Efficient Sparse Clustering of High-Dimensional Non-spherical Gaussian Mixtures

Martin Azizyan
Machine Learning Department
Carnegie Mellon University
mazizyan@cs.cmu.edu

Aarti Singh
Machine Learning Department
Carnegie Mellon University
aarti@cs.cmu.edu

Larry Wasserman
Department of Statistics
Carnegie Mellon University
larry@stat.cmu.edu

Abstract

We consider the problem of clustering data points in high dimensions, i.e., when the number of data points may be much smaller than the number of dimensions. Specifically, we consider a Gaussian mixture model (GMM) with two non-spherical Gaussian components, where the clusters are distinguished by only a few relevant dimensions. The method we propose is a combination of a recent approach for learning parameters of a Gaussian mixture model and sparse linear discriminant analysis (LDA). In addition to cluster assignments, the method returns an estimate of the set of features relevant for clustering. Our results indicate that the sample complexity of clustering depends on the sparsity of the relevant feature set, while only scaling logarithmically with the ambient dimension. Further, we require much milder assumptions than existing work on clustering in high dimensions. In particular, we do not require spherical clusters nor necessitate mean separation along relevant dimensions.

1 Introduction

The last few years have seen extensive research on developing computationally efficient and statistically sound methods that can leverage sparsity of the relevant feature set for supervised learning (classification, regression, etc.) of high-dimensional data. These methods show that learning and selection of relevant features is possible even when the number of training data n is much less than the number of features d, which is typically the case for high-dimensional data. However, similar results for the unsupervised task of clustering are largely non-existent. The task of clustering high-dimensional data and extracting relevant features arises routinely in many applications, e.g., clustering of patients based on gene expression profiles and identifying the relevant genotypes, grouping web content and identifying relevant characteristics, clustering proteins with similar drug expression profiles, etc.

While there have been recent attempts at clustering high-dimensional data and selecting relevant features, these either do not come with theoretical guarantees or assume very strong conditions that suggest that even employing marginal feature selection, using projections of the data onto individual coordinates, as a pre-processing step before clustering might suffice. Thus, while supervised learning in high dimensions requires single-step methods that can perform the learning task and select relevant features simultaneously, it is not clear whether a sophisticated single-step approach is necessary for clustering in high dimensions.

A simple example which demonstrates that preprocessing the data using a marginal (coordinate-wise) feature selection step does not suffice for clustering, is provided by a mixture of two non-spherical Gaussian components (see Figure 1). It is clear that x_1 is relevant to define the clusters, however the marginal distribution of the data when projected onto x_1 is a single unimodal Gaussian. Hence, marginal feature selection cannot be used to identify the relevant features.

Motivated by this example, we consider a simple non-spherical Gaussian mixture model (defined formally in the next section) for clustering high-dimensional data, and aim to provide a computationally efficient algorithm for simultaneous feature selection and clustering, that comes with sample complexity guarantees that depend primarily on the number of relevant features (intrinsic dimension) and only logarithmically on the total number of features (ambient dimension).

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

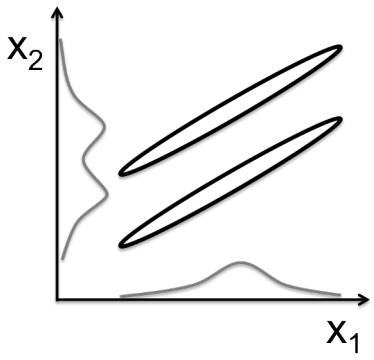


Figure 1: An example of two clusters where both features x_1 and x_2 are relevant to define the clusters, however the marginal distribution of data points along x_1 is unimodal, and hence no marginal feature selection method can work.

Related work. Before we describe our approach and results, we discuss related work in some more detail. Sparse clustering methods that perform feature selection for high-dimensional data have received attention recently.

K–means based approaches begin with the typical K–means objective and introduce some sparsity-inducing penalties [1, 2, 3, 4, 5]. While the penalization introduced in these papers is convex (akin to supervised learning approaches), the K–means objective itself is non-convex and in fact NP-hard. Thus, in general, solving any of these objectives is NP-hard and the papers propose iterative approaches akin to Llyod algorithm for solving the K–means objective. Moreover, these papers do not provide any statistical guarantees, with the exception of [1, 5]. The latter two papers do provide some consistency results, however these are for the true objective optimizers only which are NP-hard. Moreover, the notion of relevant features considered in all these papers is that the means are separated along each relevant feature, which may not necessarily be the case as demonstrated in Figure 1.

Another non-parametric approach to feature selection for clustering that is consistent in high dimensions is presented in [6], however it relies on pre-screening features which appear marginally unimodal, again failing for the example in Figure 1.

Learning Gaussian mixture models (GMMs) has a long history, particularly in computer science theory community, where the emphasis has been on relaxing the assumptions under which GMMs can be learnt under various metrics such as estimating the distribution, parameters or clustering [7, 8, 9]. However, these papers primarily focus on computational tractability and mostly have high sample complexity, particularly in high dimensions. For example, the most relevant to this paper is the work on learning non-spherical GMMs where the components are separated by a hyperplane [10], however it has sample complexity that depends as d^4 on the ambient dimension. The proposed estimator relies on first making the data isotropic (zero mean and overall identity covariance). This is achieved by pre-whitening the data by multiplying it with the inverse sample covariance matrix. However, in high dimensions when the number of samples drawn from the mixture $n \ll d$ (the number of features), the sample covariance matrix is not invertible and hence the method cannot succeed. Moreover, no work in this line, to the best of our knowledge, addresses feature selection. There is a very recent work [11] where the question of optimal sample complexity for GMM parameter estimation in ℓ_∞ norm is addressed and we build on this paper to provide clustering and feature selection guarantees.

