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
Notice: Accepted version subject to English editing.	

HYBRID HARD-SOFT SCREENING FOR HIGH DIMENSIONAL LATENT CLASS ANALYSIS

Wei Dong¹, Xingxiang Li², Chen Xu³ and Niansheng Tang¹

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

Abstract: Latent class analysis (LCA) is a powerful tool in detecting unobservable subgroups within a population. When a large number of covariates (features) are considered, LCA faces great challenges in both classification accuracy and computational efficiency. In this paper, we propose a novel feature screening procedure, which eliminates most irrelevant features before LCA is conducted. The proposed method is built upon an EM-based hybrid hard-soft thresholding updating (HHS-EM) on latent class parameters, where the joint effects between features can be naturally accounted. We show that 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 simulation studies and a real data example.

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

1. Introduction

High-dimensional datasets containing unobservable subgroups are commonly encountered in various fields such as biomedical, social, behavioral, and economical sciences. It is often of interest to discover the latent subgroups (classes) within a population and classify new cases input based on the information of their data input. In the literature, this scientific task is typically conducted by latent class analysis (LCA), which attempts to reveal the 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 by criminal actives that are possibly linked with hundreds of community indices such as household income, school engagement, insurance coverage, food insufficiency, and so forth. When the number of features p is large, LCA faces simultaneous challenges from both methodological and computational aspects due 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. With this sparsi-

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

To ease the implementation difficulties for high-dimensional LCA, a natural strategy is to remove most irrelevant features before LCA is conducted. Such a strategy is referred to as feature screening, which has attracted a great deal of attention in the past decade. With dimensionality reduced from high to low, analytical difficulties are reduced drastically. In the literature, plenty of works have been done in this area. For example, Fan and Lv (2008) proposed sure independent screening (SIS) and its iterated version (ISIS) procedure for linear models; Wang (2009) justified the use of forward regression for sequentially screening features; Fan and Song (2010) extended SIS (ISIS)

to generalized linear models based on maximum marginal likelihood estimation (MMLE); Xu and Chen (2014) proposed 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 analysis; Li et al. (2020) proposed a distributed screening framework for the divide-and-conquer setup; Xie et al. (2020) and Tang et al. (2021) developed category-adaptive and quantile correlation-based variable selection. Readers may refer to Liu et al. (2015) for an overview on feature screening.

Despite the rich literature on feature screening, the existing methods are mainly developed based on a direct measurement of correlation between the response and the data features. Thus, they may not be directly applicable to LCA, where the response is implicitly linked with the features via the unobserved class labels. In this paper, we propose a new feature screening approach for high-dimensional LCA. The proposed method is built upon a hybrid hard-soft EM thresholding procedure (HHS-EM), which 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 a LCA feature screener. The HHS-EM procedure efficiently leads to an improving mixture likelihood, where the joint effects between features are naturally accounted. With the hybrid hard-soft thresholding, we are able to reach a good

balance between sparsity and convexity in the numerical updating, such that the procedure is less sensitive to 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 conducts effective classification for both training and testing data, as if the true model were known in advance. We use extensive numerical examples to demonstrate the promising performance of HHS-EM.

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). While both methods are built upon a hybrid penalization technique, they are designed under different model setups, serve for different research purposes, and are studied from different angles. Specifically, the trimmed Lasso was proposed as a robust estimation method for the least square problem. The work in (Bertsimas et al., 2017) for the trimmed Lasso focused on finding its robustness interpretation and showing its connection with other penalized methods; their work was conducted mainly from an optimization perspective with no statistical property studied. In comparison, we propose HHS-EM as a procedure for feature screening in high-dimensional LCA, where accurate selection

by penalized methods is often difficult to achieve (Fan and Lv (2008)). Although the proposed procedure is inspired from an optimization problem, our primary interest centers on the screening accuracy and the misclassification errors, rather than the optimization itself.

The rest of this paper is organized as follows. Section 2 reviews LCA and introduces the HHS-EM procedure. We investigate the theoretical properties of the proposed procedure in Section 3. To numerically justify HHS-EM, we show the simulation results in Section 4 and a real data example in Section 5. We provide some concluding remarks in Section 6. Figures, tables, and proofs are given in Online Supplementary Material.

