Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

CHIME: CLUSTERING OF HIGH-DIMENSIONAL GAUSSIAN MIXTURES WITH EM ALGORITHM AND ITS OPTIMALITY*

By T. Tony Cai, Jing Ma, and Linjun Zhang

University of Pennsylvania

Unsupervised learning is an important problem in statistics and machine learning with a wide range of applications. In this paper, we study clustering of high-dimensional Gaussian mixtures and propose a procedure, called CHIME, that is based on the EM algorithm and a direct estimation method for the sparse discriminant vector. Both theoretical and numerical properties of CHIME are investigated. We establish the optimal rate of convergence for the excess mis-clustering error and show that CHIME is minimax rate optimal. In addition, the optimality of the proposed estimator of the discriminant vector is also established. Simulation studies show that CHIME outperforms the existing methods under a variety of settings. The proposed CHIME procedure is also illustrated in an analysis of a glioblastoma gene expression data set and shown to have superior performance.

Clustering of Gaussian mixtures in the conventional low-dimensional setting is also considered. The technical tools developed for the high-dimensional setting are used to establish the optimality of the clustering procedure that is based on the classical EM algorithm.

1. Introduction. Clustering analysis, which aims to partition unlabeled data into homogeneous groups, is an ubiquitous problem in statistics and machine learning with a broad range of applications, including pattern recognition, disease diagnostics, and information retrieval [see 6, 19, and the references therein]. A number of clustering algorithms have been proposed in the literature. The well-known k-means and k-medians algorithms [8] are centroid-based. Hierarchical clustering [37] builds a hierarchy of clusters based on the empirical measures of dissimilarity among sets of observations. Clustering algorithms have also been developed and analyzed under the probabilistic mixture model framework [31, 14]. Among the possible probability distributions for the mixture components, the Gaussian distribution is the most commonly used for both theoretical and computational considerations [15, 23, 7], and has been widely used in a range of applications for clustering and discriminant analysis [16, 30].

In the present paper, we consider clustering of data generated from Gaus-

*The research was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.

MSC 2010 subject classifications: Primary 62G15; secondary 62C20, 62H35

Keywords and phrases: High-dimensional data; Unsupervised learning; Gaussian mixture model; EM algorithm; Mis-clustering error; Minimax optimality

sian mixtures with the focus on the high-dimensional setting. We begin with the following mixture of two p-dimensional Gaussian distributions with equal covariance matrices:

(1.1)
$$Y \sim \begin{cases} 1, & \text{with probability } 1-\omega^* \\ 2, & \text{with probability } \omega^* \end{cases} \quad \text{and} \quad Z \mid Y = k \sim N_p(\boldsymbol{\mu}_k^*, \Sigma^*),\ k = 1, 2.$$

In clustering, Z is observable and Y is not. For identifiability, we assume $\omega^* \in (0, 1/2]$. Suppose we have n unlabeled observations $\boldsymbol{z}^{(i)}$ $(i = 1, \ldots, n)$ generated independently and identically from the mixture in (1.1), that is,

$$\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, ..., \boldsymbol{z}^{(n)} \overset{i.i.d.}{\sim} (1-\omega^*) N_p(\boldsymbol{\mu}_1^*, \Sigma^*) + \omega^* N_p(\boldsymbol{\mu}_2^*, \Sigma^*). \tag{1.2}$$

The goal is to cluster $\boldsymbol{z}^{(i)}$ $(i = 1, \ldots, n)$ into two groups. Although the conventional low-dimensional setting will also be considered later, we are particularly interested in the high-dimensional setting where the dimension p can be much larger than the sample size n.

Clustering analysis is closely connected to classification analysis where the goal is to construct a classifier for future unlabeled observations based on the observed labeled data. In the ideal case where the parameter $\boldsymbol{\theta}^* = \{\omega^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*, \Sigma^*\}$ is known, the optimal classification procedure is the Fisher's linear discriminant rule

$$G_{\boldsymbol{\theta}^*}(\boldsymbol{z}) = \begin{cases} 1, & (\boldsymbol{z} - \frac{\boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^*}{2})^\top \boldsymbol{\beta}^* \geq \log(\frac{\omega^*}{1-\omega^*}) \\ 2, & (\boldsymbol{z} - \frac{\boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^*}{2})^\top \boldsymbol{\beta}^* < \log(\frac{\omega^*}{1-\omega^*}), \end{cases} \tag{1.3}$$

where $\boldsymbol{\beta}^* = \Omega^* \boldsymbol{\delta}^*$, $\Omega^* = (\Sigma^*)^{-1}$ and $\boldsymbol{\delta}^* = \boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$. Let Φ be the cumulative distribution function of the standard normal distribution. Fisher's rule given in (1.3) achieves the optimal mis-classification error

$$\begin{aligned} R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) :=& \mathbb{E}[I(G_{\boldsymbol{\theta}^*}(Z) \neq Y)] \\ =& (1-\omega^*)\Phi\left(\frac{1}{\Delta}\log\frac{\omega^*}{1-\omega^*} - \frac{1}{2}\Delta\right) + \omega^*\bar{\Phi}\left(\frac{1}{\Delta}\log\frac{\omega^*}{1-\omega^*} + \frac{1}{2}\Delta\right), \end{aligned} \tag{1.4}$$

where $\Delta = \sqrt{(\boldsymbol{\delta}^*)^\top \Omega^* \boldsymbol{\delta}^*}$ and $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$. See, for example, [1].

In practice, the parameters $\omega^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*$ and Σ^* are unknown and a data driven method is needed. In the supervised case where the sample labels of $\boldsymbol{z}^{(i)}$ are known, a common approach in the low-dimensional setting is to simply plug the sample values in (1.3). Driven by a wide range of applications, recent focus in clustering and classification has shifted to the high-dimensional setting where p can be much larger than n. In this case, the sample covariance matrix may not even be invertible and it is difficult to estimate the precision matrix Ω^*. [9, 24] proposed to directly estimate the discriminant direction $\boldsymbol{\beta}^* = \Omega^* \boldsymbol{\delta}^*$. More specifically, let $\hat{\boldsymbol{\mu}}_k$ be the sample

mean for class k ($k = 1, 2$) and $\hat{\Sigma}$ be the pooled sample covariance matrix. Assuming that $\boldsymbol{\beta}^*$ is sparse, one can estimate $\boldsymbol{\beta}^*$ directly through the regularized ℓ_1 minimization

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\frac{1}{2}\boldsymbol{\beta}^\top\hat{\Sigma}\boldsymbol{\beta} - \boldsymbol{\beta}^\top(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) + \lambda_n\|\boldsymbol{\beta}\|_1\}, \tag{1.5}$$

where λ_n is a tuning parameter. The classification rule is obtained by using (1.3) with $\boldsymbol{\beta}^*$ replaced by $\hat{\boldsymbol{\beta}}$, $\boldsymbol{\mu}_k^*$ replaced by $\hat{\boldsymbol{\mu}}_k$ for $k = 1, 2$, and ω^* replaced by the sample proportion. This algorithm is easy to implement and avoids estimation of Ω^*.

For unsupervised learning, the class labels are not observed. Compared with the classification analysis, clustering high-dimensional Gaussian mixtures is significantly more complicated, both in terms of the algorithm and in terms of the theoretical analysis. It is not easy to estimate the parameters $\omega^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*$ and Σ^* in the high-dimensional case. In the classical low-dimensional setting, commonly used methods for estimating the parameters include the method of moments [28], spectral method [22], the maximum likelihood, and the Expectation-Maximization (EM) algorithm [29, 4].

In this paper, we introduce CHIME, a clustering procedure for high-dimensional Gaussian mixtures based on the EM algorithm together with the direct estimation idea introduced in [9]. The method uses the posterior probability of $\boldsymbol{z}^{(i)}$ in class k as the 'sample label' of $\boldsymbol{z}^{(i)}$ and efficiently estimates the parameters via the EM algorithm. A key component of the proposed method is to directly estimate and update the discriminant direction $\boldsymbol{\beta}^*$ in each iteration through the regularized ℓ_1 minimization algorithm (1.5). The resulting estimates are subsequently used to yield the discriminant rule as in (1.3). Instead of restricting both the mean vectors and the precision matrix to be sparse, CHIME only requires sparsity of the discriminant vector $\boldsymbol{\beta}^*$.

Both theoretical and numerical properties of the CHIME algorithm are studied. Our analysis first obtains the rate of convergence for estimating $\boldsymbol{\beta}^*$ under the ℓ_2 norm loss, and the convergence rate of the expected excess mis-clustering error $R(\hat{G}_{CHIME}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*})$ (the mis-clustering error is defined later in (2.1)). Furthermore, minimax lower bounds are obtained. The upper and lower bounds together establish the rate optimality of the estimator $\hat{\boldsymbol{\beta}}$ and the CHIME procedure. Specifically, we show that

$$R(\hat{G}_{CHIME}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) \asymp \frac{s\log p}{n},$$

where s is the sparsity of the discriminant vector $\boldsymbol{\beta}^*$, and prove that this rate is optimal. To the best of our knowledge, this is the first optimality result for clustering of high-dimensional Gaussian mixtures and the first construction of a rate-optimal clustering procedure.

In addition to its theoretical optimality, CHIME is computationally easy to implement. The updates of $\hat{\omega}$ and $\hat{\boldsymbol{\mu}}_k$ in the M-step of the EM algorithm

can be calculated analytically, and the update of $\hat{\boldsymbol{\beta}}$ can be implemented via linear programming. Simulation results show that CHIME outperforms existing clustering methods and achieves performance comparable to that of (1.5), which requires the additional label information. The effectiveness of CHIME is also illustrated through an analysis of a glioblastoma gene expression data set, and CHIME yields the smallest error when clustering heterogeneous patients into two distinct subtypes of glioblastoma.

Although the focus of the present paper is on the high-dimensional setting, we also consider clustering of low-dimensional Gaussian mixtures via the CLOME procedure. The technical tools developed for the high-dimensional setting can be used to establish the optimality for the general low-dimensional setting where the covariance matrix is not necessarily the identity matrix.

Our proposed clustering method together with its theoretical optimality guarantees extends the literature on clustering of high-dimensional Gaussian mixtures. [2] considered a special case of (1.1) with $\Sigma^* = \sigma^2\mathbf{I}_p$, $\omega^* = 1/2$, and provided both lower and upper bounds, on the mis-clustering error for sparse $\boldsymbol{\delta}^*$, but the upper bound is not tight. [36] also focused on the special case $\Sigma^* = \sigma^2\mathbf{I}_p$ and $\omega^* = 1/2$, studied the performance of the high-dimensional EM algorithm and established the convergence rate for the estimator of the sparse mean vector. [22] considered the special case where $\Sigma^* = \mathbf{I}_p$ and studied the statistical limits of clustering when the signals are "rare and weak". A phase transition diagram for the IF-PCA method is given in [21]. [3] extended the results in [2] to allow for a general covariance matrix Σ^* and directly estimated the discriminant vector $\boldsymbol{\beta}^*$ via the LPD rule [9]. Using the initial estimates of $\boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*$ and Σ^* provided by [18], they established an upper bound for the mis-clustering error as well as recovery of the support of sparse $\boldsymbol{\beta}^*$ under regularity conditions. Compared to the procedure in [36], our proposed CHIME yields a sparse estimate of $\boldsymbol{\beta}^*$ without the need of truncation, nor does it require sample splitting across iterations.

The rest of the paper is organized as follows. The proposed procedure, CHIME, for clustering high-dimensional Gaussian mixtures is described in detail in Section 2. The theoretical properties are analyzed in Section 3. Both upper and lower bounds are obtained. Together they establish the optimality of CHIME as well as the estimator of discriminant vector $\boldsymbol{\beta}^*$. Section 4 considers clustering low-dimensional Gaussian mixtures based on the classical EM algorithm and establishes the optimality of the clustering procedure by modifying our analysis for the high-dimensional setting. A simulation study is given in Section 5 where we compare the performance of CHIME to other existing clustering methods in the literature. Section 6 uses a real data application to illustrate the merit of CHIME. Section 7 discusses extensions to the multi-class setting. The proofs of the main results are given in Section 8. Proofs of other results together with additional technical details as well as additional simulations are provided in [12].

2. Methodology. In this section, we present in detail the clustering procedure CHIME under the two-component Gaussian mixture model (1.2).

We begin with notations. Throughout the paper, $X, Y, Z, \ldots$ denote random vectors and $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \ldots$ denote their realizations. For $a, b \in \mathbb{R}$, we denote by $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, 2, ..., n\}$. For a vector $\boldsymbol{x} \in \mathbb{R}^p$, the usual vector ℓ_0, ℓ_1, ℓ_2 and ℓ_∞ norms are denoted respectively by $\|\boldsymbol{x}\|_0, \|\boldsymbol{x}\|_1, \|\boldsymbol{x}\|_2$ and $\|\boldsymbol{x}\|_\infty$. Here the ℓ_0 norm counts the number of nonzero entries in a vector. We use $\mathrm{supp}(\boldsymbol{x})$ to denote the support of the vector $\boldsymbol{x}$. The Frobenius norm of a matrix $A = (a_{ij})$ is defined as $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$. The matrix ℓ_1 and ℓ_2 norms are defined, respectively, as $\|A\|_1 = \sup_{\|\boldsymbol{x}\|_1=1} \|A\boldsymbol{x}\|_1$ and $\|A\|_2 = \sup_{\|\boldsymbol{x}\|_2=1} \|A\boldsymbol{x}\|_2$. The matrix ℓ_0 norm is defined similarly to the vector ℓ_0 norm as $\|A\|_0 = |\{(i, j) : a_{ij} \neq 0\}|$, where $|\cdot|$ denotes the cardinality here. The vector ℓ_∞ norm on matrix A is $|A|_\infty = \max_{i,j} |A_{ij}|$. For a symmetric matrix A, we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote respectively the largest and smallest eigenvalue of A. We say $A \succ 0$ if A is positive definite. The inner product between two matrices A and B is defined as $\langle A, B\rangle = \mathrm{tr}(A^\top B)$. For a set $\mathcal{A}$, we use $\mathcal{A}^c$ to denote its complement, and use $I(\mathcal{A})$ to denote its corresponding indicator function. For a positive integer $s < p$, let $\Gamma(s) = \{\boldsymbol{u} \in \mathbb{R}^p : 2\|\boldsymbol{u}_{S^C}\|_1 \leq 4\|\boldsymbol{u}_S\|_1 + 3\sqrt{s}\|\boldsymbol{u}\|_2$, for some $S \subset [p]$ with $|S| = s\}$. For a vector $\boldsymbol{x} \in \mathbb{R}^p$ and a matrix $A \in \mathbb{R}^{m\times p}$, we define $\|\boldsymbol{x}\|_{2,s} = \sup_{\|\boldsymbol{y}\|_2=1, \boldsymbol{y}\in\Gamma(s)} |\boldsymbol{x}^\top \boldsymbol{y}|$ and $\|A\|_{2,s} = \sup_{\|\boldsymbol{y}\|_2=1, \boldsymbol{y}\in\Gamma(s)} \|A\boldsymbol{y}\|_2$. For two sequences of positive numbers a_n and b_n, $a_n \lesssim b_n$ means that for some constant $c > 0$, $a_n \leq cb_n$ for all n, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Finally, we use $c_0, c_1, c_2, C_1, C_2, \ldots$ to denote generic positive constants that may vary from place to place.