Apart from the work in the computer science theory community, multiple statistical approaches have also been proposed to learning Gaussian mixture models in high dimensions and feature selection [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. These employ various sparsity assumptions, e.g. that the components are spherical and have sparse mean vectors, or that the covariance matrices (or their inverses) are also sparse, etc. However, as the K–means based methods, these approaches either a) require approximating maximum likelihood parameters without providing efficient algorithms; or b) do not come with precise finite sample statistical properties of the estimators.

Assuming mixtures of equal weight spherical components with sparse mean separation, [24] provide some minimax bounds for the problem with sample complexity that scales with the number of relevant features and only logarithmically with the total number of features. Similar statistical guarantees are obtained by [25] for learning mixtures of more than two spherical components with sparse mean vectors. However, the assumption of spherical components necessitates that relevant features are characterized by mean separation, and hence the results do not apply for cases like the one described in Figure 1.

Under less restrictive assumptions on the components, [26] analyze detection of high-dimensional Gaussian mixtures (vs. a single Gaussian as null) and selection of sparse set of features along which mean separation occurs, from a minimax perspective. Their minimax optimal estimators involve combinatorial search, and while the authors also investigate some tractable procedures, they are either based on marginal feature selection or assume that the component covariance matrices are known and diagonal.

Finally, we mention that if the cluster assignments are known, the problem of clustering reduces to binary classification. And specifically, clustering using a mixture of two identical covariance Gaussians reduces to linear discriminant analysis, if the cluster assignments are known. Feature selection using a sparse linear discriminant analysis has been analyzed in [27]. We will leverage this approach, in combination with the results for ℓ_∞ parameter estimation for GMMs [11] to demonstrate a method and results for sparse clustering and feature selection under high-dimensional non-spherical GMMs.

Contributions. Our contributions can be summarized as follows:

- We present a computationally efficient method for clustering in high dimensions that comes with finite sample guarantees on the misclustering rate. Our results show that sample complexity scales quadratically with the number of relevant features (inherent dimension), but only logarithmically with total number of features (ambient dimension). As a result, the proposed method enables learning of non-spherical Gaussian mixtures of two components in high dimensions (when the number of data points may be much smaller than the number of features), without assuming sparsity of the covariance or inverse covariance matrix.
- We provide guarantees for feature selection under a very generalized notion of relevant features that does not require that clusters necessarily have mean separation along the relevant features. This allows us to handle cases like that shown in Figure 1.

The rest of the paper is organized as follows. In section 2, we formalize our setup. The proposed method combining ideas from [11] and [27] is presented in section 3. Section 4 states our results on misclustering rate, sample complexity and feature selection in high dimensions. Experimental results described in section 5 on some simulated datasets demonstrate the viability of our proposed method. We conclude with some open directions in section 6.

2 Problem Setup and Assumptions

Inspired by Figure 1, we consider the following simple model.

A1) *Data generating model:* The data points $X_1, \ldots, X_n$ are generated i.i.d. from a mixture of two Gaussians of the form $\frac{1}{2}\mathcal{N}(\mu_1, \Sigma) + \frac{1}{2}\mathcal{N}(\mu_2, \Sigma)$ in $\mathbb{R}^d$.

The assumption that the components have equal weight and equal covariance is made largely for expositional simplicity. For the reasons discussed in section 4.2, we believe that extending our results to allow for arbitrary mixture weights and differing component covariances is possible without introducing any major technical issues, and should involve no more than some additional bookkeeping. In fact, in section 4) we demonstrate a successful application of our proposed approach to a mixture with unequal weights and covariances. On the other hand, extending to more than two mixture components is a significant challenge, and addressing it is out of the scope of this paper.

The error of a clustering $\psi : \mathbb{R}^d \to \{1, 2\}$ is defined as follows. Let X be a random draw from the true mixture, and let $Y \in \{1, 2\}$ be the (latent) label of the mixture component from which X was drawn, i.e., $Y - 1 \sim \text{Bern}\left(\frac{1}{2}\right)$ and $X|Y \sim \mathcal{N}(\mu_Y, \Sigma)$. We define the *overlap* of the clustering ψ as $\Upsilon(\psi) := \min_\pi \mathbb{P}(\psi(X) \neq \pi(Y))$ where the minimum is over permutations $\pi : \{1, 2\} \to \{1, 2\}$, and the error of ψ is defined as $L(\psi) := \Upsilon(\psi) - \min_{\psi'} \Upsilon(\psi')$.

We define the optimal clustering $\psi^* := \text{argmin}_\psi \Upsilon(\psi)$, which coincides with the Bayes optimal classifier in the supervised problem of predicting Y from X:

$$\psi^*(x) = \begin{cases} 1 & \text{if} \quad (\mu_0 - x)^\intercal \beta < 0, \\ 2 & \text{o.w.} \end{cases} \tag{1}$$

where $\beta = \Sigma^{-1}\Delta_\mu$, $\mu_0 = \frac{\mu_1 + \mu_2}{2}$ and $\Delta_\mu = \frac{\mu_1 - \mu_2}{2}$. Notice that the Bayes optimal decision boundary is linear and hence the problem corresponds to linear discriminant analysis (LDA).