2. HHS-EM feature screening

2.1 Model setup and notations

We consider the following latent class model with Q classes:

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

where $\mathbf{y}_i = (y_{i1}, \dots, y_{ie})^\top$ is an $e \times 1$ vector of continuous response for the i th subject; $\boldsymbol{\mu}_m = (\mu_{m1}, \dots, \mu_{me})^\top$ and $\boldsymbol{\Sigma}_m$ represent mean vector and covariance matrix corresponding to the m th class, respectively. Notation $\boldsymbol{\omega}_i = (\omega_{i1}, \dots, \omega_{iQ})^\top$ is a $Q \times 1$ vector of binary latent allocation variables, where

$\omega_{im} = 1$ if the i th subject belongs to class m and 0 otherwise for $m = 1, \dots, Q$.

We require $\sum_{m=1}^Q \omega_{im} = 1$. Furthermore, we use $\mathcal{N}(\cdot, \cdot)$ to denote the normal distribution, and define $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_Q\}$, $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_Q\}$.

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

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

where $\boldsymbol{\beta}_m$ is a $p \times 1$ vector of LCA coefficients corresponding to class m for $m = 1, \dots, Q$. When all $\boldsymbol{\beta}_m$ s are zero, none of the features is relevant to the analysis and the above latent class model reduces to a finite Gaussian mixture model (Khalili and Chen (2007)). Following the common practice, we set $\boldsymbol{\beta}_Q = \mathbf{0}$ and assume Q is known. In applications, Q is often specified using prior information and pre-analysis of the data. Throughout this paper, we assume that the number of features p exponentially grows with the sample size n , i.e., $\log(p) = O(n^\tau)$ for some constant $0 \leq \tau < 1$, but only a small number of features have important effects on the latent classes. This amounts

to assuming that coefficient β_m has a sparse structure with many zero entries. Note that, in model (2.1), β_m may vary across classes; consequently, the probabilities of an object belonging to different latent classes may be related to different sets of relevant features.

Let $s_m^* = \{j : \beta_{mj} \neq 0 \text{ for } j = 1, \dots, p\}$ be the index set of the relevant features associated with class m , where β_{mj} is the j th component of β_m , and $s_m^{c*} = \{1, \dots, p\} \setminus s_m^*$ be the index set of the irrelevant features related with class m for $m = 1, \dots, Q - 1$. Denote $s^* = \{s_1^*, \dots, s_{Q-1}^*\}$, $\beta = \{\beta_1, \dots, \beta_{Q-1}\}$, $\Theta = \{\beta, \mu, \Sigma\}$, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ and $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$. 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 Θ for the observed data $\mathbf{Y}$ is given by

$$\ell_n(\Theta) = \sum_{i=1}^n \log \left\{ \sum_{m=1}^Q \pi_{im} \phi(\mathbf{y}_i; \mu_m, \Sigma_m) \right\}, \quad (2.2)$$

where $\phi(\cdot; \mu_m, \Sigma_m)$ denotes the Gaussian density with mean vector μ_m and covariance matrix Σ_m . A common issue with the aforementioned latent class model is the identifiability problem. That is, without further restrictions on the class parameters μ_m and Σ_m , the above considered latent class model with different class parameters may lead to the same model. To address the issue, we adopt the widely used approach to impose restrictions on the class means. Thus, we assume $\mu_{11} < \mu_{21} < \dots < \mu_{Q1}$. More discussions on the

identifiability problem can refer to Frühwirth-Schnatter (2006).

The goal of study is to screen out most irrelevant features associated with zero-effect LCA coefficients from model (2.1) by analyzing the observed data $(\mathbf{Y}, \mathbf{X})$.

2.2 Feature screening in LCA

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

Despite the success of the L_0 -based approaches, less attention has been paid for LCA with a focus on clustering. Different from a regular regression model, the response in model (2.1) is implicitly linked with features via unobservable class labels, where relevant features may vary across different classes. This complication makes the L_0 -based approaches practically less effective due to the algorithm instability. One nature strategy is to use a modified L_0

penalty that helps to improve the stability while maintaining a high screening accuracy.

With the above thinking, we consider a penalized likelihood problem for model (2.1) as follows.