2.1. *The Gaussian mixture model.* Suppose we have n observations $\{\boldsymbol{z}^{(1)} \ldots, \boldsymbol{z}^{(n)}\}$ generated independently and identically from the p-dimensional Gaussian mixture model in (1.2) without knowing labels $(y_1, ..., y_n)$, and wish to cluster the observations $\{\boldsymbol{z}^{(1)} \ldots, \boldsymbol{z}^{(n)}\}$ into two groups. The accuracy of a clustering rule $G : \boldsymbol{z}^{(i)} \to \{1, 2\}$, $i = 1, ..., n$, is measured by the expected mis-clustering error,

$$R(G) = \min_{\pi\in\mathcal{P}_2} \mathbb{E}[I(G(\boldsymbol{z}) \neq \pi(y))], \tag{2.1}$$

where $\mathcal{P}_2 = \{\pi : [1, 2] \to [1, 2]\}$ is a set of permutation function, and y is the latent label of a future observation $\boldsymbol{z}$.

As mentioned in the introduction, for this clustering problem, it is important to first estimate the parameters $\omega^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*$ and Σ^* in (1.2). In the classical setting where p is much smaller than n, it has been shown that the maximum likelihood estimator (MLE) performs well under mild conditions

[4]. The joint log-likelihood of the data $\boldsymbol{z}^{(i)}$ $(i = 1, \ldots, n)$ can be written as (2.2)

$$L(\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma; \boldsymbol{z}) = \frac{1}{n} \sum_{i=1}^{n} \log \left\{ (1-\omega) f(\boldsymbol{z}^{(i)} \mid \boldsymbol{\mu}_1, \Sigma) + \omega f(\boldsymbol{z}^{(i)} \mid \boldsymbol{\mu}_2, \Sigma) \right\},$$

where $f(\cdot \mid \boldsymbol{\mu}_k, \Sigma)$ represents the density function of $N_p(\boldsymbol{\mu}_k, \Sigma)$. The MLE maximizes the joint log-likelihood function $L(\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma; \boldsymbol{z})$.

When p is large, direct optimization of the log-likelihood in (2.2) becomes infeasible due to the nonconvexity of the objective function $L(\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma; \boldsymbol{z})$. Moreover, the MLE does not even exist in the high-dimensional setting where $p \gg n$. In this paper, we propose to explore the sparsity of the discriminant vector $\boldsymbol{\beta}^*$ as in [9] for the supervised case by noting that the discriminant rule in (1.3) depends on Σ^* only through $\boldsymbol{\beta}^*$. Further, we adopt the EM algorithm [13] to address the nonconvexity of the joint log-likelihood.

2.2. *A clustering procedure based on the EM algorithm.* To simplify the notation, under the mixture model (1.2), we denote $\boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$, and let $\boldsymbol{\beta} = \Omega(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ with $\Omega = \Sigma^{-1}$. For a given $\boldsymbol{\theta}$, we use $\mathbb{E}_{\boldsymbol{\theta}}$ and $\mathbb{P}_{\boldsymbol{\theta}}$ to denote the expectation and probability under the model (1.2) with respect to the parameter $\boldsymbol{\theta}$. In addition, sometimes we write $\mathbb{E}_{\boldsymbol{\theta}^*}$, $\mathbb{P}_{\boldsymbol{\theta}^*}$ as $\mathbb{E}$ and $\mathbb{P}$ when there is no ambiguity.

Note that if the true labels $\boldsymbol{y} = \{y_i\}_{i=1}^n \in \{1, 2\}^n$ were observed together with $\boldsymbol{z} = \{\boldsymbol{z}^{(i)}\}_{i=1}^n$, the log-likelihood of the complete data is given by

$$L_C(\boldsymbol{\theta}; y, \boldsymbol{z}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{2} I(y_i = k) \left\{ \log f(\boldsymbol{z}^{(i)} \mid \boldsymbol{\mu}_k, \Sigma) + \log \mathbb{P}_{\boldsymbol{\theta}}(y_i = k) \right\}.$$

To address the nonconvexity of the joint log-likelihood, we use the EM algorithm, which iterates between two goals: classification given the parameters, and estimation given the labels. In the t-th iteration, given the estimated $\hat{\boldsymbol{\theta}}^{(t)} = (\hat{\omega}^{(t)}, \hat{\boldsymbol{\mu}}_1^{(t)}, \hat{\boldsymbol{\mu}}_2^{(t)}, \hat{\Sigma}^{(t)})$ from the previous step, the E-step can be interpreted as classifying the observed data $\boldsymbol{z}^{(i)}$ by assuming the true parameter is $\hat{\boldsymbol{\theta}}^{(t)}$. The posterior probability of the i-th sample in class 2 given the observed data $\boldsymbol{z}^{(i)}$ can be calculated as

$$\begin{aligned} \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)}) =& \mathbb{P}_{\hat{\boldsymbol{\theta}}^{(t)}}(y_i = 2 | \boldsymbol{z}^{(i)}) \\ =& \frac{\hat{\omega}^{(t)}}{\hat{\omega}^{(t)} + (1 - \hat{\omega}^{(t)}) \exp\left\{ (\hat{\Omega}^{(t)}(\hat{\boldsymbol{\mu}}_2^{(t)} - \hat{\boldsymbol{\mu}}_1^{(t)}))^\top \left(\boldsymbol{z}^{(i)} - \frac{\hat{\boldsymbol{\mu}}_1^{(t)} + \hat{\boldsymbol{\mu}}_2^{(t)}}{2} \right) \right\}}. \end{aligned} \tag{2.3}$$

We then calculate the expectation of the log-likelihood, with respect to the conditional distribution of y given $\boldsymbol{z}$ under the current estimate of the pa-

rameters $\hat{\boldsymbol{\theta}}^{(t)}$, as

$$\begin{aligned} Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)}) &= \mathbb{E}_{\hat{\boldsymbol{\theta}}^{(t)}}[\log L_C(\boldsymbol{\theta}; y, \boldsymbol{z}) \mid \boldsymbol{z}] \\ &= -\frac{1}{2n}\sum_{i=1}^{n}\Big\{(1-\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)}))(\boldsymbol{z}^{(i)}-\boldsymbol{\mu}_1)^\top\Omega(\boldsymbol{z}^{(i)}-\boldsymbol{\mu}_1) \\ &\qquad + \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)})(\boldsymbol{z}^{(i)}-\boldsymbol{\mu}_2)^\top\Omega(\boldsymbol{z}^{(i)}-\boldsymbol{\mu}_2)\Big\} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\Big\{(1-\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)}))\log(1-\omega) + \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)})\log\omega\Big\} + \frac{1}{2}\log|\Omega|. \end{aligned}$$

The M-step proceeds by maximizing $Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$ given $\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)})$, and is interpreted as parameter estimation given the labels. The maximizer,

$$\hat{\boldsymbol{\theta}}^{(t+1)} = M_n(\hat{\boldsymbol{\theta}}^{(t)}) = \arg\max_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)}), \tag{2.4}$$

can be calculated analytically. We now derive the exact analytic form for the M-step in the t-th iteration, which is used to obtain updates of $\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and Σ. It is straightforward to define and calculate

$$\hat{\omega}^{(t+1)} = \hat{\omega}(\hat{\boldsymbol{\theta}}^{(t)}) = \frac{1}{n}\sum_{i=1}^{n}\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)}), \tag{2.5}$$

$$\hat{\boldsymbol{\mu}}_1^{(t+1)} = \hat{\boldsymbol{\mu}}_1(\hat{\boldsymbol{\theta}}^{(t)}) = \Big\{n - \sum_{i=1}^{n}\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)})\Big\}^{-1}\Big\{\sum_{i=1}^{n}(1-\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)}))\boldsymbol{z}^{(i)}\Big\}, \tag{2.6}$$

$$\hat{\boldsymbol{\mu}}_2^{(t+1)} = \hat{\boldsymbol{\mu}}_2(\hat{\boldsymbol{\theta}}^{(t)}) = \Big\{\sum_{i=1}^{n}\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)})\Big\}^{-1}\Big\{\sum_{i=1}^{n}\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)})\boldsymbol{z}^{(i)}\Big\}. \tag{2.7}$$

This leads to a solution for $\hat{\Sigma}^{(t+1)}$ given by

$$\begin{aligned} \hat{\Sigma}^{(t+1)} = \hat{\Sigma}(\hat{\boldsymbol{\theta}}^{(t)}) = \frac{1}{n}\sum_{i=1}^{n}\Big\{&(1-\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)}))(\boldsymbol{z}^{(i)}-\hat{\boldsymbol{\mu}}_1^{(t+1)})(\boldsymbol{z}^{(i)}-\hat{\boldsymbol{\mu}}_1^{(t+1)})^\top + \\ &\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)})(\boldsymbol{z}^{(i)}-\hat{\boldsymbol{\mu}}_2^{(t+1)})(\boldsymbol{z}^{(i)}-\hat{\boldsymbol{\mu}}_2^{(t+1)})^\top\Big\}. \end{aligned} \tag{2.8}$$

Note that in the high-dimensional setting where $p \gg n$, $\hat{\Sigma}^{(t+1)}$ is singular and cannot be used directly in (1.3) and (2.3) to obtain a clustering rule and $\gamma(\boldsymbol{z}^{(i)})$. Instead of estimating the covariance matrix Σ^*, we estimate the

discriminant vector $\boldsymbol{\beta}^*$ directly. The update $\hat{\boldsymbol{\beta}}^{(t+1)}$ can be obtained through the regularized ℓ_1 minimization

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\frac{1}{2}\boldsymbol{\beta}^\top\hat{\Sigma}^{(t+1)}\boldsymbol{\beta} - \boldsymbol{\beta}^\top(\hat{\boldsymbol{\mu}}_1^{(t+1)} - \hat{\boldsymbol{\mu}}_2^{(t+1)}) + \lambda_n^{(t+1)}\|\boldsymbol{\beta}\|_1\}, \tag{2.9}$$

where $\lambda_n^{(t+1)}$ is the tuning parameter. It is shown in the supplement [12] that the sequence $\lambda_n^{(t+1)} = \kappa^t \cdot C_1\frac{d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)},\boldsymbol{\theta}^*)}{\sqrt{s}} + (\frac{1-\kappa^{t+1}}{1-\kappa})C_\lambda\sqrt{\frac{\log p}{n}}$, for some constants $C_1, C_\lambda > 0$, and $\kappa \in (0, 1/2)$ is an appropriate choice for tuning parameters. In practice, $\lambda_n^{(t+1)}$ can be chosen by cross validation.

As a result, in the high-dimensional setting, the update of $\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)})$ in the E-step is different from (2.3) and proposed to be

$$\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)}) := \mathbb{P}_{\hat{\boldsymbol{\theta}}^{(t)}}(y_i = 2|\boldsymbol{z}^{(i)}) = \frac{\hat{\omega}^{(t)}}{\hat{\omega}^{(t)} + (1-\hat{\omega}^{(t)})\exp\left\{(\hat{\boldsymbol{\beta}}^{(t)})^\top\left(\boldsymbol{z}^{(i)} - \frac{\hat{\boldsymbol{\mu}}_1^{(t)}+\hat{\boldsymbol{\mu}}_2^{(t)}}{2}\right)\right\}}. \tag{2.10}$$

Given a suitable initialization, the EM algorithm iterates between the E-step and M-Step as described above, and terminates in, say T_0, steps. Once the final estimates of $\boldsymbol{\theta}^*$ and $\boldsymbol{\beta}^*$ are obtained, the clustering rule can be constructed by plugging them into the Fisher's rule (1.3). We call this procedure CHIME for **C**lustering of **HI**gh-dimensional Gaussian **M**ixtures with the **E**M, which is summarized in Algorithm 1.

REMARK 1. CHIME requires the initialization $\hat{\boldsymbol{\theta}}^{(0)}$ to be reasonably good to ensure the convergence of $\hat{\boldsymbol{\theta}}^{(t)}$ to an optimum near the true parameters $\boldsymbol{\theta}^*$. We address the issue of initialization in Section 3. The total number of iterations T_0 needs to be specified. It is shown in Section 3 that $T_0 \asymp \log n$ is sufficient to yield the optimal convergence rate for $\hat{\boldsymbol{\beta}}^{(T_0)}$. In practice, it is recommended to run Algorithm 1 until the distance between $\hat{\boldsymbol{\theta}}^{(t+1)}$ and $\hat{\boldsymbol{\theta}}^{(t)}$ is less than a pre-specified tolerance level. In addition, Algorithm 1 requires specifying the contraction constant κ as well as constants C_1 and C_λ. The choice of the tuning parameter in the form of $\lambda_n^{(0)}$ and (2.11) is necessary for establishing convergence of $\hat{\boldsymbol{\beta}}^{(T_0)}$ to the true parameter $\boldsymbol{\beta}^*$, and will be discussed in detail in Section 3.

3. Theoretical Analysis. In this section, we study the properties of the estimator $\hat{\boldsymbol{\beta}}^{(T_0)}$ and the performance of the CHIME clustering rule $\hat{G}_{CHIME}$ proposed in Algorithm 1. We first establish the rates of convergence for the estimation and mis-clustering error and then provide matching minimax lower bounds. These results together show the optimality of CHIME as well as the proposed estimator of the discriminant vector $\boldsymbol{\beta}^*$.