If the labels are known, one can simply plug-in sample estimates of class conditional means $\widehat{\mu}_Y$ and covariance matrix $\widehat{\Sigma}$ to obtain an empirical classification rule. In clustering, the labels are latent. However, if we can learn the parameters μ_1, μ_2, Σ of the Gaussian mixture model, we can plug these in and obtain a similar empirical clustering.

In the high-dimensional setting ($n \ll d$), estimates of the covariance matrix are typically not invertible, necessitating some additional assumptions to make the problem well-posed. In high-dimensional clustering, it is natural to expect that not all features are relevant for clustering. For example, in clustering proteins based on their drug expression profiles, not all drugs are responsible for differentiation of the expressions. This assumption can be captured as follows (using the notation $[d] = \{1, \ldots, d\}$).

A2) *Sparsity of relevant features:* The set of relevant features $S \subseteq [d]$, which are given by the non-zero coordinates of β, satisfy $|S| \leq s$, where $s \leq d$ is the sparsity level.

This notion of feature relevance is motivated by the fact that the optimal clustering ψ^* in Eq. 1 depends on a given feature only when the corresponding coordinate of β is non-zero.

We will demonstrate that the sample complexity of clustering in high dimensions depends on the number of non-zero coordinates $\|\beta\|_0 = |S| \leq s$, and only logarithmically on the total number of features d.

In comparison, existing work on high-dimensional clustering typically assumes Δ_μ is sparse, and the relevant features are given by its non-zero coordinates, i.e., the coordinates along which mean separation occurs. So, they cannot identify relevant features such as x_1 in Figure 1. Also, some existing work on high-dimensional GMM learning assumes sparsity of the covariance Σ or its inverse Σ^{-1}. These assumptions used in previous work are more restrictive than (and can be considered special cases of) our notion of relevant features (nonzero coordinates of $\beta \equiv \Sigma^{-1}\Delta_\mu$) which is precisely what the optimal clustering function depends on.

We make the following additional assumption which guarantees success of our computationally feasible method that uses the ℓ_1 penalty.

A3) *Restricted eigenvalue property:* The covariance matrix Σ satisfies

$$\min_{S\subseteq[d]:\,|S|\leq s, v\neq 0} \left\{ \frac{\|\Sigma v\|_2}{\|v\|_2} : \|v_{S^c}\|_1 \leq \|v_S\|_1 \right\} \geq \eta > 0$$

where v_S is the projection of v onto the coordinates in S, and $S^c = [d] \setminus S$ is the complement of S. A similar assumption is required for feature selection in supervised learning using ℓ_1 penalties (c.f. [28]). This condition ensures that there cannot exist two different values of the sparse vector β which correspond to similar values for $\Sigma\beta = \Delta_\mu$, and hence that a small error in estimating the parameters (either Δ_μ or Σ) must imply small clustering and feature selection error.

While the above assumptions suffice to evaluate the clustering performance in high dimensions, we also seek to correctly identify the set of relevant features. For this, we need to assume that each relevant feature is "relevant enough" to be detectable using a finite sample. Formally,

A4) *Signal strength along each relevant feature:* For each $i \in [d]$ such that $\beta(i) \neq 0$, let $|\beta(i)| \geq \beta_{\min}$, where $\beta_{\min} > 0$.

3 Proposed Method

Given samples $X_1, \ldots, X_n$ from the unknown mixture $\frac{1}{2}\mathcal{N}(\mu_1, \Sigma) + \frac{1}{2}\mathcal{N}(\mu_2, \Sigma)$, we propose a procedure composed of three stages. First we aquire initial estimates of mixture parameters using the algorithm of Hardt and Price [11]. Next we estimate the discriminating direction $\beta := \Sigma^{-1}\Delta_\mu$ by means of solving a convex program analogous to the proposal of Cai and Liu [27] for sparse supervised linear classification. Finally we threshold the elements of the estimate of β to recover the relevant features S.

Precisely, the steps are as follows.

1. Obtain estimates $\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\Sigma}$ by invoking Algorithm HardtPrice defined in Section 3.1 with ϵ, δ satisfying $\epsilon = C(\log(dn/\delta)/n)^{1/6}$ for some constant C.

2. For some $\lambda > 0$, set

$$\widehat{\beta}_\lambda = \operatorname*{argmin}_{z\in\mathbb{R}^d} \|z\|_1 \quad \text{subject to } \|\widehat{\Sigma}z - \widehat{\Delta}_\mu\|_\infty \leq \lambda \tag{2}$$

   where $\widehat{\Delta}_\mu = \frac{\widehat{\mu}_1 - \widehat{\mu}_2}{2}$, $\|\cdot\|_\infty$ is the elementwise absolute maximum, and λ is a tuning parameter the choice of which is discussed below. Let $\widehat{\mu}_0 = \frac{\widehat{\mu}_1 + \widehat{\mu}_2}{2}$. The estimated clustering is defined as

$$\widehat{\psi}_\lambda(x) = \begin{cases} 1 & \text{if} \quad (\widehat{\mu}_0 - x)^\intercal \widehat{\beta}_\lambda < 0, \\ 2 & \text{o.w.} \end{cases}$$

Proposition 1 in Section 4 ties the error in the estimates of $\widehat{\Sigma}$ and $\widehat{\Delta}_\mu$ to the error of $\widehat{\psi}_\lambda$. In this result, the bound on the clustering error is minimized when λ takes on the smallest value such that the true β is a feasible point of the constraints in (2). A specific value for λ is given in Corollary 3, based on a few additional technical assumptions.

