}^{(t,h)}$. The preceding two-step iteration procedure for the optimization problem (2.8) yields the following proposition.

Proposition 2. *Let $\boldsymbol{\Theta}^{(t,1)}, \boldsymbol{\Theta}^{(t,2)} \ldots$ be an updating sequence of $\boldsymbol{\Theta}$ obtained from (2.9) and (2.10) for the above two-step algorithm, where $\boldsymbol{\Theta}^{(t,h)} = \{\boldsymbol{\beta}^{(t,h)}, \boldsymbol{\psi}^{(t,h)}\}$, for $h = 1, 2, \ldots$ Denote $\varsigma = \mathbb{E}_{\max}\{(\boldsymbol{I} - \mathbf{1}\mathbf{1}^\top/Q) \otimes \boldsymbol{X}^\top \boldsymbol{X}/4\}$, where $\mathbb{E}_{\max}(A)$ is the maximum eigenvalue of the matrix A, $\otimes$ denotes the Kronecker product, $\mathbf{1}$ is a $(Q-1) \times 1$ vector with components equal to one, and $\boldsymbol{I}$ is an identity matrix. If $\nu_1 \geq \varsigma$, we have*

$$\mathcal{L}(\boldsymbol{\Theta}^{(t,h+1)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{u}^{(t)}) \geq \mathcal{L}(\boldsymbol{\Theta}^{(t,h)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{u}^{(t)}), \tag{2.11}$$

which shows that the above algorithm ensures that $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\theta}^{(t)}, \boldsymbol{u}^{(t)})$ is an increasing function in $\boldsymbol{\Theta}$, and $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\theta}^{(t)}, \boldsymbol{u}^{(t)})$ can attain its maximum at $\boldsymbol{\Theta}^{(t+1)} = \{\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\psi}^{(t+1)}\}$ under some proper regularity conditions.

We now discuss how to select the stepsize ν_1, regularization parameter λ, and scaling parameter ρ given in (2.9). Many empirical studies have shown

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

that a smaller value of ν_1 often leads to a faster convergence of the above algorithm. Proposition 2 indicates that only if ν_1 is larger than ς is the objective function $\mathcal{L}(\boldsymbol{\Theta}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{u}^{(t)})$ guaranteed to increase at each iteration. In practice, one can first use a tentatively small ν_1, and then check condition (2.11). If (2.11) is not satisfied, we take ν_1 as twice its current value. Similarly, a small ρ helps to improve the penalized log-likelihood and boost the algorithm convergence. In practice, we begin with a small ρ, and gradually increase its value by $\rho = (1+\varepsilon)\rho$, for some $\varepsilon > 0$. As indicated in Figure 1, this adaptive choice of ρ often leads to a faster convergence than using fixed ρ updating. Our empirical experience shows that this adaptive strategy works with an insensitive choice of ε. In our numerical studies, we simply set $\varepsilon = 0.1$, which seems to work well for the cases we consider. For the regularization parameter λ, a proper value can accelerate the convergence of the algorithm and lead to a robust estimation of $\boldsymbol{\Theta}$. In general, we can choose λ using the generalized cross-validation method. For the sparsity parameter κ, an appropriate value may depend on the specific data structures and the nature of the model sparsity. Following Xu and Chen (2014), we use the screening bound $\kappa = 3^{-1}\log(n)n^{1/3}$. In our simulation studies, we show the performance of the HHS-EM algorithm is robust to a wide range of κ. This facilitates using the HHS-EM algorithm by avoiding an elaborative specification of κ.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

The proposed HHS-EM algorithm is summarized in Algorithm 1.

Algorithm 1 HHS-EM for high-dimensional LCA

Step 1. Set the initial value $\boldsymbol{\varphi}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\theta}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}, \boldsymbol{\xi}^{(0)})$, and let $t = 0$.

Step 2. Take $\boldsymbol{\beta}^{(t,0)} = \boldsymbol{\beta}^{(t)}$, set $h = 0$, and implement the following steps:

(2a) E-step: Compute $Q_n(\boldsymbol{\Theta}^{(t,h)}; \boldsymbol{\Theta}^{(t,h)})$;

(2b) M-step: Update $\boldsymbol{\beta}^{(t,h+1)}$ and $\boldsymbol{\psi}^{(t,h+1)}$ using (2.9) and (2.10), respectively;

(2c) Repeat steps (2a) and (2b) T_1 times.

Step 3. Set $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t,T_1)}$ and $\boldsymbol{\psi}^{(t+1)} = \boldsymbol{\psi}^{(t,T_1)}$, and update $\boldsymbol{\theta}^{(t+1)}$ and $\boldsymbol{u}^{(t+1)}$ using (2.7).

Step 4. If $\ell_{n;\lambda}(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\psi}^{(t+1)}) < \ell_{n;\lambda}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\psi}^{(t)})$, set $\rho^{(t+1)} = (1+\varepsilon)\rho^{(t)}$.

Step 5. Repeat steps 2–4 until the algorithm converges, and set $\hat{s}_m = \{j : \hat{\theta}_{mj} \neq 0\}$, for $m = 1, \cdots, Q$.