We first introduce the parameter space. For parameters $\boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\omega}, \tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\mu}}_2, \tilde{\Sigma})$, define their $\ell_{2,s}$ distance by

$$d_{2,s}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = |\omega - \tilde{\omega}| \vee \|\boldsymbol{\mu}_1 - \tilde{\boldsymbol{\mu}}_1\|_{2,s} \vee \|\boldsymbol{\mu}_2 - \tilde{\boldsymbol{\mu}}_2\|_{2,s} \vee \|(\Sigma - \tilde{\Sigma})\tilde{\boldsymbol{\beta}}\|_{2,s}. \tag{3.1}$$

Algorithm 1 Clustering of **HI**gh-dimensional Gaussian **M**ixtures with the **E**M (CHIME)

Inputs: Initializations $\hat{\omega}^{(0)}, \hat{\boldsymbol{\mu}}_1^{(0)}, \hat{\boldsymbol{\mu}}_2^{(0)}$ and $\hat{\Sigma}^{(0)}$, maximum number of iterations T_0, and tuning parameters $\kappa \in (0,1), C_d, C_\lambda > 0$. Set

$$\hat{\boldsymbol{\beta}}^{(0)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\boldsymbol{\beta}^\top \hat{\Sigma}^{(0)}\boldsymbol{\beta} - \boldsymbol{\beta}^\top(\hat{\boldsymbol{\mu}}_1^{(0)} - \hat{\boldsymbol{\mu}}_2^{(0)}) + \lambda_n^{(0)}\|\boldsymbol{\beta}\|_1 \right\},$$

where the tuning parameter $\lambda_n^{(0)} = C_d + C_\lambda\sqrt{\log p/n}$.

for $t = 0, 1, \ldots, T_0 - 1$ **do**

E-Step: Evaluate $Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$ with $\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)})$ defined in (2.10).

M-Step: Update $\hat{\omega}^{(t+1)}, \hat{\boldsymbol{\mu}}_1^{(t+1)}, \hat{\boldsymbol{\mu}}_2^{(t+1)}$, and $\hat{\Sigma}^{(t+1)}$ via (2.5), (2.6), (2.7) and (2.8), and $\hat{\boldsymbol{\beta}}^{(t+1)}$ via

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\boldsymbol{\beta}^\top \hat{\Sigma}^{(t+1)}\boldsymbol{\beta} - \boldsymbol{\beta}^\top(\hat{\boldsymbol{\mu}}_1^{(t+1)} - \hat{\boldsymbol{\mu}}_2^{(t+1)}) + \lambda_n^{(t+1)}\|\boldsymbol{\beta}\|_1 \right\},$$

with

$$\lambda_n^{(t+1)} = \kappa\lambda_n^{(t)} + C_\lambda\sqrt{\frac{\log p}{n}}. \tag{2.11}$$

end for

Output $\hat{\omega}^{(T_0)}, \hat{\boldsymbol{\mu}}_1^{(T_0)}, \hat{\boldsymbol{\mu}}_2^{(T_0)}$ and $\hat{\boldsymbol{\beta}}^{(T_0)}$.

Construct the clustering rule

$$\hat{G}_{CHIME}(\boldsymbol{z}) = \begin{cases} 1, & \{\boldsymbol{z} - \frac{\hat{\boldsymbol{\mu}}_1^{(T_0)} + \hat{\boldsymbol{\mu}}_2^{(T_0)}}{2}\}^\top \hat{\boldsymbol{\beta}}^{(T_0)} \geq \log(\frac{\hat{\omega}^{(T_0)}}{1-\hat{\omega}^{(T_0)}}), \\ 2, & \{\boldsymbol{z} - \frac{\hat{\boldsymbol{\mu}}_1^{(T_0)} + \hat{\boldsymbol{\mu}}_2^{(T_0)}}{2}\}^\top \hat{\boldsymbol{\beta}}^{(T_0)} < \log(\frac{\hat{\omega}^{(T_0)}}{1-\hat{\omega}^{(T_0)}}). \end{cases}$$

We shall write $d_{2,s}(\boldsymbol{\theta})$ for $d_{2,s}(\boldsymbol{\theta}, \mathbf{0})$, and consider the following parameter space

$$\Theta_p(s, c_\omega, M, M_b) = \{\boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p\times p}, \Sigma \succ 0, \|\boldsymbol{\beta}\|_0 \leq s, \tag{3.2}$$
$$\omega \in (c_\omega, 1-c_\omega), M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M, \|\boldsymbol{\beta}\|_1 \leq M_b\}.$$

This is a natural parameter space to consider. The condition on the eigenvalues of Σ is standard. For example, it has been used in [11], [5] and [10] for estimation of precision matrices, covariance matrices, and regression coefficients, respectively. Condition on $\|\boldsymbol{\beta}\|_1$ were also similarly used in [27] and [32] for discriminant analysis.

3.1. *Upper bounds.* We need two technical conditions before stating the properties of the clustering algorithm.

Recall that in (1.4), $\Delta = \sqrt{(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*)^\top(\Sigma^*)^{-1}(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*)}$ is the Mahalanobis distance between $\boldsymbol{\mu}_1^*$ and $\boldsymbol{\mu}_2^*$ with covariance matrix Σ^*, and can be interpreted as the Signal-to-Noise Ratio. For constants $c_0, c_1, C_b > 0$ and

$c_0 \le c_\omega, c_1 < 1$, the contraction basin $B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)$ is defined as

(3.3)
$$\begin{aligned} B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s) = \big\{ &\boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p\times p}, \Sigma \succ 0, \\ &\omega \in (c_0, 1-c_0), (1-c_1)\Delta^2 < |\delta_1(\boldsymbol{\beta})|, |\delta_2(\boldsymbol{\beta})|, \sigma^2(\boldsymbol{\beta}) < (1+c_1)\Delta^2, \\ &\boldsymbol{\beta} - \boldsymbol{\beta}^* \in \Gamma(s), \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \le C_b\Delta, \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^*\|_{2,s}, \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^*\|_{2,s} \le C_b\Delta \big\}, \end{aligned}$$

where $\delta_1(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top(\boldsymbol{\mu}_1^* - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})$, $\delta_2(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top(\boldsymbol{\mu}_2^* - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})$, and $\sigma(\boldsymbol{\beta}) = \sqrt{\boldsymbol{\beta}^\top\Sigma^*\boldsymbol{\beta}}$.

The following conditions are needed to establish the convergence of $\hat{\boldsymbol{\beta}}^{(T_0)}$.
(C1) The initial value $\hat{\boldsymbol{\theta}}^{(0)}$ satisfies that

$$d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) \vee \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\| \le r\Delta, \quad \hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^* \in \Gamma(s)$$

with $r < \frac{|c_0 - c_\omega|}{\Delta} \wedge \frac{\sqrt{9M+16c_1} - \sqrt{9M}}{4} \wedge \sqrt{\frac{c_1}{M}} \wedge \frac{C_b}{5\sqrt{s}}$.

In fact, condition **(C1)** guarantees that $\boldsymbol{\theta}^{(t)} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)$ for $t \ge 0$ in Algorithm 1, which is shown in Lemma A.2 and proved in the supplement [12]. We will discuss in Section 3.2 an initialization algorithm whose output satisfies Condition **(C1)**.

(C2) The Signal-to-Noise Ratio Δ satisfies that

$$\Delta > C(c_0, c_1, M, C_b), \tag{3.4}$$

where $C(c_0, c_1, M, C_b)$ is a constant that only depends on the c_0, c_1, M, and C_b, and is given in (C.24) in the supplement [12].

Before we state the main results, we introduce two technical lemmas that characterize the properties of the population version of the proposed CHIME algorithm under Conditions **(C1)** and **(C2)**. We define the respective population version of M-step as follows.

Let $M(\boldsymbol{\theta}) = (\omega(\boldsymbol{\theta}), \boldsymbol{\mu}_1(\boldsymbol{\theta}), \boldsymbol{\mu}_2(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}))$ denote the population version of $M_n(\boldsymbol{\theta})$, the estimator evaluated in (2.4). More specifically,

$$M(\boldsymbol{\theta}) = \arg\max_{\tilde{\boldsymbol{\theta}}} Q(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) := \arg\max_{\tilde{\boldsymbol{\theta}}} \mathbb{E}_{\boldsymbol{\theta}^*}[Q_n(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta})]. \tag{3.5}$$

By definition, $M(\boldsymbol{\theta})$ can be analytically expressed as

(3.6)
$$\omega(\boldsymbol{\theta}) = \mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)], \quad \boldsymbol{\mu}_1(\boldsymbol{\theta}) = \frac{\mathbb{E}[(1-\gamma_{\boldsymbol{\theta}}(Z))Z]}{\mathbb{E}[1-\gamma_{\boldsymbol{\theta}}(Z)]}, \quad \boldsymbol{\mu}_2(\boldsymbol{\theta}) = \frac{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)Z]}{\mathbb{E}[\gamma_{\boldsymbol{\theta}}(Z)]},$$

(3.7)
$$\Sigma(\boldsymbol{\theta}) = \mathbb{E}\big[(1-\gamma_{\boldsymbol{\theta}}(Z))(Z-\boldsymbol{\mu}_1(\boldsymbol{\theta}))(Z-\boldsymbol{\mu}_1(\boldsymbol{\theta}))^\top + \gamma_{\boldsymbol{\theta}}(Z)(Z-\boldsymbol{\mu}_2(\boldsymbol{\theta})(Z-\boldsymbol{\mu}_2(\boldsymbol{\theta}))^\top\big].$$

Using the above definition of the population version updates, we then introduce the following two lemmas, Lemma 3.1 characterizes the linear

convergence of the population EM updates, and Lemma 3.2 captures the distance between the sample and population version estimates. These two lemmas are the key steps in the proof of the main result Theorem 3.1.

LEMMA 3.1 (Contraction on the population iteration). *Suppose* $\boldsymbol{\theta}^* \in \Theta_p(s, c_\omega, M, M_b)$. *If* $\Delta > C(c_0, c_1, M, C_b)$, *where* $C(c_0, c_1, M, C_b)$ *is given in (C.24) in the supplement [12]. Then there exists* $0 < \kappa_0 < \frac{1}{2 \vee (16M)}$, *such that for* $\boldsymbol{\theta} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)$,

$$d_2(M(\boldsymbol{\theta}), \boldsymbol{\theta}^*) \leq \kappa_0 \cdot (d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2), \tag{3.8}$$

where $d_2(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = |\omega - \tilde{\omega}| \vee \|\boldsymbol{\mu}_1 - \tilde{\boldsymbol{\mu}}_1\|_2 \vee \|\boldsymbol{\mu}_2 - \tilde{\boldsymbol{\mu}}_2\|_2 \vee \|(\Sigma - \tilde{\Sigma})\tilde{\boldsymbol{\beta}}\|_2$.

REMARK 2. This theorem implies that

$$d_{2,s}(M(\boldsymbol{\theta}), \boldsymbol{\theta}^*) \leq \kappa_0 \cdot (d_{2,s}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \vee \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2).$$

LEMMA 3.2 (Uniform concentration inequality). *Suppose* $\boldsymbol{\theta}^* \in \Theta_p(s, c_\omega, M, M_b)$ *with* $c_\omega \in (0, 1)$ *and* M, M_b *universally bounded. Under the condition* **(C1)**, *there exists a constant* $C_{con} > 0$, *such that with probability at least* $1 - o(1)$,

$$\sup_{\boldsymbol{\theta} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)} d_{2,s}(M_n(\boldsymbol{\theta}), M(\boldsymbol{\theta})) \leq C_{con} \sqrt{\frac{s \log p}{n}};$$

$$\sup_{\boldsymbol{\theta} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)} \|(\hat{\Sigma}(\boldsymbol{\theta}) - \Sigma(\boldsymbol{\theta}))\boldsymbol{\beta}^*\|_\infty \leq C_{con} \sqrt{\frac{\log p}{n}}.$$

The above two lemmas imply that at each iteration, $\hat{\boldsymbol{\theta}}^{(t)}$ converges geometrically to the truth $\boldsymbol{\theta}^*$, until their distance is indistinguishable with the statistical error, whose rate is characterized by Lemma 3.2.

In addition, we point out that the inequality in (3.8) quantifies the contraction w.r.t the $\ell_{2,s}$-norm of the distance between the population EM update and the true parameter $\boldsymbol{\theta}^*$. This contraction property is different from the ones used in [4, 36, 39]. Consequently, our subsequent analysis differs from theirs. Indeed, existing works use the ℓ_2 or ℓ_∞-norm of the distance between the EM update and the true parameter to define the contraction. The advantage with the $\ell_{2,s}$-norm is that it characterizes a more refined sparsity-based difference, which converges at the rate $\sqrt{s \log p / n}$ by Lemma 3.2. The ℓ_2 or ℓ_∞-norm used in previous works is not suitable for our purpose.

We are now ready to state the first main result. The following theorem shows that under Conditions **(C1)** and **(C2)**, the estimate $\hat{\boldsymbol{\beta}}^{(T_0)}$ provided by Algorithm 1 converges to the true parameter $\boldsymbol{\beta}^*$.

THEOREM 3.1. *Suppose we observe* n *i.i.d. samples* $\{\boldsymbol{z}^{(1)} \ldots, \boldsymbol{z}^{(n)}\}$ *from model (1.2) with the true parameter* $\boldsymbol{\theta}^* \in \Theta_p(s, c_\omega, M, M_b)$, *for some constant* $c_\omega \in (0, 1)$ *and universally bounded constants* $M, M_b > 0$ *and* $s =$

$o(\sqrt{n/\log p})$. *Assume that conditions* **(C1)** *and* **(C2)** *hold with* r *satisfying* $\sqrt{s\log p/n} = o(r)$. *Then there exist constants* $C_d, C_\lambda > 0$, $\kappa \in (0, 1/2)$, *such that the output* $\hat{\boldsymbol{\beta}}^{(T_0)}$ *of Algorithm 1 with tuning parameters* C_d, C_λ, κ *satisfies, with probability* $1 - o(1)$,

$$\|\hat{\boldsymbol{\beta}}^{(T_0)} - \boldsymbol{\beta}^*\|_2 \lesssim \kappa^{T_0} d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*) + \sqrt{\frac{s\log p}{n}}.$$

Consequently, if $T_0 \gtrsim (-\log(\kappa))^{-1}\log(n \cdot d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*))$, *then*

$$\|\hat{\boldsymbol{\beta}}^{(T_0)} - \boldsymbol{\beta}^*\|_2 \lesssim \sqrt{\frac{s\log p}{n}}. \tag{3.9}$$

The proof of Theorem 3.1 relies on Lemmas 3.1 and 3.2. The idea of proving Theorem 3.1 by establishing the contraction and uniform concentration properties is similar to that in Balakrishnan et al. [4] for the conventional low-dimensional setting. However, establishing such results in the high-dimensional setting is quite challenging. The proof of Lemmas 3.1 and 3.2 are involved and are given in the supplement [12].

REMARK 3. In comparison with the results in [36, 39], which consider the high-dimensional EM algorithm under the special Gaussian mixture model $\frac{1}{2}N_p(-\boldsymbol{\mu}^*, \sigma^2\mathbf{I}_p) + \frac{1}{2}N_p(\boldsymbol{\mu}^*, \sigma^2\mathbf{I}_p)$, Theorem 3.1 establishes a faster convergence rate under a more general model. In fact, [36] and [39] show the convergence rate $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|_2 \lesssim \sqrt{s\log p\log n/n}$ for their estimator $\hat{\boldsymbol{\mu}}$ and require sample splitting. In the present paper, we remove the $\log n$ factor and establish that $\|\hat{\boldsymbol{\beta}}^{(T_0)} - \boldsymbol{\beta}^*\|_2 \lesssim \sqrt{s\log p/n}$ by using a uniform concentration inequality (Lemma 3.2) and thus avoid the need for sample splitting. The idea of using uniform concentration results is similar to that in [4], but the techniques to prove this uniform concentration is much more involved in the high-dimensional setting.