3. Estimate the relevant features S by thresholding $\widehat{\beta}_\lambda$:

$$\widehat{S} = \{i : \widehat{\beta}_\lambda(i) > c \cdot \lambda\sqrt{s}\}$$

   where $c > 0$ is a constant.

Our results for support recovery hold when $c > 2/\eta$.

3.1 Algorithm HardtPrice

We describe below the algorithm HardtPrice proposed by Hardt and Price [11] (section 3, algorithm $\mathcal{B}$), simplified in accordance to the assumption (A1). Specifically, the version of the algorithm we state here skips the steps necessary to learn different component variances and the mixture weights.

The HardtPrice algorithm assumes the availability of another algorithm, GMFitLowDim, for mixture learning in a low-dimensional setting. The latter is

any algorithm that, like HardtPrice, takes as input a set of samples as described in (A1) together with parameters $\epsilon, \delta > 0$, and has the same properties as those of HardtPrice stated in Theorem 2, but only for up to 2 dimensional mixtures.

Hardt and Price give a candidate for GMFitLowDim, which combines a type of moment method approach and a grid search over parameters. The algorithm involves a large number of steps, and we do not restate it here due to the space constraint. Since much of the computational and statistical difficulty of general Gaussian mixture learning is not present when only considering such small dimensional cases, we believe this does not take away from the exposition. In fact, in our simulation experiments (section 5) we successfully use an "off-the-shelf" EM based maximum likelihood mixture learning algorithm as GMFitLowDim.

Input: Samples $X_1, \ldots, X_n \in \mathbb{R}^d$; $\epsilon, \delta > 0$.

1. Set $\widehat{V} = \max_{i \in [d]} \sum_{j=1}^n \frac{X_j(i)^2}{n} - \left(\sum_{j=1}^n \frac{X_j(i)}{n}\right)^2$, $\epsilon^* = \frac{\epsilon}{20}$, $\delta^* = \frac{\delta}{10d^2}$. Algorithm GMFitLowDim will always be invoked with parameters ϵ^* and δ^*.

Estimate $\widehat{\mu}_1$ and $\widehat{\mu}_2$:

2. For each $i \in [d]$, use GMFitLowDim on the univariate data $X_1(i), \ldots, X_n(i)$ obtaining estimates of the means $\xi_1(i)$ and $\xi_2(i)$.
3. If $|\xi_1(i) - \xi_2(i)| \leq \epsilon \widehat{V}/4$ for all $i \in [d]$, put $\widehat{\mu}_1 = \widehat{\mu}_2 = \xi_1$ (and skip step 4).
4. Otherwise, let i be the smallest index such that $|\xi_1(i) - \xi_2(i)| > \epsilon \widehat{V}/4$ and, for each $j \in [d] \backslash \{i\}$ do:
   a) Apply GMFitLowDim to the bivariate data $[X_1(i), X_1(j)], \ldots, [X_n(i), X_n(j)]$ to obtain mean estimates $(\nu_k(i), \nu_k(j))$ for $k = 1, 2$.
   b) Let $k \in \{1, 2\}$ such that $|\xi_1(i) - \nu_k(i)| \leq \epsilon \widehat{V}/10$. If such k does not exist, the algorithm terminates with failure.
   c) Set $\widehat{\mu}_1(j) = \nu_k(j)$ and $\widehat{\mu}_2(j) = \nu_{3-k}(j)$.

Estimate $\widehat{\Sigma}$:

5. For each $i \in [d]$, invoke GMFitLowDim on the univariate data $X_1(i), \ldots, X_n(i)$ and obtain an estimate of the diagonal element $\widehat{\Sigma}(i; i)$.
6. For each $i < j$, invoke GMFitLowDim on the bivariate data $[X_1(i), X_1(j)], \ldots, [X_n(i), X_n(j)]$ and obtain an estimate of $\widehat{\Sigma}(i, j) = \widehat{\Sigma}(j, i)$.
7. Return $\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\Sigma}$.

Intuitively, the algorithm works as follows. Note that the marginal of a Gaussian mixture is the same as a mixture of the marginals of the mixture components. So, given sufficient data, the univariate mixture mean estimates $\xi_1(i)$ and $\xi_2(i)$ learned in step 2 of the algorithm will be close to $\mu_1(i)$ and $\mu_2(i)$ for all $i \in [d]$, up to ordering. I.e., having $\xi_1(i)$ and $\xi_2(i)$, it remains only to decide whether $\xi_1(i)$ corresponds to the same mixture component as $\xi_1(j)$, or to $\xi_2(j)$ instead, for each other $j \in [d]$. To do this, in step 4 the algorithm looks at bivariate marginals of feature pairs i, j, and matches $\xi_1(i)$ to whichever one of $\xi_1(j)$ and $\xi_2(j)$ it co-occurs with in the bivariate marginal.

Similarly, if the component covariances are identical as we assume, then we only need to look at bivariate marginals to get $\Sigma(i, i)$, $\Sigma(i, j)$, and $\Sigma(j, j)$. Hardt and Price allow for different component covariances as well, where, similarly to learning the means, it is necessary to decide if $\Sigma_1(i, j)$ belongs to the same component as $\Sigma_1(k, l)$ or $\Sigma_2(k, l)$. To do this, Hardt and Price make use of up to 4-dimensional marginals.