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

where $\lambda > 0$ is a regularization parameter and κ is a 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; 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 a more effective screening in LCA. With the L_0 penalty, users can precisely control the desirable model sparsity via κ . When κ is properly chosen, the resulting sparse estimator suggests no more than κ coefficients that receive the most support from the likelihood; it thus 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 the computational perspective, the optimization problem (2.3) can be represented as the following proposition.

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

$$\begin{aligned} \max_{\Theta, \mathbf{z}} & \left\{ \ell_n(\Theta) - n\lambda \sum_{m=1}^{Q-1} \|\beta_m\|_1 - n\eta \sum_{m=1}^{Q-1} \langle \mathbf{z}_m, |\beta_m| \rangle \right\} \\ \text{s.t. } & \sum_{m=1}^{Q-1} \sum_{j=1}^p z_{mj} = p(Q-1) - \kappa, \quad \mathbf{z}_m = (z_{m1}, \dots, z_{mp})^\top \in \{0, 1\}^p, \end{aligned} \quad (2.4)$$

have the same optimal objective value and the same set of optimal solutions

$\hat{\Theta}$ *when* $\eta > (\max_j \|\mathbf{x}_j\|_2)/n$, *where* $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_{Q-1}\}$.

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 less penalties on LCA coefficients associated with important features for the desired level of sparsity.

Following the common practice in LCA, an EM algorithm can be developed to evaluate maximum likelihood estimates of model parameters. To this end, it follows from Khalili and Chen (2007) that the complete data log-likelihood function of Θ for $\mathbf{Y}$ and $\omega = \{\omega_1, \dots, \omega_n\}$ can be written as

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

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

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

subject to $\|\beta\|_0 \leq \kappa$. Similar to Khalili and Chen (2007), the EM algorithm for evaluating the MLE of Θ is implemented as follows. At the t th iteration with a current value $\Theta^{(t)}$ of Θ , we iteratively update the following two steps:

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

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

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

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

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

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

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

where

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

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

where $\psi = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. 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 constrained condition $\|\boldsymbol{\beta}\|_0 \leq \kappa$ due to the nonconvex optimization problem involved. As an alternative to the above presented EM algorithm, we develop a hybrid method for evaluating MLEs of $\boldsymbol{\Theta}$ by combining the EM iterative procedure and the ADMM algorithm to optimize (2.3). The ADMM algorithm can be regarded as a variant to the augmented Lagrangian method, which 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 considered objective function. It has been applied to solve various nonconvex optimization problems whose objective functions are potentially non-smooth and have some hard constraints. For example, see Boyd et al. (2011); Wang and Yuan (2012); Wang, Yin and Zeng (2019).

To utilize the ADMM to solve problem (2.4), we consider the corresponding augmented Lagrangian optimization problem as follows.

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

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, ρ is some positive scaling parameter. If we set $\boldsymbol{\xi} = \rho \mathbf{u}$, the ADMM for solving (2.6) is

implemented by the following gradient descent iteration procedure

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

where $\boldsymbol{\psi} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, $\mathbf{H}(\boldsymbol{\gamma}; \kappa) = \boldsymbol{\gamma} I(|\boldsymbol{\gamma}| > r_\kappa)$ in which r_κ is the κ th largest component of $|\boldsymbol{\gamma}|$ and $I(\cdot)$ is an indicator function. Generally, 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 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 (e.g., see remarks below Theorem 1). Thus, the above presented 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 presented 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

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

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

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

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

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

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

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

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

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

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

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

which shows that the above presented algorithm can ensure that $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\theta}^{(t)}, \mathbf{u}^{(t)})$ is an increasing function in $\boldsymbol{\Theta}$ and $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\theta}^{(t)}, \mathbf{u}^{(t)})$ 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 the selection problem of the stepsize ν_1 , regularization parameter λ and scaling parameter ρ given in (2.9). A large number of empirical studies show that a smaller value of ν_1 often leads to a faster convergence