We now turn to the performance of the clustering rule given by Algorithm 1. For ease of presentation, we denote the final output $\hat{\boldsymbol{\theta}}^{(T_0)}$ and $\hat{\boldsymbol{\beta}}^{(T_0)}$ of Algorithm 1 by $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ respectively. Recall that in Algorithm 1, after obtaining the final estimates $\hat{\omega}, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2$ and $\hat{\boldsymbol{\beta}}$, we construct the following clustering rule

$$\hat{G}_{CHIME}(\boldsymbol{z}) = \begin{cases} 1, & (\boldsymbol{z} - \hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\beta}} \geq \log(\frac{\hat{\omega}}{1-\hat{\omega}}), \\ 2, & (\boldsymbol{z} - \hat{\boldsymbol{\mu}})^\top\hat{\boldsymbol{\beta}} < \log(\frac{\hat{\omega}}{1-\hat{\omega}}), \end{cases} \tag{3.10}$$

where $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2$. By (2.1), we obtain

$$R(\hat{G}_{CHIME}) = (1-\omega^*)\Phi\Big(\frac{\log(\frac{\hat{\omega}}{1-\hat{\omega}}) + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1^*)^\top\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\boldsymbol{\beta}}^\top\Sigma^*\hat{\boldsymbol{\beta}}}}\Big) + \omega^*\bar{\Phi}\Big(\frac{\log(\frac{\hat{\omega}}{1-\hat{\omega}}) + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2^*)^\top\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\boldsymbol{\beta}}^\top\Sigma^*\hat{\boldsymbol{\beta}}}}\Big).$$

The following theorem shows the convergence rate of $R(\hat{G}_{CHIME})$ to $R_{\text{opt}}(G_{\boldsymbol{\theta}^*})$, where $R_{\text{opt}}(G_{\boldsymbol{\theta}^*})$ is defined in (1.4).

THEOREM 3.2. *Under the conditions of Theorem 3.1, if* $T_0 \geq (-\log(\kappa))^{-1} \cdot \log(n \cdot d_{2,s}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\theta}^*))$, *the mis-clustering error* $R(\hat{G}_{CHIME})$ *for the classifier* $\hat{G}_{CHIME}(\boldsymbol{z})$ *defined in* (3.10) *satisfies*

$$R(\hat{G}_{CHIME}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) \lesssim \frac{s \log p}{n},$$

with probability at least $1 - o(1)$.

The result in Theorem 3.2 pushes forward the convergence rate of the mis-classification error of the LPD rule [9]. In fact, Theorem 3 in [9] implies that the convergence rate is $R(\hat{G}_{LPD}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) = O((s \log p/n)^{1/2})$, over the parameter space $\Theta_p(s, c_\omega, M_1, M_2)$. Theorem 3.2 shows a faster rate and later in Section 3.3 we will show that this convergence rate in the order of $(s \log p)/n$ is indeed optimal.

3.2. *Initialization.* As mentioned earlier, CHIME requires a good initialization $\hat{\boldsymbol{\theta}}^{(0)}$ that lies in the contraction basin $B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)$, defined in (3.3). This contraction basin forces the two inner products, $\boldsymbol{\delta}^\top \boldsymbol{\beta}^*$ and $(\boldsymbol{\delta}^*)^\top \boldsymbol{\beta}$ to be of the same order as $\Delta^2 = (\boldsymbol{\delta}^*)^\top \boldsymbol{\beta}^*$. In the special case where $\Sigma^* = \mathbf{I_p}$, this constraint reduces to the boundedness condition on the relative error of $\boldsymbol{\delta}$. The latter condition was used in [4, 36, 39], where they focused on the specialized mixture model $\frac{1}{2}N_p(-\boldsymbol{\mu}^*, \sigma^2\mathbf{I_p}) + \frac{1}{2}N_p(\boldsymbol{\mu}^*, \sigma^2\mathbf{I_p})$. From a theoretical perspective, this condition guarantees that the weights $\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)})$ assigned in the E-step are close to the truth.

In the following, we introduce the initialization condition **(IC)**, which ensures that $\hat{\boldsymbol{\theta}}^{(0)} \in B_{con}(\boldsymbol{\theta}^*; c_0, c_1, C_b, s)$.

(IC) For some permutation $\pi : \{1, 2\} \to \{1, 2\}$

$$\max_{k=1,2}\{\|\hat{\boldsymbol{\mu}}_k^{(0)} - \boldsymbol{\mu}_{\pi(k)}^*\|_\infty\} \lesssim \frac{1}{s}, \quad |\hat{\Sigma}^{(0)} - \Sigma^*|_\infty \lesssim \frac{1}{s}.$$

The estimator $\hat{\boldsymbol{\theta}}^{(0)}$ satisfying **(IC)** can be obtained by the Hardt-Price algorithm. The Hardt-Price algorithm was proposed by [18, see algorithm B], which first established tight bounds for learning the parameters of a mixture of two univariate Gaussians using a variant of the method of moments [28]. They then extended the univariate result to the multivariate Gaussian mixture model and obtained the following theorem.

PROPOSITION 3.1 ([18]). *Suppose we observe* n *i.i.d. samples* $\boldsymbol{z}^{(i)}$ *from model* (1.2). *Given* $\epsilon, \nu > 0$, *if* $n = \Omega(\frac{1}{\epsilon^6}\log(\frac{p}{\nu}\log(\frac{1}{\epsilon})))$, *then with probability at least* $1 - \nu$, *the Hardt-Price algorithm produces estimates* $\hat{\boldsymbol{\mu}}_1^{(0)}$, $\hat{\boldsymbol{\mu}}_2^{(0)}$ *and* $\hat{\Sigma}^{(0)}$ *such that for some permutation* $\pi : \{1, 2\} \to \{1, 2\}$,

$$\max\left\{\|\hat{\boldsymbol{\mu}}_1^{(0)} - \boldsymbol{\mu}_{\pi(1)}^*\|_\infty^2, \|\hat{\boldsymbol{\mu}}_2^{(0)} - \boldsymbol{\mu}_{\pi(2)}^*\|_\infty^2, |\hat{\Sigma}^{(0)} - \Sigma^*|_\infty\right\} \leq \epsilon\left(\frac{1}{4}\|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_\infty^2 + |\Sigma^*|_\infty\right).$$

Using Proposition 3.1, the following lemma shows that $\hat{\boldsymbol{\theta}}^{(0)}$ given by the Hardt-Price algorithm satisfies **(IC)**, and thus guarantees that the subsequent estimators $\hat{\boldsymbol{\theta}}^{(t)}$ in Algorithm 1 are contained in the contraction basin.

LEMMA 3.3. *Let $\hat{\boldsymbol{\theta}}^{(0)}$ be the estimator constructed by the Hardt-Price algorithm. Under the conditions of Theorem 3.1, if $s(\frac{\log p}{n})^{1/12} = o(1)$, then for sufficiently large n, with probability $1 - o(1)$, $\hat{\boldsymbol{\theta}}^{(0)}$ satisfies* **(IC)** *and thus* **(C1)** *holds.*

REMARK 4. The conditions in Lemma 3.3 implies that the sample size $n \gtrsim s^{12} \log p$. To the best of our knowledge, the rate $n \gtrsim s^{12} \log p$ is so far the best for general Gaussian mixture models (without assuming spherical covariance matrix) in the literature (see, e.g., [18]). The optimality for the required sample size is an interesting problem for future work.

3.3. *Lower bounds.* We now turn to the minimax lower bounds for the estimation of $\boldsymbol{\beta}^*$ and the mis-clustering error. Our results show that CHIME yields optimal results in the minimax sense, both for estimating the discriminating direction $\boldsymbol{\beta}^*$ and for clustering.

THEOREM 3.3. *Under the conditions of Theorem 3.1, let $\mathcal{C}$ be the set of all clustering rules based on n i.i.d. samples $\{\boldsymbol{z}^{(1)}, ..., \boldsymbol{z}^{(n)}\}$ from model (1.2) with the true parameter $\boldsymbol{\theta}^* \in \Theta_p(s, c_\omega, M_1, M_2)$, for some constants $c_\omega, M_1, M_2 > 0$. If $\log p = O(\log(p/s))$, then*

(1).

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\theta}^* \in \Theta_p(s, c_\omega, M, M_b)} \mathbb{E} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \gtrsim \sqrt{\frac{s \log p}{n}},$$

(2).

$$\inf_{\hat{G} \in \mathcal{C}} \sup_{\boldsymbol{\theta}^* \in \Theta_p(s, c_\omega, M, M_b)} \mathbb{E}[R(\hat{G}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*})] \gtrsim \frac{s \log p}{n}.$$

Theorems, 3.1, 3.2 and 3.3 together show that our proposed estimator of $\boldsymbol{\beta}^*$ and the clustering rule attain the optimal rates of convergence.

REMARK 5. Although a sparsity assumption on $\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$ seems to be more appealing in the Gaussian mixture model (1.2), Theorem 3.3 demonstrates that sparsity on $\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$ alone is not sufficient as the precision matrix Ω^* also plays an important role. Indeed, Theorem 3.3 shows that the difficulty of the problem depends on the sparsity of the product $\boldsymbol{\beta}^* = \Omega^*(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*)$. Therefore, a structural assumption directly on $\boldsymbol{\beta}^*$ is the most natural.

In the proof of Theorem 3.3, while the construction of the lower bound for the estimation of $\boldsymbol{\beta}^*$ is standard, that of the mis-clustering error is

not straightforward. This is partially due to the fact that the risk function $R(\hat{G}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*})$ does not satisfy the triangle inequality. A key step is to reduce the above loss to an alternative risk function.

Let $G_{\boldsymbol{\theta}}$ be the optimal Fisher's classification rule defined with the parameter $\boldsymbol{\theta}$. For some generic classification rule G, we rewrite the risk function $R(G) - R(G_{\boldsymbol{\theta}}) = \mathbb{P}_{\boldsymbol{\theta}}(G(Z) \neq Y) - \mathbb{P}_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}(Z) \neq Y)$ and define $L_{\boldsymbol{\theta}}(G)$ by

$$L_{\boldsymbol{\theta}}(G) = \min_{\pi \in \mathcal{P}_2} \mathbb{P}_{\boldsymbol{\theta}}(G(Z) \neq \pi(G_{\boldsymbol{\theta}}(Z))).$$

The following lemma enables us to reduce the loss $R(\hat{G}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*})$ to the risk function $L_{\boldsymbol{\theta}^*}(\hat{G})$.

LEMMA 3.4. *Let* $Z \sim \frac{1}{2}N_p(\boldsymbol{\mu}_1, \Sigma) + \frac{1}{2}N_p(\boldsymbol{\mu}_2, \Sigma)$ *with parameter* $\boldsymbol{\theta} = (1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$. *Suppose* $\boldsymbol{\theta}$ *satisfies* $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq c_L$ *for some* $c_L > 0$. *Then there exists some constant* $m > 0$, *such that if* $L_{\boldsymbol{\theta}}(G) \leq 1/m$ *for some classifier* G, *then*

$$\frac{1}{2m} L^2_{\boldsymbol{\theta}}(G) \leq \mathbb{P}_{\boldsymbol{\theta}}(G(Z) \neq Y) - \mathbb{P}_{\boldsymbol{\theta}}(G_{\boldsymbol{\theta}}(Z) \neq Y).$$

Lemma 3.4 shows the relationship between the risk function $R(\hat{G}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*})$ and a more 'standard' risk function $L_{\boldsymbol{\theta}^*}(\hat{G})$. With Lemma 3.4, Theorem 3.3 can be proved by providing a lower bound for $L_{\boldsymbol{\theta}^*}(\hat{G})$. The risk function $L_{\boldsymbol{\theta}^*}(\hat{G})$ has been studied in [2] for a specialized model $\frac{1}{2}N(\boldsymbol{\mu}_1, \sigma^2 \mathbf{I}_p) + \frac{1}{2}N(\boldsymbol{\mu}_2, \sigma^2 \mathbf{I}_p)$. Although no matching upper and lower bounds were provided, the following lemma in [2] is crucial to our analysis, which shows the triangle inequality property of the risk function $L_{\boldsymbol{\theta}^*}(\hat{G})$. For two probability density functions $\mathbb{P}_{\boldsymbol{\theta}_1}$ and $\mathbb{P}_{\boldsymbol{\theta}_2}$, denote their KL divergence by

$$\text{KL}(\mathbb{P}_{\boldsymbol{\theta}_1}, \mathbb{P}_{\boldsymbol{\theta}_2}) = \int \mathbb{P}_{\boldsymbol{\theta}_1}(z) \log \frac{\mathbb{P}_{\boldsymbol{\theta}_1}(z)}{\mathbb{P}_{\boldsymbol{\theta}_2}(z)} \, dz.$$

LEMMA 3.5 ([2]). *For any* $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta_p(s, c_\omega, M, M_b)$ *and any clustering* $\hat{G}$, *if* $L_{\boldsymbol{\theta}}(G_{\tilde{\boldsymbol{\theta}}}) + L_{\boldsymbol{\theta}}(\hat{G}) + \sqrt{\frac{\text{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})}{2}} \leq 1/2$, *then*

$$L_{\boldsymbol{\theta}}(G_{\tilde{\boldsymbol{\theta}}}) - L_{\boldsymbol{\theta}}(\hat{G}) - \sqrt{\frac{\text{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})}{2}} \leq L_{\tilde{\boldsymbol{\theta}}}(\hat{G}) \leq L_{\boldsymbol{\theta}}(G_{\tilde{\boldsymbol{\theta}}}) + L_{\boldsymbol{\theta}}(\hat{G}) + \sqrt{\frac{\text{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}})}{2}}.$$

After applying Lemmas 3.4 and 3.5, we then use Fano's inequality to complete the proof of Theorem 3.3. The details are shown in Section 8.

4. Low-dimensional Gaussian Mixtures. Although the focus of the present paper is on the high-dimensional setting, our analysis can be modified to establish the optimality of the clustering procedure for the low-dimensional Gaussian mixtures that is based on the classical EM algorithm.