In the first stage, the algorithm fits mixtures independently to the univariate projections of the data (step 2). However, it is important to note that the subsequent steps, which use bivariate projections of the data, recover information that cannot be obtained by purely marginal methods. For instance, the marginal method of [24] would fail to identify feature x_1 in the example in Figure 1 as relevant to the mixture, even with infinite data. The Hardt and Price algorithm, on the other hand, succeeds in this case, as demonstrated in section 5.

4 Main Result

Our first result states that if the parameters of the Gaussian mixture model in (A1) can be learnt accurately in ℓ_∞ norm, then the misclustering rate of the proposed method is small.

Proposition 1. *Assume (A1). For any ϵ, if* $\max\left(\|\mu_1 - \widehat{\mu}_{\pi(1)}\|_\infty^2, \|\mu_2 - \widehat{\mu}_{\pi(2)}\|_\infty^2, \|\Sigma - \widehat{\Sigma}\|_\infty\right) \leq \epsilon$ *for some permutation $\pi : \{1, 2\} \to \{1, 2\}$, and if $\epsilon\|\beta\|_1 + \sqrt{\epsilon} \leq \lambda$, then*

$$L(\widehat{\psi}_\lambda) \leq \phi\left(\max\left(\frac{\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu - \epsilon_1}{\sqrt{\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu + \epsilon_2}}, 0\right)\right) \frac{\epsilon_1 + \epsilon_2}{\sqrt{\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu}}$$

where $\epsilon_1 = (2\lambda + 3\sqrt{\epsilon})\|\beta\|_1$, $\epsilon_2 = \epsilon\|\beta\|_1^2 + 3(\lambda + \sqrt{\epsilon})\|\beta\|_1$, and ϕ is the standard normal density.

This result is similar to the classical results in classification error analysis of Fisher's Linear Discriminant, but with the key difference that the misclustering rate is bounded in terms of the ℓ_∞ norms of the errors of

the parameter estimates. This is crucial, as it will subsequently allow us to obtain a rate that is only logarithmic in the ambient dimension d. Before giving the proof, we notice that the misclustering rate depends on $\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu$ which can be regarded as the signal energy.

Proof. Since the clustering error does not change upon flipping the labels assigned by $\widehat{\psi}_\lambda$, WLOG we assume π is the identity permutation.

It is easy to verify that

$$\Upsilon(\psi^*) = \Phi\left(-\frac{\Delta_\mu^\intercal \beta}{\sqrt{\beta^\intercal \Sigma \beta}}\right) = \Phi\left(-\sqrt{\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu}\right)$$

where Φ is the standard normal CDF (whereas ϕ denotes the PDF). Also,

$$\Upsilon(\widehat{\psi}_\lambda) = \frac{1}{2}\Phi\left(-\frac{|\Delta_\mu^\intercal \widehat{\beta}_\lambda| + |(\mu_0 - \widehat{\mu}_0)^\intercal \widehat{\beta}_\lambda|}{\sqrt{\widehat{\beta}_\lambda^\intercal \Sigma \widehat{\beta}_\lambda}}\right) + \\ + \frac{1}{2}\Phi\left(-\frac{|\Delta_\mu^\intercal \widehat{\beta}_\lambda| - |(\mu_0 - \widehat{\mu}_0)^\intercal \widehat{\beta}_\lambda|}{\sqrt{\widehat{\beta}_\lambda^\intercal \Sigma \widehat{\beta}_\lambda}}\right)$$

where the appearance of the absolute values is to account for the minimum over permutations in the definition of Υ – the overlap must not change if $\widehat{\beta}_\lambda$ is negated.

So,

$$\begin{aligned} L(\widehat{\psi}_\lambda) =& \Upsilon(\widehat{\psi}_\lambda) - \Upsilon(\psi^*) \\ \leq& \Phi\left(-\frac{|\Delta_\mu^\intercal \widehat{\beta}_\lambda| - |(\mu_0 - \widehat{\mu}_0)^\intercal \widehat{\beta}_\lambda|}{\sqrt{\widehat{\beta}_\lambda^\intercal \Sigma \widehat{\beta}_\lambda}}\right) - \\ & - \Phi\left(-\sqrt{\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu}\right). \end{aligned} \tag{3}$$

Clearly, $\|\mu_0 - \widehat{\mu}_0\|_\infty \leq \sqrt{\epsilon}$ and $\|\Delta_\mu - \widehat{\Delta}_\mu\|_\infty \leq \sqrt{\epsilon}$. Since $\Delta_\mu = \Sigma\beta$,

$$\begin{aligned} \|\widehat{\Sigma}\beta - \widehat{\Delta}_\mu\|_\infty &\leq \|\widehat{\Sigma}\beta - \Delta_\mu\|_\infty + \|\Delta_\mu - \widehat{\Delta}_\mu\|_\infty \\ &\leq \|\widehat{\Sigma}\beta - \Sigma\beta\|_\infty + \sqrt{\epsilon} \\ &\leq \|\widehat{\Sigma} - \Sigma\|_\infty \|\beta\|_1 + \sqrt{\epsilon} \\ &\leq \epsilon\|\beta\|_1 + \sqrt{\epsilon} \leq \lambda \end{aligned}$$

which implies that β is a feasible point for the optimization problem (2). Hence, since $\widehat{\beta}_\lambda$ is an optimum for (2), $\|\widehat{\beta}_\lambda\|_1 \leq \|\beta\|_1$, and

$$|(\mu_0 - \widehat{\mu}_0)^\intercal \widehat{\beta}_\lambda| \leq \|\mu_0 - \widehat{\mu}_0\|_\infty \|\widehat{\beta}_\lambda\|_1 \leq \sqrt{\epsilon}\|\beta\|_1.$$