2.3 Post-screening classification

Using the HHS-EM algorithm, one can effectively screen out most irrelevant features for a high-dimensional LCA. With the retained features and their estimated LCA coefficients from Algorithm 1, we can then uncover the hidden class membership $\boldsymbol{\omega}_i$ in the training data using a Bayesian classification rule.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

Specifically, let $\hat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ be the estimate of $\boldsymbol{\Theta}$ from the HHS-EM procedure. Given the training data set $\{\boldsymbol{y}_i, \boldsymbol{x}_i\}_{i=1}^n$ and $\hat{\boldsymbol{\Theta}}$, it follows from Bayesian formula that the posterior probability of the event $\omega_{im} = 1$ can be estimated as

$$\widehat{\Pr}(\omega_{im} = 1|\boldsymbol{x}_i, \boldsymbol{y}_i) = \frac{\hat{\pi}_{im}\phi(\boldsymbol{y}_i; \hat{\boldsymbol{\mu}}_m, \hat{\boldsymbol{\Sigma}}_m)}{\sum_{h=1}^{Q} \hat{\pi}_{ih}\phi(\boldsymbol{y}_i; \hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\Sigma}}_h)},$$

where

$$\hat{\pi}_{im} = \frac{\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_m)}{\sum_{k=1}^{Q} \exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_k)}. \tag{2.12}$$

Accordingly, we assign the ith unit in the training data to class M, based on the following classification rule:

$$G_{1;\hat{\boldsymbol{\Theta}}}(\boldsymbol{x}_i, \boldsymbol{y}_i) = M \quad \text{with} \quad M = \underset{m \in \{1,\cdots,Q\}}{\operatorname{argmax}} \hat{H}_m(\boldsymbol{x}_i, \boldsymbol{y}_i), \tag{2.13}$$

where $\hat{H}_m = \hat{B}_m(\boldsymbol{x}_i, \boldsymbol{y}_i) - \hat{B}_Q(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and $\hat{B}_m(\boldsymbol{x}_i, \boldsymbol{y}_i) = -(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_m)^\top \hat{\boldsymbol{\Sigma}}_m^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_m) - \log|\hat{\boldsymbol{\Sigma}}_m| + 2\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_m$.

With the HHS-EM-based LCA, one can also conveniently classify a new case using the information on $\boldsymbol{x}$ only. Similarly to (2.12), we can estimate the posterior probability of the class membership as

$$\widehat{\Pr}(\omega_m = 1|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_m)}{\sum_{k=1}^{Q} \exp(\boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_k)}.$$

The corresponding classifier is given by

$$G_{2;\hat{\boldsymbol{\Theta}}}(\boldsymbol{x}) = M \quad \text{with} \quad M = \underset{m \in \{1,\ldots,Q\}}{\operatorname{argmax}} \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_m. \tag{2.14}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

3. Theoretical properties

In this section, we provide theoretical support for the proposed screening method. We first introduce some regularity conditions needed for our analysis. For simplicity, we denote the gradient and the Hessian matrix of $\ell_1(\boldsymbol{\beta}; \boldsymbol{\Theta}^{(t)})$ with respect to $\boldsymbol{\beta}$ as

$$\dot{\ell}_1(\boldsymbol{\beta}; \boldsymbol{\Theta}^{(t)}) = (\boldsymbol{X}^\top(\boldsymbol{\delta}_{.1}^{(t)} - \boldsymbol{\pi}_{.1}), \boldsymbol{X}^\top(\boldsymbol{\delta}_{.2}^{(t)} - \boldsymbol{\pi}_{.2}), \ldots, \boldsymbol{X}^\top(\boldsymbol{\delta}_{.Q-1}^{(t)} - \boldsymbol{\pi}_{.Q-1}))^\top \text{and}$$

$$\Xi_1(\boldsymbol{\beta}; \boldsymbol{\Theta}^{(t)}) = \frac{\partial^2 \ell_1(\boldsymbol{\beta}; \boldsymbol{\Theta}^{(t)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\sum_{i=1}^{n} (\boldsymbol{\Lambda}_i - \boldsymbol{\pi}_{i.}\boldsymbol{\pi}_{i.}^\top) \otimes \boldsymbol{x}_i \boldsymbol{x}_i^\top,$$

respectively, and the Hessian matrix of $\ell_n(\boldsymbol{\Theta})$ and the Fisher information matrix as

$$\Xi_n(\boldsymbol{\Theta}) = \frac{\partial^2 \ell_n(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}^\top} \quad \text{and} \quad I(\boldsymbol{\Theta}) = \mathrm{E}\left\{ \left[\frac{\partial \ell_n(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} \right] \left[\frac{\partial \ell_n(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} \right]^\top \right\},$$

respectively, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{Q-1})^\top$, $\boldsymbol{\delta}_{.m}^{(t)} = (\delta_{1m}^{(t)}, \ldots, \delta_{nm}^{(t)})^\top$, $\boldsymbol{\pi}_{.m} = (\pi_{1m}, \ldots, \pi_{nm})^\top$, $\boldsymbol{\Lambda}_i = \mathrm{diag}(\pi_{i1}, \ldots, \pi_{i,Q-1})$, and $\boldsymbol{\pi}_{i.} = (\pi_{i1}, \ldots, \pi_{i,Q-1})^\top$.

(C1) $\log(p) = O(n^\tau)$, for some $0 \leq \tau < 1$.

(C2) The number of classes $Q = O(n^\zeta)$, for some $0 < \zeta < 1$.

(C3) There exist some positive constants $\gamma_1, \gamma_2, \tau_1, \tau_2$, and α such that

$$\min_{j \in s^*} |\boldsymbol{\beta}_j^*| \geq \gamma_1 n^{-\tau_1}, \quad q < \kappa \leq \gamma_2 n^{\tau_2}, \quad \lambda = O(n^{-\alpha}),$$

$$\tau_1 + \max(\tau_2, \zeta) < \alpha < (1 - \tau)/2.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

(C4) There exists some positive constant c_0 such that $\max_{i,j} |x_{ij}| \leq c_0$.