of the above presented algorithm. Proposition 2 indicates that only if ν_1 is larger than ς , the objective function $\mathcal{L}(\Theta^{(t)}, \theta^{(t)}, \mathbf{u}^{(t)})$ is guaranteed to increase at each iteration. In applications, one may first initiate a tentatively small ν_1 , and then check condition (2.11). If (2.11) is unsatisfied, we take ν_1 as twice of its current value. Similarly, a small ρ helps to improve the penalized log-likelihood and boost the algorithm convergence. In practice, one may first 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 in comparison with the 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 considered. For the regularization parameter λ , a proper value of λ can accelerate the convergence of the algorithm and lead to a robust estimation of Θ . Generally, one may choose λ by the generalized cross-validation method. For the sparsity parameter κ , an appropriate κ may depend on the specific data structures as well as the nature of model sparsity. Following the argument of Xu and Chen (2014), we take the screening bound $\kappa = 3^{-1} \log(n)n^{1/3}$. Actually, in our simulation studies, we have shown the performance of the HHS-EM algorithm is robust to a wide choices of κ , which facilitates the usage of the HHS-EM algorithm by avoiding an elabo-

rative specification of κ . The proposed HHS-EM algorithm is summarized in Algorithm 1 below.

Algorithm 1 HHS-EM for high-dimensional LCA

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

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

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

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

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

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

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

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

2.3 Post-screening classification

By HHS-EM, one can effectively screen out most irrelevant features for high-dimensional LCA. With the retained features and their estimated LCA co-

efficients from Algorithm 1, one can uncover the hidden class membership ω_i in training data using a Bayesian classification rule. Specifically, let $\hat{\Theta} = (\hat{\beta}, \hat{\mu}, \hat{\Sigma})$ be the estimate of Θ by the HHS-EM procedure. Given the training dataset $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^n$ and $\hat{\Theta}$, it follows from Bayesian formula that the posterior probability of the event $\omega_{im} = 1$ can be estimated by

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

where

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

Accordingly, one may assign the i th unit in the training data to class M based on the following classification rule:

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

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

With the HHS-EM-based LCA, one can also conveniently classify a new case with the information of $\mathbf{x}$ only. Similar to (2.12), one can estimate the posterior probability of the class membership by

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

The corresponding classifier is given by

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

3. Theoretical Properties

In this section, we provide some 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

$$\dot{\ell}_1(\boldsymbol{\beta}; \boldsymbol{\Theta}^{(t)}) = (\mathbf{X}^\top (\boldsymbol{\delta}_{\cdot 1}^{(t)} - \boldsymbol{\pi}_{\cdot 1}), \mathbf{X}^\top (\boldsymbol{\delta}_{\cdot 2}^{(t)} - \boldsymbol{\pi}_{\cdot 2}), \dots, \mathbf{X}^\top (\boldsymbol{\delta}_{\cdot Q-1}^{(t)} - \boldsymbol{\pi}_{\cdot Q-1}))^\top,$$

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

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

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

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

(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

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

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

(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 , $\Xi_n(\Theta)$ is non-positive definite on a given domain $\mathbf{D}_\Theta$ and

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

for any $\Delta \neq 0$ satisfying $\|\Delta_{s_c^*}\|_1 \leq 3\|\Delta_{s^*}\|_1$, and $\Theta_{s^*} = \{\beta_{s^*}, \psi\}$.

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

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 of the HHS-EM algorithm. It implies that the minimum signal is bounded away from zero. Condition (C4) is widely used in high-dimensional data analysis literature (e.g., Xu and Chen (2014)), it holds naturally with the rescaling operation.

Condition (C5) allows that when n is sufficiently large, the observed information matrix $\Xi_n(\Theta)$ is non-positive definite on a given domain D_Θ , which is weaker than that given in Jordan and Xu (1995). Also, the restriction puts a limitation on the set of Δ for which (3.1) is required, so that it is in fact weaker than imposing non-zero eigenvalues. The condition can match restricted eigenvalue condition given in Bickel et al. (2009). Actually, by using the assumption of Wang et al. (2015) for high dimensional mixture model that $\mathbb{E}_{\min}(I(\Theta^*)) > 0$ under the condition (C2), and $\|\Delta_{s_c^*}\|_1 \leq 3\|\Delta_{s^*}\|_1$, we can obtain that the left hand of (3.1) is lower bound by $n\|\Delta_{s^*}\|_2^2\{\mathbb{E}_{\min}(I(\Theta^*)) + o_p(1)\}$ using the similar process of Bickel et al. (2009). Thus the assumption (C5) of restricted eigenvalues for latent class model is reasonable. Condition (C6) assumes that the parameters in the mixture model are in a bounded compact space. The same condition as (C6) can be found in Jiang et al. (2015) and Huang, Peng and Zhang (2017).