Next,

$$\begin{aligned} |\Delta_\mu^\intercal \widehat{\beta}_\lambda| &\geq |\Delta_\mu^\intercal \beta| - |\Delta_\mu^\intercal (\widehat{\beta}_\lambda - \beta)| \\ &= \Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu - |\Delta_\mu^\intercal (\widehat{\beta}_\lambda - \beta)| \end{aligned}$$

where

$$\begin{aligned} |\Delta_\mu^\intercal (\widehat{\beta}_\lambda - \beta)| &\leq |\widehat{\beta}_\lambda^\intercal (\widehat{\Sigma}\beta - \Delta_\mu)| + |\beta^\intercal (\widehat{\Sigma}\widehat{\beta}_\lambda - \Delta_\mu)| \\ &\leq \|\beta\|_1 \left(\|\widehat{\Sigma}\beta - \Delta_\mu\|_\infty + \|\widehat{\Sigma}\widehat{\beta}_\lambda - \Delta_\mu\|_\infty\right) \\ &\leq \|\beta\|_1 \left(\|\widehat{\Sigma}\beta - \widehat{\Delta}_\mu\|_\infty + \|\widehat{\Sigma}\widehat{\beta}_\lambda - \widehat{\Delta}_\mu\|_\infty + \right. \\ &\qquad \left. + 2\|\Delta_\mu - \widehat{\Delta}_\mu\|_\infty\right) \\ &\leq 2(\lambda + \sqrt{\epsilon})\|\beta\|_1 \end{aligned}$$

i.e.,

$$\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu - \left(|\Delta_\mu^\intercal \widehat{\beta}_\lambda| - |(\mu_0 - \widehat{\mu}_0)^\intercal \widehat{\beta}_\lambda|\right) \leq \\ \leq (2\lambda + 3\sqrt{\epsilon})\|\beta\|_1 \equiv \epsilon_1.$$

And

$$\begin{aligned} \widehat{\beta}_\lambda^\intercal \Sigma \widehat{\beta}_\lambda &\leq \beta^\intercal \Sigma \beta + |\widehat{\beta}_\lambda^\intercal \Sigma \widehat{\beta}_\lambda - \beta^\intercal \Sigma \beta| \\ &\leq \beta^\intercal \Sigma \beta + |\widehat{\beta}_\lambda^\intercal \Sigma \widehat{\beta}_\lambda - \widehat{\beta}_\lambda^\intercal \Delta_\mu| + |(\widehat{\beta}_\lambda - \beta)^\intercal \Delta_\mu| \\ &\leq \Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu + \|\Sigma\widehat{\beta}_\lambda - \Delta_\mu\|_\infty \|\beta\|_1 + \\ &\quad + 2(\lambda + \sqrt{\epsilon})\|\beta\|_1 \end{aligned}$$

where

$$\begin{aligned} \|\Sigma\widehat{\beta}_\lambda - \Delta_\mu\|_\infty &\leq \|\Sigma\widehat{\beta}_\lambda - \widehat{\Sigma}\widehat{\beta}_\lambda\|_\infty + \|\widehat{\Sigma}\widehat{\beta}_\lambda - \widehat{\Delta}_\mu\|_\infty + \\ &\quad + \|\widehat{\Delta}_\mu - \Delta_\mu\|_\infty \\ &\leq \|\Sigma - \widehat{\Sigma}\|_\infty \|\beta\|_1 + \lambda + \sqrt{\epsilon} \\ &\leq \epsilon\|\beta\|_1 + \lambda + \sqrt{\epsilon} \end{aligned} \tag{4}$$

so

$$\widehat{\beta}_\lambda^\intercal \Sigma \widehat{\beta}_\lambda - \Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu \leq \epsilon\|\beta\|_1^2 + 3(\lambda + \sqrt{\epsilon})\|\beta\|_1 \equiv \epsilon_2.$$

Let $L = \frac{\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu - \epsilon_1}{\sqrt{\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu + \epsilon_2}}$; combining these with (3),

$$\begin{aligned} L(\widehat{\psi}_\lambda) &\leq \Phi(-L) - \Phi(-\sqrt{\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu}) \\ &\leq \phi(\max(L,\ 0))\left(\sqrt{\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu} - L\right) \\ &\leq \phi(\max(L,\ 0))\frac{\epsilon_1 + \epsilon_2}{\sqrt{\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu}}. \end{aligned}$$

□

The following result from [11] provides us ℓ_∞ control over the GMM parameters.