(C5) There exists some positive constant $c_1 > 0$ such that, for a sufficiently large n, $\Xi_n(\boldsymbol{\Theta})$ is non-positive definite on a given domain $\boldsymbol{D}_{\boldsymbol{\Theta}}$ and

$$-[\ell_n(\boldsymbol{\Theta}^* + \boldsymbol{\Delta}) - \ell_n(\boldsymbol{\Theta}^*) - \dot{\ell}_n(\boldsymbol{\Theta}^*)^\top \boldsymbol{\Delta}] \geq c_1 n \|\boldsymbol{\Delta}_{s^*}\|_2^2, \tag{3.1}$$

for any $\boldsymbol{\Delta} \neq 0$ satisfying $\|\boldsymbol{\Delta}_{s_c^*}\|_1 \leq 3\|\boldsymbol{\Delta}_{s^*}\|_1$ and $\boldsymbol{\Theta}_{s^*} = \{\boldsymbol{\beta}_{s^*}, \boldsymbol{\psi}\}$.

(C6) There exist some positive constants c_μ and c_Σ such that $\|\boldsymbol{\mu}_m^*\|_\infty \leq c_\mu$ and $c_\Sigma^{-1} \leq \mathbb{E}_{\min}(\boldsymbol{\Sigma}_m^*) \leq \mathbb{E}_{\max}(\boldsymbol{\Sigma}_m^*) \leq c_\Sigma$, for all $m = 1, \ldots, Q$.

Condition (C1) indicates that the number of covariates can be exponentially high compared with the sample size n. Condition (C2) allows the number of classes for the response to diverge as n increases. Condition (C3) includes a few requirements for establishing the sure screening property of the HHS-EM algorithm. It implies that the minimum signal is bounded away from zero. Condition (C4) is widely used in the high-dimensional data analysis literature (e.g., Xu and Chen (2014)), and holds naturally with the rescaling operation. Condition (C5) states that when n is sufficiently large, the observed information matrix $\Xi_n(\boldsymbol{\Theta})$ is non-positive definite on a given domain $\boldsymbol{D}_{\boldsymbol{\Theta}}$, which is weaker than the condition given in Jordan and Xu (1995). Furthermore, the restriction puts a limitation on the set of $\boldsymbol{\Delta}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

for which (3.1) is required, so that it is weaker than imposing nonzero eigenvalues. The condition matches the restricted eigenvalue condition given in Bickel et al. (2009). Using the assumption of Wang et al. (2015) for the high–dimensional mixture model that $\mathbb{E}_{\min}(I(\boldsymbol{\Theta}^*)) > 0$ under condition (C2) and $\|\boldsymbol{\Delta}_{s_c^*}\|_1 \leq 3\|\boldsymbol{\Delta}_{s^*}\|_1$, we can obtain that the left-hand side of (3.1) is bounded below by $n\|\boldsymbol{\Delta}_{s^*}\|_2^2\{\mathbb{E}_{\min}(I(\boldsymbol{\Theta}^*)) + o_p(1)\}$ using a similar process to that of Bickel et al. (2009). Thus, assumption (C5) of restricted eigenvalues for the latent class model is reasonable. Condition (C6) assumes that the parameters in the mixture model are in a bounded compact space. The same condition appears in Jiang et al. (2015) and Huang, Peng, and Zhang (2017).

Theorem 1. (Sure screening). *Under conditions (C1)–(C6), if the initial values satisfy* $\rho^{(0)}\|\boldsymbol{\beta}^* - \boldsymbol{\theta}^{(0)} + \boldsymbol{u}^{(0)}\|_\infty \leq n\lambda/4$, *with* $\|\boldsymbol{u}^{(0)}\|_1 = o(n^{-\tau_1})$, *there exist optimal solutions* $\hat{\boldsymbol{\Theta}} = \{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\}$ *and* $\hat{\boldsymbol{\theta}}$ *in (2.7) satisfying*

$$\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_1 = o_p(n^{-\tau_1}) \quad \textit{and} \quad \|\hat{\boldsymbol{\theta}} - \boldsymbol{\beta}^*\|_1 = o_p(n^{-\tau_1}).$$

Moreover, let $\hat{s}_m = \{j : \hat{\theta}_{mj} \neq 0\}$ *and* $\hat{s} = \{\hat{s}_1, \cdots, \hat{s}_{Q-1}\}$. *Then, we have*

$$\Pr(s^* \subset \hat{s}) \to 1 \quad \textit{as} \quad n \to \infty,$$

where $s^* \subset \hat{s}$ *denotes* $s_1^* \subset \hat{s}_1, \ldots, s_{Q-1}^* \subset \hat{s}_{Q-1}$.

Theorem 1 shows that, with a good initial value, the proposed HHS-EM procedure leads to a consistent estimate for the LCA parameters under the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

high-dimensional setup. The procedure enjoys the sure screening property; that is, with probability tending to one, all relevant features are retained after screening. Note that, because $\lambda = O(n^{-\alpha})$, for some $\alpha \in (0, 1)$, $n\lambda/4$ is in the same order of $n^{1-\alpha}$. By choosing a sufficiently small $\rho^{(0)}$, the requirement of the initial values in Theorem 1 can be easily satisfied, even when $\boldsymbol{\theta}^{(0)}$ is not very close to $\boldsymbol{\beta}^*$. Thus, we can simply set $\boldsymbol{\theta}^{(0)} = \mathbf{0}$ when using Algorithm 1 in practice. Note that, in L_0-based methods, a root-n consistent initial estimation is usually needed to ensure the sure screening property (Xu and Chen (2014)). However, such an accurate initial estimation might not be easy to obtain in a high-dimensional LCA. By comparison, using the hybrid L_0-L_1 penalty in the proposed method helps to relax this requirement and improves the practical screening accuracy.