In the general low-dimensional setting, we consider the model

$$\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, ..., \boldsymbol{z}^{(n)} \overset{i.i.d.}{\sim} (1-\omega^*)N_p(\boldsymbol{\mu}_1^*, \Sigma^*) + \omega^* N_p(\boldsymbol{\mu}_2^*, \Sigma^*), \tag{4.1}$$

without imposing any assumption on the sparsity of the discriminant direction. In such case, direct estimation of $\boldsymbol{\beta}^*$ is not needed. The clustering procedure under model (4.1), which uses the classical EM algorithm to estimate $\omega^*, \boldsymbol{\mu}_1^*$, $\boldsymbol{\mu}_2^*$ and Σ^*, is summarized in Algorithm 2. We call it CLOME for **C**lustering of **LO**w-dimensional Gaussian **M**ixtures with the **E**M.

Algorithm 2 Clustering of **LO**w-dimensional Gaussian **M**ixtures with the **E**M (CLOME)

1: **Inputs:** Initializations $\hat{\omega}^{(0)}, \hat{\boldsymbol{\mu}}_1^{(0)}, \hat{\boldsymbol{\mu}}_2^{(0)}$ and $\hat{\Sigma}^{(0)}$, maximum number of iterations T_0.
2: **for** $t = 0, 1, \ldots, T_0 - 1$ **do**
3: **E-Step:** Evaluate $Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)})$ with $\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{z}^{(i)})$ defined in (2.3).
4: **M-Step:** Update $\hat{\omega}^{(t+1)}, \hat{\boldsymbol{\mu}}_1^{(t+1)}, \hat{\boldsymbol{\mu}}_2^{(t+1)}$, and $\hat{\Sigma}^{(t+1)}$ via (2.5), (2.6), (2.7) and (2.8).
5: **end for**
6: Output $\hat{\omega}^{(T_0)}, \hat{\boldsymbol{\mu}}_1^{(T_0)}, \hat{\boldsymbol{\mu}}_2^{(T_0)}$ and $\hat{\Sigma}^{(T_0)}$.
7: Construct the clustering rule

$$\hat{G}_{EM}(\boldsymbol{z}) = \begin{cases} 1, & \{\boldsymbol{z} - \frac{\hat{\boldsymbol{\mu}}_1^{(T_0)} + \hat{\boldsymbol{\mu}}_2^{(T_0)}}{2}\}^\top (\hat{\Sigma}^{(T_0)})^{-1}(\hat{\boldsymbol{\mu}}_1^{(T_0)} - \hat{\boldsymbol{\mu}}_2^{(T_0)}) \geq \log(\frac{\hat{\omega}^{(T_0)}}{1-\hat{\omega}^{(T_0)}}), \\ 2, & \{\boldsymbol{z} - \frac{\hat{\boldsymbol{\mu}}_1^{(T_0)} + \hat{\boldsymbol{\mu}}_2^{(T_0)}}{2}\}^\top (\hat{\Sigma}^{(T_0)})^{-1}(\hat{\boldsymbol{\mu}}_1^{(T_0)} - \hat{\boldsymbol{\mu}}_2^{(T_0)}) < \log(\frac{\hat{\omega}^{(T_0)}}{1-\hat{\omega}^{(T_0)}}). \end{cases}$$

The technical tools developed for the proofs of Theorems 3.1, 3.2 and 3.3 can be used to establish the optimality of CLOME in Algorithm 2. We consider the theoretical performance of estimation and the CLOME clustering procedure over the parameter space $\Theta_p(c_\omega, M_1, M_2)$, defined by

$$\begin{aligned}\Theta_p(c_\omega, M_1, M_2) = \{\boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p\times p}, \Sigma \succ 0, \\ \omega \in (c_\omega, 1-c_\omega), \|\Sigma\|_2 \leq M_1, \|\boldsymbol{\mu}_k\|_2 \leq M_2, k = 1, 2\}.\end{aligned}$$

Similar to the high-dimensional setting, CLOME requires a good initialization. The initial value $\hat{\boldsymbol{\theta}}^{(0)}$ should lie in the contraction basin $\tilde{B}_{con}(\boldsymbol{\theta}^*; c_0)$,

$$\begin{aligned}\tilde{B}_{con}(\boldsymbol{\theta}^*; c_0) = \{\boldsymbol{\theta} = (\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \omega \in (c_0, 1-c_0), \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p\times p}, \\ \Sigma \succ 0, \|\Sigma - \Sigma^*\|_2 \leq \frac{1}{4}\phi_{\min}(\Sigma^*), \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_k^*\|_2 \leq \frac{1}{4\|\Sigma\|_2}\|\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*\|_2, k = 1, 2\}.\end{aligned}$$

Indeed, in the low-dimensional regime, the algorithm proposed by [17], which is based on the method of moments, was proved to satisfy the above condition [see Theorem 3.4 of 17].

We are ready to provide the upper bound results of CLOME under the low-dimensional Gaussian mixture model (4.1). For any $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta_p(c_\omega, M_1, M_2)$, define the ℓ_2 distance between $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ by

$$d_2(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = |\omega - \tilde{\omega}| + \|\boldsymbol{\mu}_1 - \tilde{\boldsymbol{\mu}}_1\|_2 + \|\boldsymbol{\mu}_2 - \tilde{\boldsymbol{\mu}}_2\|_2 + \|\Sigma - \tilde{\Sigma}\|_2.$$

THEOREM 4.1. *Consider the model (4.1) over the parameter space* $\Theta_p(c_\omega, M_1, M_2)$ *where* $p = o(n)$. *Suppose the initialization* $\hat{\boldsymbol{\theta}}^{(0)} \in \tilde{B}_{con}(\boldsymbol{\theta}^*; c_0)$ *and* $\Delta^2 > \log(16M_1/3+64M_2/3)$. *Then there exist constants* $\kappa \in (0,1), C_1, C_2 > 0$, *such that with probability at least* $1 - n^{-1}$, *the outputs* $\hat{\boldsymbol{\mu}}_1^{(T_0)}, \hat{\boldsymbol{\mu}}_2^{(T_0)}$ *and* $\hat{\Sigma}^{(T_0)}$ *of Algorithm 2 satisfy*

$$\|\hat{\boldsymbol{\mu}}_k^{(T_0)} - \boldsymbol{\mu}_k^*\|_2 \le \kappa^{T_0} d_2(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(0)}) + C_1\sqrt{\frac{p}{n}}, \quad k = 1, 2;$$
$$\|\hat{\Sigma}^{(T_0)} - \Sigma^*\|_2 \le \kappa^{T_0} d_2(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(0)}) + C_2\sqrt{\frac{p}{n}}.$$

In particular, if $T_0 \ge 2(-\log(\kappa))^{-1}\log(nd_2(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(0)})/p)$, *then there exists a constant* $C_3 > 0$, *such that*

$$d_2(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}^{(T_0)}) \le C_3\sqrt{\frac{p}{n}} \quad \text{and} \quad R(\hat{G}_{EM}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) \le C_3\frac{p}{n}.$$

REMARK 6. Theorem 4.1 provides upper bound results for the estimators given in Algorithm 2 under a general Gaussian mixture model in (4.1), and shows that CLOME is consistent if the initialization $\hat{\boldsymbol{\theta}}^{(0)}$ lies in the contraction basin $\tilde{B}_{con}(\boldsymbol{\theta}^*; c_0)$. Applying Theorem 4.1 to the special case $\frac{1}{2}N_p(-\boldsymbol{\mu}^*, \sigma^2\mathbf{I}_p) + \frac{1}{2}N_p(\boldsymbol{\mu}^*, \sigma^2\mathbf{I}_p)$ leads to the same result as that in [4].

We establish the optimality of Algorithm 2 for both the estimators and the clustering rule by providing the following lower bound results.

THEOREM 4.2. *Under the conditions of Theorem 4.1, we have*

$$\inf_{\hat{\boldsymbol{\mu}}_k} \sup_{\boldsymbol{\theta}^* \in \Theta_p(c_\omega, M_1, M_2)} \mathbb{E}\|\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k^*\|_2 \gtrsim \sqrt{\frac{p}{n}}, \quad k = 1, 2;$$
$$\inf_{\hat{\Sigma}} \sup_{\boldsymbol{\theta}^* \in \Theta_p(c_\omega, M_1, M_2)} \mathbb{E}\|\hat{\Sigma} - \Sigma^*\|_2 \gtrsim \sqrt{\frac{p}{n}},$$
$$\inf_{\hat{G} \in \mathcal{C}} \sup_{\boldsymbol{\theta}^* \in \Theta_p(c_\omega, M_1, M_2)} R(\hat{G}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) \gtrsim \frac{p}{n}.$$

Theorems 4.1 and 4.2 together characterize the optimality of CLOME. Note that in the low-dimensional case the estimators $\hat{\boldsymbol{\mu}}_k^{(T_0)}$ and $\hat{\Sigma}^{(T_0)}$ achieve the same convergence rate as the MLE obtained with known sample labels.

5. Simulations. The proposed CHIME procedure is easily implementable. In this section we conduct simulation studies to investigate the numerical properties of CHIME under various settings.

We compare the performance of CHIME with the k-means (KM), sparse k-means [SKM, 38], Influential Feature PCA [IF-PCA, 21], penalized model-based clustering with common covariance matrices [PCCM, 40], sparse clustering via HardtPrice [SHP, 3], the linear programming discriminant rule

[LPD, 9] and the oracle Fisher's rule obtained by plugging in the true parameters (Oracle) on a suite of three simulated examples. Three methods including SKM, PCCM and SHP were implemented in `R`, whereas the others were implemented in `MATLAB`. We refer readers to [12] for additional simulation scenarios–including unequal mixing proportion case and settings with discriminant vectors of different sparsity levels–and subsequent discussion.

In all simulations, the sample size is $n = 200$ while the number of variables p varies from $100, 200, 500$ to 800. The probability of being in either of the two classes is equal, i.e. $\omega^* = 0.5$. The discriminant vector $\boldsymbol{\beta}^* \propto (1, \ldots, 1, 0, \ldots, 0)^\top$ is sparse such that only the first $s = 10$ entries are nonzero. We consider the following three models for the inverse covariance matrix Ω^*.

Model 1 Erdős-Rényi random graph: Let $\tilde{\Omega} = (\tilde{\omega}_{ij})$ where $\tilde{\omega}_{ij} = u_{ij}\delta_{ij}$, $\delta_{ij} \sim \text{Ber}(1, 0.05)$ being the Bernoulli random variable with success probability 0.05 and $u_{ij} \sim \text{Unif}[0.5, 1] \cup [-1, -0.5]$. After symmetrizing $\tilde{\Omega}$, set $\Omega^* = \tilde{\Omega} + \{\max(-\phi_{\min}(\tilde{\Omega}), 0) + 0.05\}\mathbf{I}_p$ to ensure the positive definiteness. Finally, Ω^* is standardized to have unit diagonals.

Model 2 Block sparse model: $\Omega^* = (\mathbf{B} + \delta\mathbf{I}_p)/(1 + \delta)$ where $b_{ij} = b_{ji} = 0.5 * \text{Ber}(1, 0.3)$ for $1 \le i \le s, i < j \le p$; $b_{ij} = b_{ji} = 0.5$ for $s + 1 \le i < j \le p$; $b_{ii} = 1$ for $1 \le i \le p$. In other words, only the first s rows and columns of Ω^* are sparse, whereas the rest of the matrix is not sparse. Here $\delta = \max(-\phi_{\min}(\mathbf{B}), 0) + 0.05$. The matrix Ω^* is also standardized to have unit diagonals.

Model 3 AR(1) model: $\Omega^* = (\Omega^*_{ij})_{p\times p}$ with $\Omega^*_{ij} = 0.8^{|i-j|}$.

In both Model 1 and Model 2, the vector $\boldsymbol{\beta}^* = (1, \ldots, 1, 0, \ldots, 0)^\top$. To ensure sufficiently strong signals in Model 3, we increase the magnitude of nonzero entries in $\boldsymbol{\beta}^*$ such that $\boldsymbol{\beta}^* = 2.5 \cdot (1, \ldots, 1, 0, \ldots, 0)^\top$. Given the inverse covariance matrix Ω^*, the mean of class 1 is $\boldsymbol{\mu}^*_1 = \mathbf{0}$ and mean of class 2 is $\boldsymbol{\mu}^*_2 = \boldsymbol{\mu}^*_1 - (\Omega^*)^{-1}\boldsymbol{\beta}^*$.

To find initializations for use in CHIME, we first run the k-means algorithm to find the initial class labels, and calculate $\boldsymbol{\mu}^{(0)}_1$ and $\boldsymbol{\mu}^{(0)}_2$. The pooled sample covariance matrices $\hat{\Sigma}^{(0)}$ is used as the initial value for the covariance matrix. We recommend running CHIME with multiple random initial class labels to obtain the best possible clustering and estimation results. In the case of Model 3, class labels estimated from SKM are sometimes more accurate than those from the k-means algorithm, and are thus used as candidates for initializing the parameters needed in CHIME.

As with any other penalization-based methods, CHIME, SKM, SHP, PCCM and LPD all require selecting a tuning parameter. To this end, we generated independently training data and test data from the same distribution. For a given λ, the training data were first used to estimate the parameters, with mis-clustering error evaluated based on the test data. The optimal λ was selected as the one that minimizes the mis-clustering errors over the test

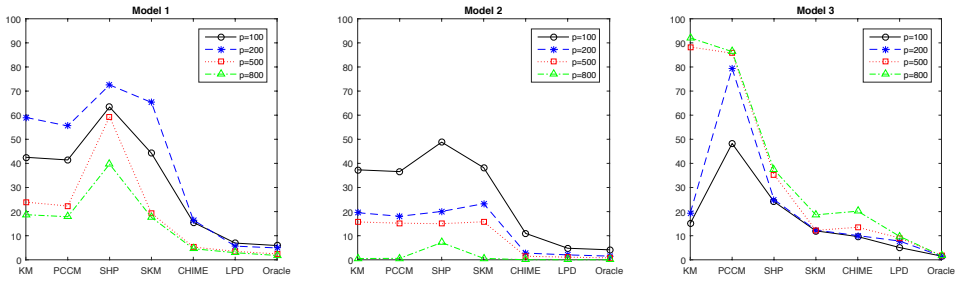

FIG 1. *Average mis-clustering errors based on* $n = 200$ *test samples from* 100 *replications under Model 1 (left), Model 2 (middle) and Model 3 (right). CHIME performs well in all three models.*

data. If there are multiple λ's that yield the same mis-clustering error, then the largest one will be selected. The tuning for SKM follows a slightly different procedure as the penalty parameter is specified in terms of an upper bound for a sequence of weights. The training data were first used to find the optimal upper bound, with mis-clustering error further evaluated on the test data under the optimal upper bound.