Theorem 2 (Hardt and Price [11])**.** *Given $\epsilon, \delta > 0$ and n samples from the model (A1), if*

$$n = O\left(\frac{1}{\epsilon^6} \log\left(\frac{d}{\delta} \log\left(\frac{1}{\epsilon}\right)\right)\right),$$

then, with probability at least $1 - \delta$, Algorithm HardtPrice *in Section 3.1 produces estimates $\widehat{\mu}_1, \widehat{\mu}_2$ and $\widehat{\Sigma}$ such that, for some permutation $\pi : \{1, 2\} \to \{1, 2\}$,*

$$\max\left(\max_{i=1,2}\left(\|\mu_i - \widehat{\mu}_{\pi(i)}\|_\infty^2\right),\ \|\Sigma - \widehat{\Sigma}\|_\infty\right) \leq \\ \leq \epsilon\left(\frac{1}{4}\|\mu_1 - \mu_2\|_\infty^2 + \|\Sigma\|_\infty\right).$$

Combining Proposition 1 and Theorem 2, we have the following result under (A2) and (A3). We defer the proof to the supplement.

Corollary 3. *Assume (A1), (A2), (A3), $\|\Sigma\|_2 \le D_0$, and $\|\mu_1 - \mu_2\|_\infty^2 < D$. Given $\delta > 0$, there is some constant c_1 such that, setting*

$$\lambda = c_1 \left(\frac{\log(dn/\delta)}{n}\right)^{1/6} \frac{\sqrt{D_0 s(\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu)}}{\eta} + \\ + \sqrt{c_1}\left(\frac{\log(dn/\delta)}{n}\right)^{1/12},$$

with probability at least $1-\delta$,

$$L(\widehat{\psi}_\lambda) \le C_0 \phi \left(\sqrt{\frac{\Delta_\mu^\intercal \Sigma^{-1} \Delta_\mu}{6}}\right) \times \\ \times \max\left[\frac{s\sqrt{\Delta_\mu^\intercal \Sigma^{-1}\Delta_\mu}}{\eta^2}\left(\frac{\log(dn/\delta)}{n}\right)^{1/6}, \right. \\ \left. \frac{\sqrt{s}}{\eta}\left(\frac{\log(dn/\delta)}{n}\right)^{1/12}\right]$$

for some constant C_0.

Remark: Hence it follows that $n = \Omega\left(s^6 \log(d)\right)$, suppressing the dependence on other parameters.

4.1 Recovery of relevant features

We derive a bound for $\|\beta - \widehat{\beta}_\lambda\|_\infty$ under (A3) and then guarantee recovery of relevant features under (A4), i.e., when the non-zero components of β are large enough.

Theorem 4. *Assume the conditions of Proposition 1 hold. We have*

$$\|\beta - \widehat{\beta}_\lambda\|_\infty \le \frac{2\lambda\sqrt{s}}{\eta}.$$

If, in addition, (A4) holds, $\eta > 2/c$, and $\beta_{\min} > 2c\lambda\sqrt{s}$, then

$$\widehat{S} = S.$$

Proof. We first establish two results that are crucial. First, if β is sparse and S denotes the support of β, then

$$\begin{aligned}\|(\beta - \widehat{\beta}_\lambda)_{S^c}\|_1 &= \|\widehat{\beta}_{\lambda,S^c}\|_1 = \|\widehat{\beta}_\lambda\|_1 - \|\widehat{\beta}_{\lambda,S}\|_1 \\ &\le \|\beta\|_1 - \|\widehat{\beta}_{\lambda,S}\|_1 = \|\beta_S\|_1 - \|\widehat{\beta}_{\lambda,S}\|_1 \\ &\le \|(\beta - \widehat{\beta}_\lambda)_S\|_1.\end{aligned}$$

Second, note that

$$\|\Sigma(\beta - \widehat{\beta}_\lambda)\|_\infty = \|\Delta_\mu - \Sigma\widehat{\beta}_\lambda\|_\infty \le 2\lambda$$

using Eq. (4).

Therefore, we can write

$$\begin{aligned}\|\beta - \widehat{\beta}_\lambda\|_\infty &\le \|\beta - \widehat{\beta}_\lambda\|_2 \\ &\le \frac{2\lambda}{\|\Sigma(\beta - \widehat{\beta}_\lambda)\|_\infty}\|\beta - \widehat{\beta}_\lambda\|_2 \\ &\le \frac{2\lambda\sqrt{s}}{\|\Sigma(\beta - \widehat{\beta}_\lambda)\|_2}\|\beta - \widehat{\beta}_\lambda\|_2 \\ &\le \frac{2\lambda\sqrt{s}}{\min_v\{\|\Sigma v\|_2/\|v\|_2 : \|v_{S^c}\|_1 \le \|v_S\|_1\}} \\ &\le \frac{2\lambda\sqrt{s}}{\eta}.\end{aligned}$$

The result follows using (A4) since $\beta_{\min} > 2c\lambda\sqrt{s} > \lambda\sqrt{s}\left(c + \frac{2}{\eta}\right)$. □

Corollary 5. *Assume (A1)-(A4). Under the conditions of Corollary 3, given $\delta > 0$, if*

$$\beta_{\min} = \omega\left(s\left(\frac{\log(dn/\delta)}{n}\right)^{1/12}\right),$$