We now investigate the classification accuracy of the HHS-EM-based LCA, as discussed in Section 2.3. We assess the accuracy of the classifier $G_{1;\widehat{\boldsymbol{\Theta}}}$ (as defined in (2.13)) using the training misclassification rate $\hat{R}_1$, which measures the goodness of fit of a model-based classifier on the training data. A low $\hat{R}_1$ indicates that the classifier is able to correctly classify the cases in the training data set used to screen the features and estimate the parameters of

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

the classifier. Mathematically, $\hat{R}_1$ is given by

$$\begin{aligned}\hat{R}_1 &= \Pr\{G_{1;\widehat{\boldsymbol{\Theta}}}(\boldsymbol{x},\boldsymbol{y}) \neq \mathbb{L}(\boldsymbol{y})\} \\ &= \sum_{m=1}^{Q} \sum_{G_{1;\widehat{\boldsymbol{\Theta}}}(\boldsymbol{x},\boldsymbol{y})\neq m} \mathrm{E}\{\pi_m(\boldsymbol{x},\boldsymbol{\beta}^*) R(G_{1;\widehat{\boldsymbol{\Theta}}}(\boldsymbol{x},\boldsymbol{y})|\mathbb{L}(\boldsymbol{y})=m)\}, \quad (3.2)\end{aligned}$$

where $R(r|m)$ denotes the probability that a unit in class m is misclassified to class r by rule (2.13), and $\mathbb{L}(\boldsymbol{y})$ denotes the true class membership of a unit with response $\boldsymbol{y}$. When $Q = 2$, $\hat{R}_1$ can be explicitly expressed as

$$\hat{R}_1 = \mathrm{E}\left\{\pi_1(\boldsymbol{x},\boldsymbol{\beta}^*)\int_{\hat{H}_1(\boldsymbol{x},\boldsymbol{y})<0} f_1(\boldsymbol{y})d\boldsymbol{y} + \pi_2(\boldsymbol{x},\boldsymbol{\beta}^*)\int_{\hat{H}_1(\boldsymbol{x},\boldsymbol{y})\geq 0} f_2(\boldsymbol{y})d\boldsymbol{y}\right\},$$

where $f_m(\boldsymbol{y})$ is the probability density of the response for class m, with $m = 1, 2$.

To assess $\hat{R}_1$, we compare it with the misclassification rate of the conceptually best classifier $G_{1;\boldsymbol{\Theta}^*}$, which is constructed based on the true model s^*. This optimal rate is given by

$$\begin{aligned}R_{1,\mathrm{opt}} &= \Pr\{G_{1;\boldsymbol{\Theta}^*}(\boldsymbol{x},\boldsymbol{y}) \neq \mathbb{L}(\boldsymbol{y})\} \\ &= \sum_{m=1}^{Q} \sum_{G_{1;\boldsymbol{\Theta}^*}(\boldsymbol{x},\boldsymbol{y})\neq m} \mathrm{E}\{\pi_m(\boldsymbol{x},\boldsymbol{\beta}^*) R(G_{1;\boldsymbol{\Theta}^*}(\boldsymbol{x},\boldsymbol{y})|\mathbb{L}(\boldsymbol{y})=m)\}.\end{aligned}$$

We assess the classifier $G_{2;\widehat{\boldsymbol{\Theta}}}$ (as defined in (2.14)) using the testing misclassification rate $\hat{R}_2$. This rate measures the accuracy of a classifier in predicting the class label of a new case based on information on $\boldsymbol{x}$ only. Mathematically, $\hat{R}_2$ is expressed as

$$\hat{R}_2 = \Pr\{G_{2;\widehat{\boldsymbol{\Theta}}}(\boldsymbol{x}) \neq \mathbb{L}(\boldsymbol{y})\}. \quad (3.3)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

Similarly, we compare $\hat{R}_2$ with the optimal rate $R_{2,\text{opt}}$, which is defined by

$$R_{2,\text{opt}} = \Pr\{G_{2;\boldsymbol{\Theta}^*}(\boldsymbol{x}) \neq \mathbb{L}(\boldsymbol{y})\}.$$

In general, $\hat{R}_1(R_{1,\text{opt}})$ tends to be relatively smaller than $\hat{R}_2$ ($R_{2,\text{opt}}$), owing to its in-sample classification. We derive the misclassification error bounds for both $\hat{R}_1$ and $\hat{R}_2$ in the following theorem.

Theorem 2. (Misclassification error). *Let $f_{H_m^*}(x)$ be the probability density of $H_m^*(\boldsymbol{x}, \boldsymbol{y})$, for $m = 1, \ldots, Q$. Suppose $\sup_{|x - H_l^*(\boldsymbol{x},\boldsymbol{y})|<c} f_{H_m^*}(x) < A$, for some positive constants c and A. Then, under the conditions of Theorem 1, we have*

$$\hat{R}_1 - R_{1,\text{opt}} = O_p[(Q-2)n^{\tau_3-\tau_1} + n^{2\tau_3-2\tau_1}] \quad \text{and} \quad \hat{R}_2 - R_{2,\text{opt}} = O_p(n^{-\tau_1}),$$

for some $\tau_3 \in (0, \tau_1)$.