Figure 1 summarizes the average mis-clustering errors for different methods under the three aforementioned settings, with respective standard errors (s.e.) presented in Table 1. All comparisons were evaluated from 100 replications based on $n = 200$ test samples. Note the LPD rule is a supervised method for classification and is included as a benchmark comparison with the proposed method CHIME.

CHIME outperforms all other unsupervised clustering methods in both Models 1 and 2. Moreover, the mis-clustering errors from CHIME are comparable to those from LPD for $p = 500, 800$ in Model 1 and $p = 200, 500, 800$ in Model 2. In comparison, KM, SKM and PCCM yield rather similar performances, with IF-PCA showing the worst performances in all three models, since IF-PCA is designed for the case of "rare and weak" signal [21] and requires a diagonal covariance matrix. Clustering with SHP generally returns large mis-clustering errors and large standard errors, including in Model 3. This is due to its use of the moment-based estimator from the Hardt-Price algorithm for parameter initializations. The Hardt-Price algorithm requires a good pivot, i.e. one out of the p variables that shows the largest difference between the two cluster centers, to get a reasonable initialization. Such a pivot might be especially difficult to find in Model 1 as a majority of entries in $\boldsymbol{\mu}_2^*$ are randomly distributed around zero.

Clustering with Model 3 is more challenging due to the special structure of the inverse covariance matrix. Indeed, Ω^* in Model 3 is not sparse. Nonetheless CHIME maintains its good performance and achieves mis-clustering errors that are comparable to those from SKM and smaller than those from other clustering methods. On the other hand, since $\boldsymbol{\mu}_2^*$ is exactly sparse with

TABLE 1
Average mis-clustering errors (s.e.) based on $n = 200$ test samples from 100 replications under three different models

	p	100	200	500	800
Model 1	KM	42.53(6.81)	59.07(7.67)	23.94(6.53)	18.72(4.32)
	PCCM	41.43(5.63)	55.53(7.41)	22.31(5.60)	17.87(3.93)
	SHP	64.33(16.38)	72.34(13.91)	58.28(18.79)	51.33(16.78)
	SKM	44.20(6.02)	65.30(8.87)	19.29(6.42)	17.59(3.91)
	IF-PCA	92.73(5.87)	94.50(4.58)	94.98?4.59?	94.03(3.94)
	CHIME	16.21(6.21)	15.37(9.97)	5.21(3.03)	4.79(1.99)
	LPD	6.94(2.49)	5.67(2.22)	3.51(2.02)	2.94(1.58)
	Oracle	5.92(2.46)	4.92(2.13)	2.44(1.64)	1.79(1.24)
Model 2	KM	37.33(5.82)	19.54(4.33)	15.71(3.57)	0.54(0.72)
	PCCM	36.59(6.26)	18.05(4.23)	15.20(3.34)	0.60(0.75)
	SHP	51.54(20.14)	20.07(16.71)	14.98(9.84)	7.16(6.75)
	SKM	38.23(6.18)	23.28(4.89)	15.78(3.67)	0.60(0.72)
	IF-PCA	75.25(22.16)	82.65(15.99)	87.55(10.40)	91.15(8.94)
	CHIME	9.62(4.92)	3.35(2.18)	2.07(1.46)	0.03(0.21)
	LPD	4.80(2.42)	2.04(1.39)	1.09(0.96)	0.03(0.17)
	Oracle	4.14(2.20)	1.53(1.23)	0.81(0.86)	0.01(0.10)
Model 3	KM	15.08(4.49)	19.39(9.47)	47.68(22.73)	65.19(20.57)
	PCCM	48.52(36.81)	79.38(18.67)	85.72(3.47)	86.37(3.64)
	SHP	24.13(20.92)	24.96(19.90)	35.37(22.17)	37.34(24.64)
	SKM	12.00(3.17)	12.32(3.28)	12.21(3.28)	18.66(20.99)
	IF-PCA	88.17(11.19)	93.00(6.50)	92.98(7.22)	93.83(5.2456)
	CHIME	8.96(2.89)	9.75(2.87)	12.94(3.76)	19.97(20.14)
	LPD	5.08(2.41)	7.77(2.69)	8.98(2.85)	9.65(2.76)
	Oracle	1.53(1.34)	1.52(1.29)	1.73(1.24)	1.69(1.19)

$s + 1$ nonzero entries by construction, SKM shows significant improvement over KM, especially for large p, by taking advantage of sparsity in the true mean parameters. PCCM performs poorly in Model 3 and worse than KM for $p = 500, 800$, mainly because of its poor performance in estimating the non-sparse precision matrix. In the case of large p, it also suffers from poor initializations with the k-means algorithm.

6. Applications to Glioblastoma Gene Expression Data. To illustrate the proposed CHIME procedure, we consider in this section an application based on glioblastoma gene expression data. Glioblastoma (GBM) is the most common and aggressive form of brain cancer in adults. In order to provide the best treatments for patients with glioblastoma, an important question is classification of GBM subtypes, as different subtypes may respond to treatments differently. In a well-known paper, [35] introuced a robust gene expression-based molecular classification of GBM into Proneural, Neural, Classical and Mesenchymal subtypes. The data are available at https://tcga-data.nci.nih.gov/docs/publications/gbm_exp/. In this study, 200 GBM and two normal brain samples assayed across three gene expression platforms were first integrated into a single unified dataset. After

further filtering, there remain 1740 genes with consistent but highly variable expression across the platforms. The 202 samples were hierarchically clustered using the consensus average linkage. Based on the silhouette width, 173 of the 202 samples were selected as the "core samples" for being most representative of the clusters. Thus our following analysis was based solely on the core samples.

To validate the performance of CHIME in recovering the labels of a two-component Gaussian mixture, we focused on two of the four identified GBM subtypes: Mesenchymal and Neural, yielding a total of 82 samples among which 56 are from the Mesenchymal group. For the purpose of clustering, one can use the full set of 1740 genes, or select a subset of them. As the samples are pre-selected, direct application of any clustering methods on the full set yields almost perfect match between the estimated clusters and the true ones. We thus followed the latter approach and chose $p = 200$ genes from the full set of 1740 genes. In particular, we considered gene selection as follows. First we calculated the variances of all the genes and ranked them in a decreasing order. The top 20 genes with the largest variances and the last 180 genes with the smallest variances were then selected as the training set. We anticipate that the high variance genes are more informative than low variance genes, although this is not always true as the results below show.

Since we do not have a separate test data with labels, we propose to select the tuning parameter required in CHIME via a stability approach, motivated by [33]. The idea is to first randomly split the data into the training set and the test set. For a given λ, we run CHIME on the test data and obtain class labels of the test data, run CHIME on the training data, and finally measure how well the parameters estimated from the training data predict the class labels of the test data. Formally, let $f(X)$ be a clustering operation learned from data X and $G[f(\cdot), X]$ be the class labels estimated on X based on the clustering operation $f(\cdot)$. The prediction strength is then defined as the average adjusted random index when comparing $G[f(X_{train}), X_{test}]$ to $G[f(X_{test}), X_{test}]$ over B replications:

$$ps(\lambda) = \frac{1}{B}\sum_{i=1}^{B} \mathrm{ARI}(G[f(X^i_{train}), X^i_{test}], G[f(X^i_{test}), X^i_{test}]). \tag{6.1}$$

The optimal λ^* is selected as $\arg\max_\lambda ps(\lambda)$. Note the adjusted rand index is preferred over the rand index as the former has the advantage of being corrected-for-chance. This is especially important since if $\hat{\boldsymbol{\beta}} = 0$ due to the large penalty, $G[f(X_{train}), X_{test}]$ can randomly coincide with $G[f(X_{test}), X_{test}]$, resulting in a large value in rand index, but not in terms of the adjusted rand index. In addition, we define the prediction strength in terms of the adjusted rand index rather than the original one proposed in [33], as the former favors a larger penalty parameter and thus returns a sparser estimate that is more interpretable.

To apply CHIME, SHP and PCCM, we first selected the tuning parameters by maximizing the prediction strength defined in (6.1). The tuning parameter required in SKM was selected via criteria defined in [38]. Sparse clustering of the 200 genes with CHIME at the optimal λ yields 2 errors. A comparison with other clustering methods reveals that CHIME performs the best in recovering the correct sample labels, as shown in Table 2. Among all other methods, SHP yields the largest error, possibly due to incorrect parameter initializations with the Hardt-Price algorithm.

TABLE 2
Clustering results for the GBM gene expression data with $p = 200$ genes and 82 samples

	CHIME		KM		PCCM		SHP		SKM	
Class	1	2	1	2	1	2	1	2	1	2
Neural	26	0	26	0	26	0	12	14	25	1
Mesenchymal	2	54	7	49	5	51	10	46	6	50

To understand the performance of CHIME better, we also looked at the selected informative variables, i.e. genes with nonzero coefficients in $\hat{\boldsymbol{\beta}}$. Figure 2 shows that large marginal variances do not necessarily imply large coefficients in $|\hat{\boldsymbol{\beta}}|$. In fact, a significant number of the low variance genes (59 of 180) turn out to be informative for the clustering. This again confirms that direct estimation of the discriminant vector with CHIME yields a better characterization of the clustering boundary than estimating separately the cluster mean differences and (partial) correlations among variables.

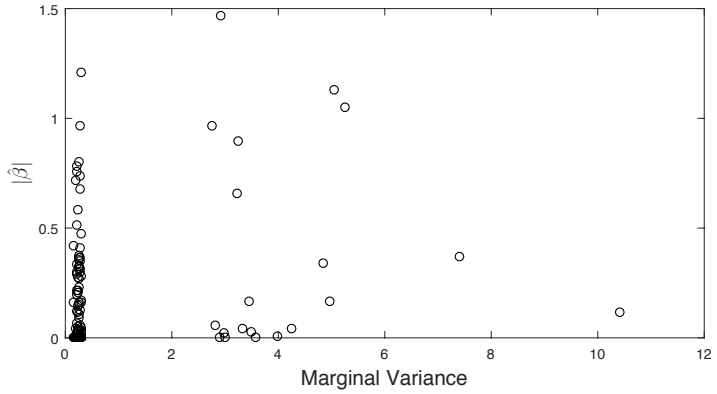

FIG 2. *The discriminant vector $|\hat{\boldsymbol{\beta}}|$ is plotted against the marginal variances.*

7. Extensions to Multi-class Gaussian Mixtures. The proposed method can be readily extended to Gaussian mixtures with K ($K \geq 2$) components. Consider the model

$$\mathbb{P}(Y = k) = \omega_k^*, \quad Z \mid Y = k \sim N_p(\boldsymbol{\mu}_k^*, \Sigma^*), \quad k = 1, \ldots, K.$$

Here $\sum_{k=1}^{K} \omega_k^* = 1$. We assume K is fixed and known. In the ideal case where the parameters are known, the oracle Bayes rule yields the label assignment

$$\hat{Y} = \arg\max_{k=1,\ldots,K} \left\{ \boldsymbol{\beta}_k^{*\top}(Z - (\boldsymbol{\mu}_k^* + \boldsymbol{\mu}_1^*)/2) + \log \omega_k^* \right\}, \tag{7.1}$$

where $\boldsymbol{\beta}_k^* = (\Sigma^*)^{-1}(\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_1^*)$ $(k = 1, 2, \ldots, K)$ are the discriminant directions. By definition, the vector $\boldsymbol{\beta}_1^*$ is trivial.

When neither the parameters nor the sample labels are known, under the assumption that the discriminant directions $\boldsymbol{\beta}_k^*$ $(k = 2, \ldots, K)$ are sparse, CHIME can be generalized for clustering multi-class Gaussian mixtures. Specifically, denote the posterior probability of the i-th sample in class k by

$$\hat{\gamma}_{ik}^{(t)} := \mathbb{P}(y_i = k | \boldsymbol{z}^{(i)}, \hat{\boldsymbol{\theta}}^{(t)}) = \frac{\hat{\omega}_k^{(t)} f(\boldsymbol{z}^{(i)} \mid \hat{\boldsymbol{\mu}}_k^{(t)}, \hat{\boldsymbol{\beta}}^{(t)})}{\sum_{\ell=1}^{K} \hat{\omega}_\ell^{(t)} f(\boldsymbol{z}^{(i)} \mid \hat{\boldsymbol{\mu}}_\ell^{(t)}, \hat{\boldsymbol{\beta}}^{(t)})}.$$

The conditional log-likelihood at the t-th step becomes

$$Q_n(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(t)}) = -\frac{1}{2n} \sum_{\substack{i\in[n] \\ k\in[K]}} \hat{\gamma}_{ik}^{(t)} (\boldsymbol{z}^{(i)} - \hat{\boldsymbol{\mu}}_k^{(t)})^T \Omega (\boldsymbol{z}^{(i)} - \hat{\boldsymbol{\mu}}_k^{(t)}) + \frac{1}{n} \sum_{\substack{i\in[n] \\ k\in[K]}} \hat{\gamma}_{ik}^{(t)} \log \hat{\omega}_k^{(t)} + \frac{1}{2} \log |\Omega|.$$

The updates of $\omega_k, \boldsymbol{\mu}_k$ and Σ in the M-step are respectively,

$$\hat{\omega}_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\gamma}_{ik}^{(t)}, \hat{\boldsymbol{\mu}}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{\gamma}_{ik}^{(t)} \boldsymbol{z}^{(i)}}{\sum_{i'=1}^{n} \hat{\gamma}_{i'k}^{(t)}},$$

$$\hat{\Sigma}^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\gamma}_{ik}^{(t)} (\boldsymbol{z}^{(i)} - \hat{\boldsymbol{\mu}}_k^{(t+1)})(\boldsymbol{z}^{(i)} - \hat{\boldsymbol{\mu}}_k^{(t+1)})^\top.$$

Finally, $\hat{\boldsymbol{\beta}}_k$'s are updated by solving the following optimizations:

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{2} \boldsymbol{\beta}^\top \hat{\Sigma}^{(t+1)} \boldsymbol{\beta} - \boldsymbol{\beta}^\top (\hat{\boldsymbol{\mu}}_k^{(t+1)} - \hat{\boldsymbol{\mu}}_1^{(t+1)}) + \lambda_n^{(t+1)} \|\boldsymbol{\beta}\|_1 \right\}, \ k = 2, \ldots, K.$$

This algorithm assumes sparsity of each discriminant direction $\boldsymbol{\beta}_k^*$ $(k = 2, \ldots, K)$, but no conditions on their joint support. If it is believed that the discriminant vectors have similar support, one might impose a group lasso penalty for their estimation, as done in [25].