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

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

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

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

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

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 high-dimensional setup. The procedure enjoys the sure screening property; that is, with probability tending to one, all relevant features will be retained after screening. Note that, since $\lambda = O(n^{-\alpha})$ 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 initial values in Theorem 1 can be easily satisfied even with a $\boldsymbol{\theta}^{(0)}$ not very close to $\boldsymbol{\beta}^*$. One may thus simply set $\boldsymbol{\theta}^{(0)} = \mathbf{0}$ for 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 (Xu and Chen (2014)); such an accurate initial estimation might not be easy to obtain in high-dimensional LCA. In comparison, the use of hybrid L_0 - L_1 penalty in the proposed method helps to relax this requirement and improve the practical screening accuracy.

We now turn to investigate the classification accuracy for the HHS-EM-based LCA as discussed in Section 2.3. We assess the accuracy of classifier $G_{1;\hat{\Theta}}$ (as defined in (2.13)) by 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 dataset that was used to screen features and estimate the parameters of the classifier. Mathematically, $\hat{R}_1$ is given by

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

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

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

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

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

This optimal rate is given by

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

We assess the classifier $G_{2;\hat{\Theta}}$ (as defined in (2.14)) by the testing misclassification rate $\hat{R}_2$. This rate measures the accuracy of a classifier to predict the

class label of a new case based on its information of $\mathbf{x}$ only. Mathematically, $\hat{R}_2$ is can be expressed as

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

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

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

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

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

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

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

With most irrelevant features screened out by the HHS-EM, we obtain a refined LCA with a manageable set of important features. By Theorem 2, we show that the HHS-EM-based LCA leads to viable classifiers $G_{1;\hat{\Theta}}$ and

$G_{2;\hat{\Theta}}$, which mimic the oracle classifiers $G_{1;\Theta^*}$ and $G_{2;\Theta^*}$, respectively. Our classifiers are asymptotically optimal in the sense that, as $n \rightarrow \infty$, their misclassification rates $\hat{R}_1, \hat{R}_2$ converge to the optimal rates $R_{1,\text{opt}}, R_{2,\text{opt}}$ in probability. In other words, when the sample size is large, the proposed method could conduct effective classification, as if the true model were known in advance.

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 $\tau \rightarrow 0$, 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 $\hat{R}_1 - R_{1,\text{opt}}$ can be nearly up to $O_p(n^{-1})$, which echoes the minmax error bound for clustering of high-dimensional Gaussian mixtures with two components (Cai et al. (2019)).

4. Numerical studies

In this section, several simulation studies are conducted to investigate the finite sample performance of the proposed HHS-EM procedure in terms of screening accuracy and misclassification rate. The simulations are conducted using software **R** on a Windows computer with a 8-core 2.21 GHz CPU and a 16 GB RAM.

In particular, we compare the proposed method with a few popular screen-

ing approaches: MMLE (Fan and Song (2010)), MV-SIS (Cui, Li and Zhong (2015)), and FR (Wang (2009)). Since the above methods are mainly designed for regression models, we modify them for LCA under a EM-framework and call the modified methods as MMLE-EM, MV-EM, FR-EM, respectively. In the modified screening methods, the screening indicator is sequentially estimated based on a marginal LCA model involving one feature at a time. Due to the nature of those methods, we treat the retained features as the relevant ones for all classes and conduct classification using those features based on a refitted LCA. We also include the L_1 -penalized selection method discussed in Houseman et al. (2006) and Wu (2013) in our study and refer to it as Lasso-EM.

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

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

where $\hat{s}_\nu$ denotes the set of retained features for the ν -th 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; PSR is the

averaged proportion of the relevant features that are retained. As relevant features may vary across classes in LCA, we use RC and PSR to indicate the overall value, and use RC_m and PSR_m to denote the class-specific value for class m . A good screening method in LCA is expected to have high RC_m and PSR_m for all classes.