With most irrelevant features screened out by the HHS-EM algorithm, we obtain a refined LCA with a manageable set of important features. From Theorem 2, the HHS-EM-based LCA leads to viable classifiers $G_{1;\widehat{\boldsymbol{\Theta}}}$ and $G_{2;\widehat{\boldsymbol{\Theta}}}$, which mimic the oracle classifiers $G_{1;\boldsymbol{\Theta}^*}$ and $G_{2;\boldsymbol{\Theta}^*}$, respectively. Our classifiers are asymptotically optimal in the sense that, as $n \to \infty$, their misclassification rates $\hat{R}_1$ and $\hat{R}_2$, respectively, converge to the optimal rates $R_{1,\text{opt}}$ and $R_{2,\text{opt}}$, respectively, in probability. In other words, when the sample size is large, the proposed method performs an effective classification, as if the true model were known in advance.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

In particular, when $Q = 2$, we have $\hat{R}_1 - R_{1,\text{opt}} = O_p(n^{2(\tau_3-\tau_1)})$. When Condition (C1) is satisfied with $\tau \to 0$, Condition (C3) implies that the value of τ_1 can get arbitrarily close to 0.5 with sufficiently small τ_2 and ζ. Consequently, the bound of $\hat{R}_1 - R_{1,\text{opt}}$ is nearly $O_p(n^{-1})$, which echoes the minmax error bound for the clustering of high-dimensional Gaussian mixtures with two components (Cai et al. (2019)).

4. Numerical studies

In this section, we present several simulation studies to investigate the finite–sample performance of the proposed HHS-EM procedure in terms of its screening accuracy and misclassification rate. The simulations are conducted using the software **R** on a Microsoft Windows computer with an eight-core 2.21 GHz CPU and 16 GB RAM.

In particular, we compare the proposed method with several popular screening approaches: the marginal maximum likelihood estimation(MMLE) (Fan and Song (2010)), mean variance sure independence screening (MV-SIS)(Cui, Li, and Zhong (2015)), and forward regression (FR) (Wang (2009)). Because these methods are designed mainly for regression models, we modify them for an LCA under an EM-framework, and refer to these modified methods as MMLE-EM, MV-EM, and FR-EM, respectively. In the modified

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

screening methods, the screening indicator is sequentially estimated using a marginal LCA model involving one feature at a time. Owing to the nature of these methods, we treat the retained features as relevant for all classes, and conduct the classification using these features based on a refitted LCA. We also include the L_1-penalized selection method discussed in Houseman et al. (2006) and Wu (2013), which we refer to as the Lasso-EM.

In our numerical studies, the HHS-EM is implemented based on Algorithm 1, with $\lambda = 0.1\sqrt{\log p/n}$ and $\varepsilon = 0.1$. We set $\boldsymbol{\beta}^{(0)} = \boldsymbol{\theta}^{(0)} = \mathbf{0}$, and set $\boldsymbol{\mu}_m^{(0)}$ and $\boldsymbol{\Sigma}_m^{(0)}$ using the naive K-means method. We terminate the iterations when $\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\|_2 < 10^{-3}$, and evaluate the screening accuracy of each method using the following two indices:

$$\text{RC} = \frac{1}{\mathbb{V}} \sum_{\nu=1}^{\mathbb{V}} I(s^* \subset \hat{s}_\nu) \quad \text{and} \quad \text{PSR} = \frac{1}{\mathbb{V}\|s^*\|_0} \sum_{\nu=1}^{\mathbb{V}} \|s^* \cap \hat{s}_\nu\|_0,$$

where $\hat{s}_\nu$ denotes the set of retained features for the νth independent experiment, and $\mathbb{V}$ is the total number of experiments. RC measures the proportion of times that all relevant features are retained after screening, and PSR is the averaged proportion of the retained relevant features. Because relevant features may vary across classes in an LCA, we use RC and PSR to indicate the overall values, and use RC_m and PSR_m to denote the class-specific values for class m. A good screening method in an LCA is expected to have a high RC_m and PSR_m for all classes.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

In addition, for each of the aforementioned methods, we assess the associated post-screening classification using both the training and the testing misclassification rates, as discussed in Section 3. Specifically, let $\mathcal{D}_{\text{train}} = \{\boldsymbol{y}_i, \boldsymbol{x}_i\}_{i=1}^n$ denote a training data set of size n, based on which $G_{1;\widehat{\boldsymbol{\Theta}}}$ and $G_{2;\widehat{\boldsymbol{\Theta}}}$ are constructed. Let $\mathcal{D}_{\text{test}} = \{\boldsymbol{y}_i, \boldsymbol{x}_i\}_{i=1}^n$ be a testing data set of the same sample size corresponding to $\mathcal{D}_{\text{train}}$.

We estimate the training misclassification rate as

$$\text{TrMR} = \frac{1}{n\mathbb{V}} \sum_{\nu=1}^{\mathbb{V}} \sum_{i \in \mathcal{D}_{\text{train}}} I\{\mathbb{L}(\boldsymbol{y}_i) \neq G_{1;\hat{\boldsymbol{\Theta}}}(\boldsymbol{x}_i, \boldsymbol{y}_i)\},$$

and estimate the testing misclassification rate as

$$\text{TeMR} = \frac{1}{n\mathbb{V}} \sum_{\nu=1}^{\mathbb{V}} \sum_{i \in \mathcal{D}_{\text{test}}} I\{\mathbb{L}(\boldsymbol{y}_i) \neq G_{2;\hat{\boldsymbol{\Theta}}}(\boldsymbol{x}_i)\}.$$

As a benchmark, we also report the TrMR and TeMR of $G_{1;\boldsymbol{\Theta}^*}$ and $G_{2;\boldsymbol{\Theta}^*}$, which are constructed based on the true model s^*.