The final clustering rule is constructed by plugging the estimates $\omega^*, \boldsymbol{\mu}_k^*$ $(k = 1, \ldots, K)$ and $\boldsymbol{\beta}_k^*$ $(k = 2, \ldots, K)$ into the optimal rule (7.1). Provided with a good initialization, similar techniques introduced in previous sections can be used to establish the convergence rate of $\hat{\boldsymbol{\beta}}_k$ as well as the upper and lower bounds of the mis-clustering error under suitable regularity conditions. The initialization for clustering multi-class Gaussian mixtures can be obtained by algorithms in Moitra and Valiant [26] or Ge et al. [17]. It was shown that the

estimate lies in B_{con} with probability at least $1-\delta$ when $n > \text{poly}(p, \frac{1}{\delta}, \frac{1}{\Delta})$, where poly$(\cdot)$ denotes the polynomial dependence. We also note here that the initialization step is of much importance in the multi-class setting, since it has been shown in Jin et al. [20] that the EM algorithm could stuck at a local optimum without a good initialization.

8. Proofs. In this section, we prove the optimality for the mis-clustering error, i.e. Theorem 3.2 and the part (2) of Theorem 3.3. The proof of the optimality for the estimation error, Theorem 3.1 and part (1) of Theorem 3.3, is given in the supplement [12]. A few technical lemmas are needed for the proof of the main results. These technical lemmas as well as some other minor results are proved in the supplement [12].

8.1. *Proof of Theorem 3.2.* We start with the following lemma.

LEMMA 8.1. *For two vectors $\boldsymbol{\gamma}^*$ and $\hat{\boldsymbol{\gamma}}$, if $\|\boldsymbol{\gamma}^* - \hat{\boldsymbol{\gamma}}\|_2 \le \|\boldsymbol{\gamma}^*\|_2$, and $\|\boldsymbol{\gamma}^*\|_2 \ge c$ for some constant $c > 0$, then*

$$(\boldsymbol{\gamma}^*)^\top \hat{\boldsymbol{\gamma}} - \|\boldsymbol{\gamma}^*\|_2 \cdot \|\hat{\boldsymbol{\gamma}}\|_2 \asymp \|\boldsymbol{\gamma}^* - \hat{\boldsymbol{\gamma}}\|_2^2.$$

Consider the model (1.2). Given the estimators $\hat{\omega}, \hat{\boldsymbol{\mu}}_k$, and $\hat{\boldsymbol{\beta}}$, the sample $\boldsymbol{z}$ is classified as

$$\hat{G}(\boldsymbol{z}) = \begin{cases} 1, & (\boldsymbol{z} - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2)^\top \hat{\boldsymbol{\beta}} \ge \log(\frac{\hat{\omega}}{1-\hat{\omega}}) \\ 2, & (\boldsymbol{z} - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2)^\top \hat{\boldsymbol{\beta}} < \log(\frac{\hat{\omega}}{1-\hat{\omega}}). \end{cases}$$

Let $\tau^* = \frac{\omega^*}{1-\omega^*}$, $\hat{\tau} = \frac{\hat{\omega}}{1-\hat{\omega}}$ and $\hat{\Delta} = \hat{\boldsymbol{\beta}}^\top \Sigma^* \hat{\boldsymbol{\beta}}$. The mis-clustering error is

$$R(\hat{G}) = (1-\omega^*)\Phi\Big(\frac{\log\hat{\tau} + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}}\Big) + \omega^* \bar{\Phi}\Big(\frac{\log\hat{\tau} + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2^*)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}}\Big),$$

with $R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) = (1-\omega^*)\Phi\Big(\frac{\log\tau^* - \Delta^2/2}{\Delta}\Big) + \omega^*\bar{\Phi}\Big(\frac{\log\tau^* + \Delta^2/2}{\Delta}\Big)$. Define an intermediate quantity

$$R^* = (1-\omega^*)\Phi\Big(\frac{\log\tau^* - (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\Big) + \omega^*\bar{\Phi}\Big(\frac{\log\tau^* + (\boldsymbol{\delta}^*)^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\Big).$$

We first show that $R^* - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) \lesssim \frac{s\log p}{n}$. Applying Taylor's expansion to the two terms in R^* at $\frac{\log\tau^*}{\Delta} - \frac{\Delta}{2}$ and $\frac{\log\tau^*}{\Delta} + \frac{\Delta}{2}$ respectively, we obtain

$$\begin{aligned} R^* - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) &= (1-\omega^*)\Big(\frac{\log\tau^*}{\hat{\Delta}} - \frac{(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\log\tau^*}{\Delta} + \frac{\Delta}{2}\Big)\Phi'\Big(\frac{\log\tau^*}{\Delta} - \frac{\Delta}{2}\Big) \\ &\quad - \omega^*\Big(\frac{\log\tau^*}{\hat{\Delta}} + \frac{(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\log\tau^*}{\Delta} - \frac{\Delta}{2}\Big)\Phi'\Big(\frac{\log\tau^*}{\Delta} + \frac{\Delta}{2}\Big) + O_P\Big(\frac{s\log p}{n}\Big), \end{aligned} \tag{8.1}$$

where the remaining term is bounded by using the facts that

$$\Big(\frac{\log\tau^*}{\hat\Delta}+\frac{(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{2\hat\Delta}-\frac{\log\tau^*}{\Delta}-\frac{\Delta}{2}\Big)^2=O_p\Big(\frac{s\log p}{n}\Big),\ \text{and}\ \Phi''=O(1).$$

In fact,

$$\begin{aligned}
&|\frac{\log\tau^*}{\hat\Delta}+\frac{(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{2\hat\Delta}-\frac{\log\tau^*}{\Delta}-\frac{\Delta}{2}|\le|\frac{\log\tau^*}{\hat\Delta}-\frac{\log\tau^*}{\Delta}|+|\frac{(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{2\hat\Delta}-\frac{\Delta}{2}|\\
\le&|\frac{\log\tau^*}{\hat\Delta}-\frac{\log\tau^*}{\Delta}|+|\frac{(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{2\hat\Delta}-\frac{\Delta^2}{2\hat\Delta}|+|\frac{\Delta^2}{2\hat\Delta}-\frac{\Delta}{2}|\lesssim|\hat\Delta-\Delta|+|(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}-(\boldsymbol{\delta}^*)^\top\boldsymbol{\beta}^*|\\
\lesssim&|\hat\Delta-\Delta|+\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|\lesssim\sqrt{\frac{s\log p}{n}}.
\end{aligned}\tag{8.2}$$

Recall that $\tau^*=\frac{\omega^*}{1-\omega^*}$, (8.1) can be further expanded such that

$$\begin{aligned}
\frac{R^*-R_{\text{opt}}(G_{\boldsymbol{\theta}^*})}{\sqrt{(1-\omega^*)\omega^*}}\asymp&\Big(\frac{\log\tau^*}{\hat\Delta}-\frac{(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{2\hat\Delta}-\frac{\log\tau^*}{\Delta}+\frac{\Delta}{2}\Big)e^{-\frac12\left(\frac{\log\tau^*}{\Delta}-\frac{\Delta}{2}\right)^2-\frac{\log\tau^*}{2}}\\
&-\Big(\frac{\log\tau^*}{\hat\Delta}+\frac{(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{2\hat\Delta}-\frac{\log\tau^*}{\Delta}-\frac{\Delta}{2}\Big)e^{-\frac12\left(\frac{\log\tau^*}{\Delta}+\frac{\Delta}{2}\right)^2+\frac{\log\tau^*}{2}}\\
=&\exp\Big(-\frac{\log^2\tau^*}{2\Delta^2}-\frac{\Delta^2}{8}\Big)\cdot\Big(\Delta-\frac{(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{\hat\Delta}\Big)\lesssim|\Delta-\frac{(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{\hat\Delta}|\\
\lesssim&\|\boldsymbol{\beta}^*-\hat{\boldsymbol{\beta}}\|_2^2.
\end{aligned}$$

In fact, for the last step, we can obtain this inequality by letting $\boldsymbol{\gamma}=(\Sigma^*)^{1/2}\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\gamma}}=(\Sigma^*)^{1/2}\hat{\boldsymbol{\beta}}$. Then

$$\Big|\Delta-\frac{(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{\hat\Delta}\Big|=\big|\|\boldsymbol{\gamma}\|_2-\frac{\boldsymbol{\gamma}^\top\hat{\boldsymbol{\gamma}}}{\|\hat{\boldsymbol{\gamma}}\|_2}\big|=|\frac{\|\boldsymbol{\gamma}\|_2\|\hat{\boldsymbol{\gamma}}\|_2-\boldsymbol{\gamma}^\top\hat{\boldsymbol{\gamma}}}{\|\hat{\boldsymbol{\gamma}}\|_2}|.$$

By Lemma 8.1, eventually we obtain $R^*-R_{\text{opt}}(G_{\boldsymbol{\theta}^*})\asymp\|\boldsymbol{\beta}^*-\hat{\boldsymbol{\beta}}\|_2^2$.

To upper bound $R(\hat G)-R^*$, applying Taylor's expansion to $R(\hat G)$,

$$\begin{aligned}
R(\hat G)=&(1-\omega^*)\Bigg\{\Phi\Big(\frac{\log\tau^*-(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2}{\hat\Delta}\Big)\\
&+\frac{\log\hat\tau-\log\tau^*+(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_1^*)^\top\hat{\boldsymbol{\beta}}+(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2}{\hat\Delta}\Phi'\Big(\frac{\log\tau^*-(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2}{\hat\Delta}\Big)+O_P(\frac{s\log p}{n})\Bigg\}\\
&+\omega^*\Bigg\{\bar\Phi\Big(\frac{\log\tau^*+(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2}{\hat\Delta}\Big)\\
&-\frac{\log\hat\tau-\log\tau^*+(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}_2^*)^\top\hat{\boldsymbol{\beta}}-(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2}{\hat\Delta}\Phi'\Big(\frac{\log\tau^*+(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2}{\hat\Delta}\Big)+O_P(\frac{s\log p}{n})\Bigg\},
\end{aligned}$$

where the remaining term $O_P(\frac{s\log p}{n})$ can be obtained similarly as (8.1).

This leads to

$$\begin{aligned}
&\frac{R^* - R(\hat{G})}{\sqrt{(1-\omega^*)\omega^*}} \\
\lesssim&\sqrt{\frac{1-\omega^*}{\omega^*}} \cdot \frac{\log\tau^* - \log\hat{\tau} - (\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1^*)^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}}\Phi'\Big(\frac{\log\tau^* - (\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\Big) \\
&- \sqrt{\frac{\omega^*}{1-\omega^*}} \cdot \frac{\log\tau^* - \log\hat{\tau} + (\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2^*)^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}}\Phi'\Big(\frac{\log\tau^* + (\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\Big) \\
=&\frac{\log\tau^* - \log\hat{\tau} - (\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1^*)^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}} e^{-\frac{1}{2}\left\{\frac{\log\tau^* - (\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right\}^2 - \frac{\log\tau^*}{2}} \\
&- \frac{\log\tau^* - \log\hat{\tau} + (\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2^*)^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}} e^{-\frac{1}{2}\left\{\frac{\log\tau^* + (\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right\}^2 + \frac{\log\tau^*}{2}}.
\end{aligned}$$

Then it follows that

$$\begin{aligned}
\frac{R^* - R(\hat{G})}{\sqrt{(1-\omega^*)\omega^*}} \lesssim& \Big|\frac{\log\tau^* - \log\hat{\tau} - (\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1^*)^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}}\Big| \\
&\cdot \Big|e^{-\frac{(\log\tau^* - (\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2)^2}{2\hat{\Delta}^2} - \frac{\log\tau^*}{2}} - e^{-\frac{(\log\tau^* + (\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2)^2}{2\hat{\Delta}^2} + \frac{\log\tau^*}{2}}\Big| \\
=& \underbrace{\Big|\frac{\log\tau^* - \log\hat{\tau} - (\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1^*)^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}}\Big|}_{(i)} \cdot \underbrace{e^{-\frac{\log^2\tau^* + (\boldsymbol{\delta}^{*\top}\hat{\boldsymbol{\beta}}/2)^2}{2\hat{\Delta}^2}}}_{(ii)} \\
&\cdot \underbrace{\Big|e^{\frac{\log\tau^*\cdot(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{2\hat{\Delta}^2} - \frac{\log\tau^*}{2}} - e^{-\frac{\log\tau^*\cdot(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{2\hat{\Delta}^2} + \frac{\log\tau^*}{2}}\Big|}_{(iii)} \\
\lesssim&\sqrt{\frac{s\log p}{n}} \cdot O_p(1) \cdot \sqrt{\frac{s\log p}{n}} \lesssim \frac{s\log p}{n},
\end{aligned}$$

where the last inequality uses the following facts

$$(i) \lesssim \sqrt{\frac{s\log p}{n}}, \quad (ii) = O(1), \text{ and } (iii) \lesssim \sqrt{\frac{s\log p}{n}}.$$

In fact, the bound on (i) follows the same idea of (8.2). (ii) uses the fact that $e^{-x} \le 1$ when $x \ge 0$. (iii) uses the fact that $|e^x - e^{-x}| \lesssim x$ when $x = o(1)$, and thus can be bounded as

$$|e^{\frac{\log\tau^*\cdot(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{2\hat{\Delta}^2} - \frac{\log\tau^*}{2}} - e^{-\frac{\log\tau^*\cdot(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{2\hat{\Delta}^2} + \frac{\log\tau^*}{2}}| \lesssim |\frac{(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}^2} - 1| \lesssim \sqrt{\frac{s\log p}{n}},$$

where the last inequality also follows the same idea as (8.2). Combining the pieces, we obtain

$$R(\hat{G}) - R_{\text{opt}}(G_{\boldsymbol{\theta}^*}) \lesssim \frac{s\log p}{n}. \quad \square$$

8.2. *Proof of Theorem 3.3.* We focus on mis-classification error. Consider the model $\frac{1}{2}N_p(\boldsymbol{\mu}_1, \Sigma)+\frac{1}{2}N_p(\boldsymbol{\mu}_2, \Sigma)$ with $\boldsymbol{\theta} = (1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) \in \Theta_p(s, c_\omega, M, M_b)$. Let $G_{\boldsymbol{\theta}}$ be the Fisher's rule defined in (1.3) with parameter $\boldsymbol{\theta}$, and the risk function for a generic parameter $\boldsymbol{\theta}$ and classification rule G is defined as

$$L_{\boldsymbol{\theta}}(G) = \mathbb{P}_{\boldsymbol{\theta}}(G \neq G_{\boldsymbol{\theta}}). \tag{8.3}$$

The proof of lower bound requires the generalized version of Fano's lemma.