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

We estimate the training misclassification rate by

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

and estimate the testing misclassification rate by

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

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

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

(M1) $\mathbf{y}$ is a 3×1 vector of response variables; the true values of $\boldsymbol{\mu}_m$'s ($m = 1, 2, 3$) 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 $\Sigma_{jj} = 1$, $\Sigma_{ij} = 0.5$ for $i \neq j$, $i, j = 1, 2, 3$. This is a case that the groups have the same correlation structure, however have different mean structures.

(M2) $\mathbf{y}$ is an univariate response variable; the true values of μ_m 's ($m = 1, 2, 3$) are taken as $\mu_1 = -5.0$, $\mu_2 = 0.0$ and $\mu_3 = 5.0$; and the true values of $\Sigma_1, \Sigma_2, \Sigma_3$ are set as $\Sigma_1 = \Sigma_3 = 4.0$ and $\Sigma_2 = 1.0$. This indicates that the considered three groups have completely different means of each other.

With (M1) and (M2), we further consider the following 4 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 4 generated from the index set $\{1, \dots, p\}$ for $m = 1, 2$. Components of $\boldsymbol{\beta}_{s_m^*}$ are independently generated by $U\{6 \log(n)/\sqrt{n} + |Z|/4\}$ and components of $\boldsymbol{\beta}_{s_m^{c*}}$ are taken as zero for $m = 1, 2$, where U is a bernoulli random variable with $\Pr(U = 1) = \Pr(U = -1) = 0.5$, Z is sampled from the standard normal distribution $\mathcal{N}(0, 1)$. Responses are generated from case (M1), while features x_{ik} 's are independently generated from the standard normal distribution $\mathcal{N}(0, 1)$ for $i = 1, \dots, n$ and $k = 1, \dots, p$. Obviously, these features are independent of each other. In this case, we set $(n, p) = (320, 1500)$.

(S2) $s_1^* = \{1, 3, 5\}$, $s_2^* = \{7, 9, 11\}$, $\beta_{s_1^*} = (2.0, 2.0, -1.5)$, and $\beta_{s_2^*} = (-1.5, 2.0, 2.0)$, while components of $\beta_{s_1^{c*}}$ and $\beta_{s_2^{c*}}$ are set as zero. Responses are generated from case (M1), while $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$ are independently sampled from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Upsilon)$ for $i = 1, \dots, n$, where $\Upsilon = (\gamma_{jk})$ with $\gamma_{jj} = 1.0$, $\gamma_{j,j-1} = 2/3$, $\gamma_{j,j-2} = 1/3$ for $j \geq 3$, and $\gamma_{jk} = 0.0$ for $|j - k| \geq 3$. In this case, we take $(n, p) = (300, 1500)$.

(S3) $s_1^* = \{1, 2, 3, 4\}$, $s_2^* = \{1, 2, 3, 5\}$, $\beta_{s_m^*} = (2.0, 2.0, 2.0, 2.0)$ for $m = 1, 2$. Responses are generated from case (M2). Features $\mathbf{x}_i$'s are independently generated from case (S2) with Υ specified by AR(1) structure, i.e., $\gamma_{jk} = 0.5^{|j-k|}$. In this case, we set $(n, p) = (320, 2000)$.

(S4) s_1^* and s_2^* are taken the same values as case (S3), while $\beta_{s_m^*} = (3.0, 3.0, 3.0, -3.0)$ for $m = 1, 2$. Responses are generated from case (M2). Features $\mathbf{x}_i$'s are independently generated from case (S2) with Υ having a compound symmetry (CS) structure, i.e., $\gamma_{jj} = 1.0$, $\gamma_{jk} = 0.5$ 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, while in (S3) and (S4), features 1, 2, 3 are influential on both classes.

In scenario 2, we further consider two model setups (S5) and (S6), where $Q = 6$ and $(n, p) = (400, 1000)$. In these two setups, we generate a univariate response from model (2.1) with $\mu_1 = -4.0$, $\mu_2 = -2.0$, $\mu_3 = 0.0$, $\mu_4 = 2.0$,

$\mu_5 = 4.0$ and $\mu_6 = 6.0$, $\Sigma_1 = \Sigma_6 = 4.0$, $\Sigma_2 = \dots = \Sigma_5 = 1.0$. We set $\beta_{s_m^*} = (2, -2)$, where s_m^* contains two randomly chosen features from $\{1, \dots, p\}$ for $m = 1, \dots, 5$. In (S5), features $\mathbf{x}_i$'s are independently generated 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), features $\mathbf{x}_i$'s are independently generated based on (S4).