We generate $\mathcal{D}_{\text{train}}$ based on model (2.1) under two scenarios. In scenario 1, we set $Q = 3$, and generate the response based on the following two cases:

(M1) $\boldsymbol{y}$ is a 3×1 vector of response variables. The true values of $\boldsymbol{\mu}_m$ $(m = 1, 2, 3)$ are taken as $\boldsymbol{\mu}_1 = (-1.0, 0.0, 1.0)^\top$, $\boldsymbol{\mu}_2 = (0.0, 1.5, -1.0)^\top$, and $\boldsymbol{\mu}_3 = (1.0, -1.5, 0.0)^\top$, and the true values of $\boldsymbol{\Sigma}_m$ $(m = 1, 2, 3)$ are set as $\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{jj} = 1$, $\boldsymbol{\Sigma}_{ij} = 0.5$, for $i \neq j$, $i, j = 1, 2, 3$. In this case, the groups have the same correlation structure, but different mean structures.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

(M2) $\boldsymbol{y}$ is a univariate response variable. The true values of μ_m ($m = 1, 2, 3$) are taken as $\mu_1 = -5.0$, $\mu_2 = 0.0$, and $\mu_3 = 5.0$, and the true values of Σ_1, Σ_2, and Σ_3 are set as $\Sigma_1 = \Sigma_3 = 4.0$ and $\Sigma_2 = 1.0$. This indicates that the three groups considered have different means.

With (M1) and (M2), we further consider the following four model setups:

(S1) $s^* = \{s_1^*, s_2^*\}$, with $\|s_1^*\|_0 = \|s_2^*\|_0 = 4$, and s_m^* is a simple random sample of size four generated from the index set $\{1, \ldots, p\}$, for $m = 1, 2$. The components of $\boldsymbol{\beta}_{s_m^*}$ are generated independently from $U\{6\log(n)/\sqrt{n} + |Z|/4\}$, and the components of $\boldsymbol{\beta}_{s_m^{c*}}$ are taken as zero, for $m = 1, 2$, where U is a Bernoulli random variable with $\Pr(U = 1) = \Pr(U = -1) = 0.5$, and Z is sampled from the standard normal distribution $\mathcal{N}(0, 1)$. The responses are generated from case (M1), and the features x_{ik} are generated independently from the standard normal distribution $\mathcal{N}(0, 1)$, for $i = 1, \ldots, n$ and $k = 1, \ldots, p$. Obviously, these features are independent of each other. In this case, we set $(n, p) = (320, 1500)$.

(S2) $s_1^* = \{1, 3, 5\}$, $s_2^* = \{7, 9, 11\}$, $\boldsymbol{\beta}_{s_1^*} = (2.0, 2.0, -1.5)$, and $\boldsymbol{\beta}_{s_2^*} = (-1.5, 2.0, 2.0)$, and the components of $\boldsymbol{\beta}_{s_1^{c*}}$ and $\boldsymbol{\beta}_{s_2^{c*}}$ are set as zero. The responses are generated from case (M1), and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ are independently sampled from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Upsilon})$, for $i = 1, \ldots, n$, where $\boldsymbol{\Upsilon} = (\gamma_{jk})$, with $\gamma_{jj} = 1.0$, $\gamma_{j,j-1} = 2/3$, $\gamma_{j,j-2} = 1/3$ for

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

$j \geq 3$, and $\gamma_{jk} = 0.0$ for $|j-k| \geq 3$. In this case, we take $(n, p) = (300, 1500)$.

(S3) $s_1^* = \{1, 2, 3, 4\}$, $s_2^* = \{1, 2, 3, 5\}$, and $\boldsymbol{\beta}_{s_m^*} = (2.0, 2.0, 2.0, 2.0)$, for $m = 1, 2$. The responses are generated from case (M2). The features $\boldsymbol{x}_i$ are generated independently from case (S2), with $\boldsymbol{\Upsilon}$ specified by an AR(1) structure, that is, $\gamma_{jk} = 0.5^{|j-k|}$. In this case, we set $(n, p) = (320, 2000)$.

(S4) s_1^* and s_2^* take the same values as those in case (S3), and $\boldsymbol{\beta}_{s_m^*} = (3.0, 3.0, 3.0, -3.0)$, for $m = 1, 2$. The responses are generated from case (M2). The features $\boldsymbol{x}_i$ are generated independently from case (S2), with $\boldsymbol{\Upsilon}$ having a compound symmetry (CS) structure, that is, $\gamma_{jj} = 1.0$ and $\gamma_{jk} = 0.5$, for $j \neq k$. In this case, we set $(n, p) = (350, 2000)$.

Note that in (S1) and (S2), the two sets of relevant features have no overlap, whereas in (S3) and (S4), features 1, 2, and 3 affect both classes.