LEMMA 8.2 ([34]). *Let $M \geq 0$ and $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M \in \Theta_p(s, c_\omega, M, M_b)$. For some constants $\alpha \in (0, 1/8), \gamma > 0$, and any classifier $\hat{G}$, if $\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_i}, \mathbb{P}_{\boldsymbol{\theta}_0}) \leq \alpha \log M/n$ for all $1 \leq i \leq M$, and $L_{\boldsymbol{\theta}_i}(\hat{G}) < \gamma$ implies $L_{\boldsymbol{\theta}_j}(\hat{G}) \geq \gamma$ for all $0 \leq i \neq j \leq M$, then*

$$\inf_{\hat{G}} \sup_{i \in [M]} \mathbb{E}_{\boldsymbol{\theta}_i}[L_{\boldsymbol{\theta}_i}(\hat{G})] \gtrsim \gamma.$$

LEMMA 8.3 ([34]). *Let $\mathcal{A}_s = \{\boldsymbol{u} : \boldsymbol{u} \in \{0, 1\}^p, \|\boldsymbol{u}\|_0 \leq s\}$. If $p \geq 4s$, then there exists $\{\boldsymbol{u}_0, \boldsymbol{u}_1, ..., \boldsymbol{u}_M\} \subset \mathcal{A}_s$ such that $\boldsymbol{u}_0 = \{0, ..., 0\}^\top$, $\rho_H(\boldsymbol{u}_i, \boldsymbol{u}_j) \geq s/2$ and $\log(M + 1) \geq \frac{s}{5} \log(\frac{p}{s})$, where ρ_H is the Hamming distance.*

LEMMA 8.4. *For any $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta_p(s, c_\omega, M, M_b)$, let $\mathbb{P}_{\boldsymbol{\theta}} = (1-\omega)N_p(-\boldsymbol{\mu}/2, \mathbf{I}_p)+\omega N_p(\boldsymbol{\mu}/2, \mathbf{I}_p)$ and $\mathbb{P}_{\tilde{\boldsymbol{\theta}}} = (1-\omega)N_p(-\tilde{\boldsymbol{\mu}}/2, \mathbf{I}_p) + \omega N_p(\tilde{\boldsymbol{\mu}}/2, \mathbf{I}_p)$ with $\|\boldsymbol{\mu}\|_2 = \|\tilde{\boldsymbol{\mu}}\|_2$. Then $\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) \leq (\|\boldsymbol{\mu}\|_2^2 + \log \tau/2)(\|\boldsymbol{\mu}\|_2^2 - |\boldsymbol{\mu}^\top \tilde{\boldsymbol{\mu}}|)$, where $\tau = \frac{\omega}{1-\omega}$. In particular, if $\omega = 1/2$, we have*

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\tilde{\boldsymbol{\theta}}}) \leq \|\boldsymbol{\mu}\|_2^4 \cdot \Big(1 - \frac{|\boldsymbol{\mu}^\top \tilde{\boldsymbol{\mu}}|}{\|\boldsymbol{\mu}\|_2}\Big).$$

Define the function $g(x) = \phi(x)\{\phi(x)-x\Phi(-x)\}$, where $\phi(x)$ is the probability density function of the standard normal distribution, i.e. $\phi(x) = \Phi'(x)$.

LEMMA 8.5 ([2]). *For any $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta_p(s, c_\omega, M, M_b)$ and $\cos\psi = |\boldsymbol{\mu}^\top \tilde{\boldsymbol{\mu}}|/\|\boldsymbol{\mu}\|_2$, we have*

$$2g\Big(\frac{\|\boldsymbol{\mu}\|}{2\sigma}\Big) \sin\psi \cos\psi \leq L_{\boldsymbol{\theta}}(G_{\tilde{\boldsymbol{\theta}}}).$$

PROOF OF THEOREM 3.3. First we construct a subset of the parameter space Θ that characterizes the hardness of the problem. Let $\boldsymbol{e}_1 = \{1, 0, ..., 0\}^\top \in \mathbb{R}^p$. By Lemma 8.3, there exist $\boldsymbol{u}_1, ..., \boldsymbol{u}_M \in \tilde{\mathcal{A}}_s = \{\boldsymbol{u} \in \{0, 1\}^p :, \boldsymbol{u}^\top \boldsymbol{e}_1 = 0, \|\boldsymbol{u}\|_0 = s\}$, such that $\rho_H(\boldsymbol{u}_i, \boldsymbol{u}_j) > s/2$ and $\log(M+1) \geq \frac{s}{5} \log(\frac{p-1}{s})$. Note the first entry in $\boldsymbol{u}_j$ is 0 for all $j = 1, \ldots, M$.

Define the parameter space

$$\Theta_1 = \{\boldsymbol{\theta} = (1/2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1 = \epsilon \boldsymbol{u} + \lambda \boldsymbol{e}_1, \boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1, \Sigma = \sigma^2 \mathbf{I}_p; \boldsymbol{u} \in \tilde{\mathcal{A}}_s\}.$$

Here $\epsilon = \sigma\sqrt{\log p/n}$, $\sigma^2 = O(1)$ and $\lambda = O(1)$ are chosen to ensure $\boldsymbol{\theta} \in \Theta_p(s, c_\omega, M, M_b)$ and $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = \frac{4\|\epsilon \boldsymbol{u}+\lambda \boldsymbol{e}_1\|_2^2}{\sigma^2} \geq c_1$, as required

in Lemma 3.4.To apply Lemma 8.2, we need to verify two conditions: (i) the upper bound on the KL divergence between $\mathbb{P}_{\boldsymbol{\theta}_u}$ and $\mathbb{P}_{\boldsymbol{\theta}_v}$, and (ii) the lower bound of $L_{\boldsymbol{\theta}_u}(\hat{G}) + L_{\boldsymbol{\theta}_v}(\hat{G})$ for $\boldsymbol{u} \neq \boldsymbol{v}$.

We calculate the KL divergence first. For $\boldsymbol{u} \in \tilde{\mathcal{A}}_s$, denote $\boldsymbol{\mu}_u = \epsilon \boldsymbol{u} + \lambda \boldsymbol{e}_1$. For $\boldsymbol{\theta}_u = (1/2, \boldsymbol{\mu}_u, -\boldsymbol{\mu}_u, \sigma^2 \mathbf{I}_p) \in \Theta_1$, the model parameterized by $\boldsymbol{\mu}_u$ is $\frac{1}{2}N_p(\boldsymbol{\mu}_u, \sigma^2\mathbf{I}_p) + \frac{1}{2}N_p(-\boldsymbol{\mu}_u, \sigma^2\mathbf{I}_p)$. For $\boldsymbol{u}, \boldsymbol{v} \in \tilde{\mathcal{A}}_s$, since

$$\epsilon^2 \cdot \rho_H(\boldsymbol{u}, \boldsymbol{v}) = \langle \boldsymbol{\mu}_u - \boldsymbol{\mu}_v, \boldsymbol{\mu}_u - \boldsymbol{\mu}_v \rangle = \|\boldsymbol{\mu}_u\|_2^2 + \|\boldsymbol{\mu}_v\|_2^2 - 2\boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v = 2\|\boldsymbol{\mu}_u\|_2^2 - 2\boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v,$$

we have $\|\boldsymbol{\mu}_u\|_2^2 - \boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v = \frac{1}{2}\epsilon^2 \cdot \rho_H(\boldsymbol{u}, \boldsymbol{v}) \asymp \frac{s \log p}{n}$. Lemma 8.4 then yields

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_u}, \mathbb{P}_{\boldsymbol{\theta}_v}) \leq \|\boldsymbol{\mu}_u\|_2^2 (\|\boldsymbol{\mu}_u\|_2^2 - \boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v) \lesssim \frac{s \log p}{n}. \tag{8.4}$$

Consider $L_{\boldsymbol{\theta}}(G)$ defined in (8.3). Recall that in Lemma 8.5, $\cos\psi = \boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v / \|\boldsymbol{\mu}_u\|_2^2$. For the choice of ϵ and $\boldsymbol{\mu}_u$, we have $\frac{\|\boldsymbol{\mu}_u\|_2}{2\sigma} = O(1)$, which implies that $2g(\frac{\|\boldsymbol{\mu}_u\|_2}{2\sigma}) = O(1)$ under the condition $s = o(n/\log p)$. Also,

$$1 - \cos\psi = 1 - \frac{\boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v}{\|\boldsymbol{\mu}_u\|_2^2} = \frac{\|\boldsymbol{\mu}_u\|_2^2 - \boldsymbol{\mu}_u^\top \boldsymbol{\mu}_v}{\|\boldsymbol{\mu}_u\|_2^2} = \frac{\rho_H(\boldsymbol{u}, \boldsymbol{v})\epsilon^2}{2(\lambda^2 + s\epsilon^2)} \asymp \frac{s \log p}{n}.$$

Therefore, by Lemma 8.5,

$$L_{\boldsymbol{\theta}_u}(G_{\boldsymbol{\theta}_v}) \geq 2g\Big(\frac{\|\boldsymbol{\mu}\|_2}{2\sigma}\Big) \sin\psi \cos\psi \geq g\Big(\frac{\|\boldsymbol{\mu}\|_2}{2\sigma}\Big) \sqrt{1 + \cos\psi} \sqrt{1 - \cos\psi} \geq \sqrt{\frac{s \log p}{n}}.$$

Applying Lemma 3.5 with a proper choice of ϵ, we have, for any $\boldsymbol{u}, \boldsymbol{v} \in \tilde{\mathcal{A}}_s$,

$$L_{\boldsymbol{\theta}_u}(\hat{G}) + L_{\boldsymbol{\theta}_v}(\hat{G}) \geq L_{\boldsymbol{\theta}_u}(G_{\boldsymbol{\theta}_v}) - \sqrt{\frac{\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_u}, \mathbb{P}_{\boldsymbol{\theta}_v})}{2}} \gtrsim \sqrt{\frac{s \log p}{n}}.$$

So far we have verified the aforementioned conditions (i) and (ii). Lemma 8.2 immediately implies that

$$\inf_{\hat{G} \in \mathcal{C}} \sup_{\boldsymbol{\theta} \in \Theta_p(s, c_\omega, M, M_b)} L_{\boldsymbol{\theta}}(\hat{G}) \gtrsim \sqrt{\frac{s \log p}{n}}. \tag{8.5}$$

Finally combining (8.5) with Lemma 3.4, we obtain the desired lower bound for the mis-clustering error. □

Acknowledgments. We thank the Editor, the Associate Editor, and two referees for their detailed and constructive comments which have helped to improve the presentation of the paper.

References.

[1] Theodore Wilbur Anderson. *An Introduction To Multivariate Statistical Analysis*. Wiley-Intersceince, 3rd ed, New York, 2003.

[2] Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In *NIPS*, pages 2139–2147, 2013.

[3] Martin Azizyan, Aarti Singh, and Larry A Wasserman. Efficient sparse clustering of high-dimensional non-spherical Gaussian mixtures. In *AISTATS*, 2015.

[4] Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Stat.*, 45(1): 77–120, 2017.

[5] Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *Ann. Stat.*, 36(6):2577–2604, 2008.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

[7] Charles Bouveyron and Camille Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Comput. Stat. Data Anal.*, 71:52–78, 2014.

[8] Paul S Bradley, Usama M Fayyad, and Olvi L Mangasarian. Mathematical programming for data mining: Formulations and challenges. *INFORMS Journal on Computing*, 11(3):217–238, 1999.

[9] T Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *J. Am. Stat. Assoc.*, 106(496):1566–1577, 2011.

[10] T Tony Cai and Linjun Zhang. High-dimensional Gaussian copula regression: Adaptive estimation and statistical inference. *Statistica Sinica*, to appear, 2017.

[11] T Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.*, 106(494):594–607, 2011.

[12] T Tony Cai, Jing Ma, and Linjun Zhang. Supplement to "CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality". 2016.

[13] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. B (Statistical Methodology)*, 39(1):1–38, 1977.

[14] Richard O Duda and Peter E Hart. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

[15] Brian S Everitt. *Finite mixture distributions*. Wiley Online Library, 1981.

[16] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, 97(458):611–631, 2002.

[17] Rong Ge, Qingqing Huang, and Sham M Kakade. Learning mixtures of Gaussians in high dimensions. pages 761–770, 2015.

[18] Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 753–760. ACM, 2015.

[19] T Hastie, R Tibshirani, and J Friedman. *The elements of statistical learning 2nd edition*. New York: Springer, 2009.

[20] Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright, and Michael I Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *NIPS*, pages 4116–4124, 2016.

[21] Jiashun Jin, Wanjie Wang, et al. Influential features pca for high dimensional clustering. *Ann. Stat.*, 44(6):2323–2359, 2016.

[22] Jiashun Jin, Zheng Tracy Ke, and Wanjie Wang. Phase transitions for high dimensional clustering and related problems. *Ann. Stat.*, to appear, 2017.

[23] Bruce G Lindsay. Mixture models: Theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 5, 1995.

[24] Qing Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis

in ultra-high dimensions. *Biometrika*, 99(1):29–42, 2012.
[25] Qing Mai, Yi Yang, and Hui Zou. Multiclass sparse discriminant analysis. *Statistica Sinica*, to appear, 2017.
[26] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *FOCS, 2010 51st Annual IEEE Symposium on Theory of Computing*, pages 93–102. IEEE, 2010.
[27] Matey Neykov, Yang Ning, Jun S Liu, and Han Liu. A unified theory of confidence regions and testing for high dimensional estimating equations. *arXiv preprint arXiv:1510.08986*, 2015.
[28] Karl Pearson. Contributions to the mathematical theory of evolution. *Phil. Trans. R. Soc. A*, 185:71–110, 1894.
[29] Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239, 1984.
[30] Douglas Reynolds. Gaussian mixture models. *Encyclopedia of Biometrics*, pages 827–832, 2015.
[31] Allen J Scott and Michael J Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27(2):387–397, 1971.
[32] Lu Tian and Quanquan Gu. Communication-efficient distributed sparse linear discriminant analysis. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pages 1178–1187. PMLR, 20–22 Apr 2017.
[33] Robert Tibshirani and Guenther Walther. Cluster validation by prediction strength. *J. Comput. Graph. Stat.*, 14(3):511–528, 2005.
[34] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.
[35] Roel GW Verhaak, Katherine A Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D Wilkerson, C Ryan Miller, Li Ding, Todd Golub, Jill P Mesirov, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell*, 17(1): 98–110, 2010.
[36] Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. In *NIPS*, pages 2521–2529, 2015.
[37] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58(301):236–244, 1963.
[38] Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *J. Am. Stat. Assoc.*, 105(490):713–726, 2010.
[39] Xinyang Yi and Constantine Caramanis. Regularized EM algorithms: A unified framework and statistical guarantees. In *NIPS*, pages 1567–1575, 2015.
[40] Hui Zhou, Wei Pan, and Xiaotong Shen. Penalized model-based clustering with unconstrained covariance matrices. *Electron. J. Stat.*, 3:1473–1496, 2009.

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: tcai@wharton.upenn.edu
jinma@wharton.upenn.edu
linjunz@wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/∼tcai/