We use the proposed HHS-EM as well as its competitors to retain $\kappa = 3^{-1} \log(n)n^{1/3}$ features on datasets 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. It is observed that all methods perform well in (S1), which is the most straightforward setup for feature screening. As the correlation structure among $\mathbf{x}$ gets more complex, the screening accuracy of marginal-effect-based methods (i.e. MMV-EM and MMLE-EM) deteriorates drastically. While Lasso-EM and FR-EM maintain a decent RC in (S2), they seem to not 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 feature screening.

In those two cases, HHS-EM still give us encouraging results by showing the “best-in-group” accuracy for both screening and classification. By the nature of Algorithm 1, the high accuracy of HHS-EM comes with a computational cost, but this is moderate in most cases. Considering the improved accuracy brought by HHS-EM, this small computational investment seems to be quite worthwhile.

We repeat the above simulation study in (S3) with 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 . This merit makes it a viable and robust approach in practice.

5. An real data example

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

The goal here is to identify important social indices (e.g. income, school

engagement, etc.) that could help to partition the communities into sub-groups by their crime rates. To this end, a LCA is conducted based on model (2.1), where the crime rate of a community is used as the response $\mathbf{y}$ and the 100 social indices are treated the covariates (features) $\mathbf{x}$. In the model, we postulate $Q = 3$ latent classes with low, high, and general criminal risk, where the unknown class membership $\mathbf{w}_i$ of community i is to be estimated for $i = 1, \dots, 1994$.

Since we have $p = 100$ features in this case, conducting feature screening would be 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 used HHS-EM and its competitors to retain the most important $\kappa = 10$ features using the training set. We repeat the above procedure over 100 times; a feature is considered to be 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 see that three features are suggested by all the methods. They are

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.

By HHS-EM, we find that x_4 and x_{45} show a positive effect of 0.8 for class **1**, while x_{51} show a positive effect of 1.28 in class **2**. Accordingly, communities with higher percentages of Caucasian population and kids in family with two parents are more likely to belong to class **1**, while communities with higher percentage of kids born to never married couples are more likely to belong to class **2**.

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

To further test HHS-EM, we combine the original 100 features with 1200 synthetic random features. We then re-conduct feature screening on the combined data based on the same procedure as described before with a 50% or 70% of the training data. In Table 3, we show the proportion of times (i.e. RC) that x_4 , x_{45} , x_{51} are still selected as important ones based on 100 repetitions. The results of other screening methods are also reported for the purpose of comparison. It is seen that the proposed HHS-EM shows its superior ro-

bustness 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 by the new screening results based on the combined data. In Table 4, we report the proportion of times that a selected community is classified into the same class as in the original analysis based on 100 repetitions. This is to test whether a screening method leads to robust classification against changes in the data. From Table 4, the promising performance of HHS-EM is again observed.

6. Conclusion and discussion

In this paper, we proposed a new feature screening method HHS-EM for high-dimensional LCA, where accurate selection is often difficult to achieve. The proposed method is inspired from solving a hybrid L_0 - L_1 penalized problem; it allows users to precisely control the number of features to be retained with an significantly enhanced stability. The promising performance of the method is supported by both theory and extensive numerical examples.

Our current work focuses on the cases with normally distributed responses. We have briefly discussed the extension of HHS-EM to LCA with categorical responses in the Online Supplementary Material. It would be an interesting

topic for future research to further extend the existing work to the cases with mixed-type responses. Also, it would be promising to explore the possibility of using hybrid L_0 - L_1 penalization in non-parametric (model-free) clustering.

Supplementary Materials

Technical proofs, tables, and figures in this paper are provided in the Online Supplementary Material.

Acknowledgements

This work was supported in part by NSFC grants 11731011, 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.
- 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.
- 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 and 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*. In press.
- Tang, W., Xie, J., Lin, Y., Tang, N. (2021). Quantile correlation-based variable selection. *Journal of Business & Economic Statistics* DOI:10.1080/07350015.2021.1899932
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Sta-*

tistical 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 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 heterogeneous 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.

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