In scenario 2, we further consider two model setups, (S5) and (S6), in which $Q = 6$ and $(n, p) = (400, 1000)$. In these two setups, we generate a univariate response from model (2.1) with $\mu_1 = -4.0$, $\mu_2 = -2.0$, $\mu_3 = 0.0$, $\mu_4 = 2.0$, $\mu_5 = 4.0$, $\mu_6 = 6.0$, $\Sigma_1 = \Sigma_6 = 4.0$, and $\Sigma_2 = \cdots = \Sigma_5 = 1.0$. We set $\boldsymbol{\beta}_{s_m^*} = (2, -2)$, where s_m^* contains two randomly chosen features from $\{1, ...p\}$, for $m = 1, \cdots, 5$. In (S5), the features $\boldsymbol{x}_i$ are generated independently based on (S2), with $\gamma_{jj} = 1.0$, $\gamma_{jk} = 0.15$ for $j, k \in s_m^*$, and $\gamma_{jk} = 0.3$ for $j, k \in s_m^{c*}$. In (S6), the features $\boldsymbol{x}_i$ are generated independently based on (S4).

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

We use the proposed HHS-EM and its competitors to retain $\kappa = 3^{-1}\log(n)n^{1/3}$ features on the data sets generated from setups (S1)–(S6). The simulation results are summarized in Table 1, based on $\mathbb{V} = 100$ repetitions, where the running time in seconds (Time(s)) for a single repetition is also reported. All methods perform well in (S1), which is the most straightforward setup for feature screening. As the correlation structure among $\boldsymbol{x}$ becomes more complex, the screening accuracy of the marginal-effect-based methods (i.e., MMV–EM and MMLE-EM) deteriorates drastically. Although Lasso-EM and FR-EM maintain a decent RC in (S2), they seem not to work well for retaining all s_m^* in (S3), where some features are important for both classes. In comparison, the proposed HHS-EM shows promising accuracy by achieving the highest RC and the lowest TrMR/TeMR for all cases. In particular, it leads to a high RC of 0.92 for a very challenging case (S4), where the irrelevant features are strongly correlated with the relevant features. By setting $Q = 6$ in (S5) and (S6), we further increase the difficulty of the feature screening. In those two cases, the HHS-EM still yields encouraging results by showing the "best-in-group" accuracy for both screening and classification. By the nature of Algorithm 1, the high accuracy of the HHS-EM comes with a computational cost, but this is moderate in most cases. Considering the improved accuracy achieved by the HHS-EM, this small computational investment seems

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

worthwhile.

We repeat the above simulation study in (S3) with the screening size κ varying from $\|s^*\|_0$ to $2\log(n)n^{1/3}$. We report the RC of different screening methods over κ in Figure 2. We observe that the proposed method is able to maintain a high RC over a wide range of κ, making it a viable and robust approach in practice.

5. An real–data example

To further demonstrate the proposed HHS-EM method, we apply it to a crime data set that contains the crime rate and $p = 100$ other social and economic indices measured on $n = 1994$ US communities. This data set incorporates data from the 1990 US census, 1990 US LEMAS survey, and 1995 FBI UCR. For more information about this data set, please refer to http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime.

Our goal is to identify important social indices (e.g., income, school engagement, etc.) that could help to partition the communities into subgroups based on their crime rates. To this end, an LCA is performed using model (2.1), where the crime rate of a community is used as the response $\boldsymbol{y}$, and the 100 social indices are treated as the covariates (features) $\boldsymbol{x}$. In the model, we postulate $Q = 3$ latent classes with low, high, and general criminal risk,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

respectively, where the unknown class membership $\boldsymbol{w}_i$ of community i is to be estimated for $i = 1, \cdots, 1994$.

Because we have $p = 100$ features in this case, feature screening is beneficial for the analysis. To obtain a reliable result, we generate a training set by randomly selecting 50% of the observations from the full data, and use the HHS-EM and its competitors to retain the most important $\kappa = 10$ features using the training set. We repeat the above procedure 100 times. A feature is considered important if it is retained after screening by at least one method and at least once.

In Table 2, we report the identities of the important features, along with their (averaged) estimated LCA coefficients and standard errors. We find that three features are suggested by all the methods:

x_4 : Percentage of Caucasian population;

x_{45} : Percentage of kids in family housing with two parents;

x_{51} : Percentage of kids born to never married.

The HHS-EM shows that x_4 and x_{45} have a positive effect of 0.8 for class **1**, and x_{51} has a positive effect of 1.28 in class **2**. Accordingly, communities with higher percentages of a Caucasian population and kids in families with two parents are more likely to belong to class **1**, whereas communities with a

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

higher percentage of kids born to couples who were never married are more likely to belong to class **2**.

To determine the identities of the two classes, we classified the $n = 1994$ communities based on an LCA using x_4, x_{45}, and x_{51} only. We find that the communities classified as class **1** tend to have a low crime rate, and the communities classified as class **2** tend to have a high crime rate. Intuitively, this seems to confirm that class **1** and class **2** are associated with a low and high crime risk, respectively; class **3** is treated as a class with a general risk. Our findings match the results in the relevant literature (e.g., Li et al. (2017)).

To further test the HHS-EM, we combine the original 100 features with 1200 synthetic random features. We then perform the feature screening on the combined data using the same procedure as before with 50% or 70% of the training data. In Table 3, we show the proportion of times (i.e., RC) that x_4, x_{45}, and x_{51} are still selected as important, based on 100 repetitions. The results of the other screening methods are also reported for comparison purpose. The proposed HHS-EM shows its superior robustness and stability over its competitors by achieving the highest RC for all three important features.

In addition, we select a set of communities from each class based on the original LCA, and then repeat the membership classification using the new screening results based on the combined data. In Table 4, we report the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

proportion of times that a selected community is classified into the same class as in the original analysis, based on 100 repetitions. This tests whether a screening method leads to robust classification against changes in the data. Table 4 again demonstrates the promising performance of the HHS-EM.

6. Conclusion

We have proposed a new feature screening method, the HHS-EM, for high-dimensional LCAs, in which accurate selection is often difficult. The proposed method is inspired from solving a hybrid L_0-L_1 penalized problem, allowing users to precisely control the number of features to be retained, with significantly enhanced stability. The promising performance of the method is supported by both theory and extensive numerical examples.

Our current work focuses on cases with normally distributed responses. We briefly discuss extending the HHS-EM to LCAs with categorical responses in the Supplementary Material. It would also be interesting to extend the existing work further to cases with mixed-type responses. Finally, it would be promising to explore the possibility of using a hybrid L_0-L_1 penalization in nonparametric (model-free) clustering.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

Supplementary Material

The proofs, tables, and figures in this paper are provided in the online Supplementary Material.

Acknowledgments

This work was supported in part by NSFC grants 11731011 and 11690014, and by NSERC grant RGPIN-2016-05024. The content is solely the responsibility of the authors, and does not necessarily represent the official views of the aforementioned funding agencies.

References

Bertsimas, D., Copenhaver, M. S. and Mazumder, R. (2017). The trimmed lasso: sparsity and robustness. arXiv preprint arXiv:1708.04527.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37**, 1705–1732.

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

Cai, T. T., Ma, J., Zhang, L., et al. (2019). CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *The Annals of Statistics* **47**, 1234–1267.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

Cui, H., Li, R. and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association* **110**, 630–641.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B* **70**, 849–911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38**, 3567–3604.

Fop, M., and Murphy, T. B. (2018). Variable selection methods for model based clustering. *Statistics Surveys* **12**, 18–65.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York.

Ghosh, J., Herring, A. H. and Siega-Riz, A. M. (2011). Bayesian variable selection for latent class models. *Biometrics* **67**, 917–925.

Huang, T., Peng, H. and Zhang, K. (2017). Model selection for Gaussian mixture models. *Statistica Sinica* **27**, 147–169.

Houseman, E. A., Coull, B. A., and Betensky, R. A. (2006). Feature-specific penalized latent class analysis for genomic data. *Biometrics* **62**, 1062–1070.

Jiang, B., Wang, X. and Leng, C. (2015). Quda: A direct approach for sparse quadratic discriminant analysis. *Journal of Machine Learning Research* **19**, 1098–1134.

Jordan, M. I. and Xu, L. (1995). Convergence results for the EM approach to mixtures of experts architectures. *Neural networks* **8**, 1409–1431.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

Khalili, A. (2010). New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics* **38**, 519–539.

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

Li, X., Li, R., Xia, Z. and Xu, C. (2020). Distributed Feature Screening via Componentwise Debiasing. *Journal of Machine Learning Research* **21**, 1–32.

Li, X., Cheng, G., Wang, L., Lai, P., Song, F. (2017). Ultrahigh dimensional feature screening via projection. *Computational Statistics & Data Analysis* **114**, 88–104.

Liu, J., Zhong, W. and Li, R. (2015). A selective overview of feature screening for ultra-high dimensional data. *Science in China: Mathematics* **58**, 2033–2054.

Qu L, Hao M, Sun L (2021). Sparse composite quantile regression with ultra-high dimensional heterogeneous data. *Statistica Sinica* DOI:10.5705ss.202020.0115

Tang, W., Xie, J., Lin, Y., Tang, N. (2021). Quantile correlation-based variable selection. *Journal of Business & Economic Statistics* DOI:10.108007350015.2021.1899932

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

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104**, 1512–1524.

Wang, X. and Yuan, X. (2012). The linearized alternating direction method of multipliers for

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

dantzig selector. *SIAM Journal on Scientific Computing* **34**, 2792–2811.

Wang, Y., Yin, W. and Zeng, J. (2019). Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing* **78**, 29–63.

Wang, Z., Gu, Q., Ning, Y. and Liu, H. (2015). High Dimensional EM Algorithm: Statistical Optimization and Asymptotic Normality. *Advances in neural information processing systems* **28**, 2521–2529.

Weller, B. E., Bowen, N. K., and Faubert, S. J. (2020). Latent class analysis: a guide to best practice. *Journal of Black Psychology* **46**, 287–311.

Wu, B. (2013). Sparse cluster analysis of large-scale discrete variables with application to single nucleotide polymorphism data. *Journal of applied statistics* **40**, 358–367.

Xie, J., Lin, Y., Yan, X., Tang, N. (2020). Category-adaptive variable screening for ultrahigh dimensional heterogenous categorical data. *Journal of the American Statistical Association*, **115** ,747–760.

Xu, C. and Chen, J. (2014). The sparse MLE for ultrahigh dimensional feature screening. *Journal of the American Statistical Association* **109**, 1257–1269.

Yang, G., Yu, Y., Li, R., Buu, A. (2016). Feature screening in ultrahigh dimensional Cox's model. *Statistica Sinica* **26**, 881–901

Zhang, Q. and Ip, E. (2014). Variable assessment in latent class models. *Computational Statistics and Data Analysis* **77**, 146–156.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0061

Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kunming 650500, P. R. of China

E-mail: 928762571@qq.com

School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, P. R. of China

E-mail: xli396@uottawa.ca

Department of Mathematics and Statistics, University of Ottawa, ON, K1N 6N5, Canada

E-mail: cx3@uottawa.ca

Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kunming 650500, P. R. of China

E-mail: nstang@ynu.edu.cn Tel: +86-871-65032416 Fax: +86-871-65033700