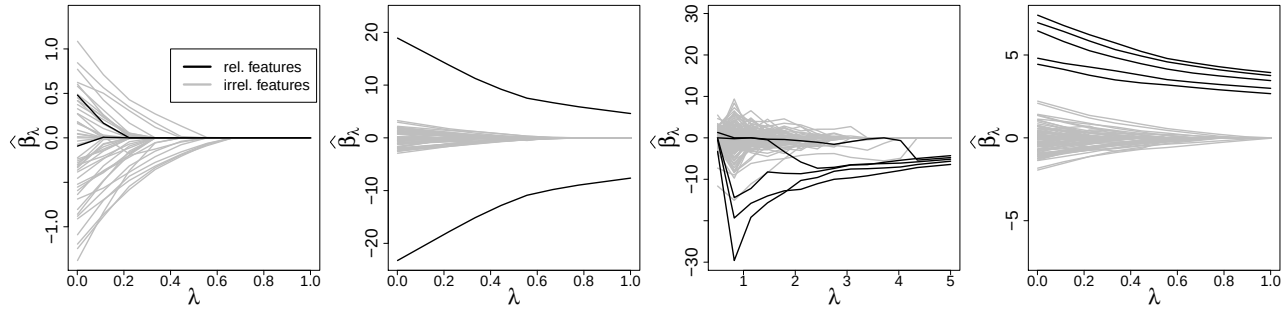
then $\widehat{S} = S$ with probability at least $1-\delta$.

Thus, in order to recover the support, we require at least $n = \Omega\left((s/\beta_{\min})^{12}\log(d)\right)$, suppressing the dependence on other parameters.

4.2 Relaxing model assumptions

When $\Sigma_1 \ne \Sigma_2$, the true clustering ψ^* defined in section 2 is no longer equivalent to the Bayes optimal decision rule for the latent classification problem, since the latter is in general quadratic. Hence, our results on the clustering error (Corollary 3) do not have a straightforward extension to the unequal covariance case.

However, the quantity $\beta := \Sigma_w^{-1}\Delta_\mu$, where now $\Sigma_w = p_1\Sigma_1 + p_2\Sigma_2$, does still have a meaningful interpretation when $\Sigma_1 \ne \Sigma_2$. Namely, it is the Fisher discriminant direction, which has been used in the Gaussian mixture learning literature (e.g. [10]) as a key quantity of interest to estimate, since projecting on this direction maximizes the between cluster variance. Corollary 5, which establishes a sample complexity for estimating the support of the vector β which is logarithmic in the dimension, depends only on the estimation of Δ_μ and Σ_w with ℓ_∞ norm control on the error. Hardt and Price [11] do provide results with precisely such a guarantee for learning Gaussian mixtures of the more general form $p_1\mathcal{N}(\mu_1, \Sigma_1) + p_2\mathcal{N}(\mu_2, \Sigma_2)$. Hence, the same proof technique should lead to an analogous result *without* assuming that $p_1 = p_2 = \frac{1}{2}$ or $\Sigma_1 = \Sigma_2 = \Sigma$.



(a) dataset 1, EM (b) dataset 1, Hardt-Price (c) dataset 2, Hardt-Price (d) dataset 3, Hardt-Price

Figure 2: Values of entries of $\widehat{\beta}_\lambda$ given by solving (2) using either EM or the Hardt and Price parameter estimates as input.

5 Simulation Results

In this section, we show the viability of our proposed method using simulated datasets. We implemented a version of the Hardt and Price algorithm, with minor modifications motivated by practicality, as well as a linear programming formulation for the problem in (2). We used the EM-based maximum likelihood Gaussian mixture learning algorithm in the R package EMCluster [29] as the subroutine GMFitLowDim in the Hardt and Price algorithm. R code is available in the supplementary material.

The first simulated dataset captures the issue depicted in Figure 1. We set $d = 50$, $\Delta_\mu = (2, 0, \ldots, 0)$, and $\Sigma = I$ except for $\Sigma(1,2) = \Sigma(2,1) = 0.85$. Clearly, the true support of $\beta = \Sigma^{-1}\Delta_\mu$ is $\{1, 2\}$, so there are $s = 2$ relevant variables. After sampling $n = 200$, we first apply the EM algorithm in the EMCluster package to the full dimensional dataset, and use the estimated parameters as input to (2). The results (Figure 5a) clearly show that this approach fails to identify the relevant features. In contrast, using the Hardt and Price parameter estimates (Figure 5b), the relevant features stand out for a wide range of values for λ.

The second dataset is high-dimensional – we draw $n = 150$ samples from a mixture with $d = 200$, where we set β to have $s = 5$ non-zero coordinates each set to 5, Σ was generated randomly from a Wishart distribution with degrees of freedom $2d$ and subsequently rescaled to have eigenvalues in $[0.5, 2]$, and $\mu_1 = -\mu_2 = \Sigma\beta$. The results in Figure 5c show the importance of the sparsifying effect from using $\lambda > 0$ in (2), since the coefficients of the relevant features are not isolated until λ is increased to ≈ 4.5, at which point each of the 5 relevant features is identified.

Finally, the third dataset serves to show that the equal component weight and equal component covariance assumptions are not crucial to the proposed method, when using the original algorithm of Hardt and Price rather than the simplified version in section 3.1. Specifically, we violate the model in (A1) and draw $n = 500$ samples from $p_1\mathcal{N}(\mu_1, \Sigma_1) + p_2\mathcal{N}(\mu_2, \Sigma_2)$, where $p_1 = 0.4 = 1 - p_2$, $d = 100$, $\Sigma_1 \neq \Sigma_2$ each generated independently as above, β is set to have $s = 5$ non-zero values each of which is 2, and $\mu_1 = -\mu_2 = (p_1\Sigma_1 + p_2\Sigma_2)\beta$. As demonstrated by the results in Figure 5d, the proposed method is indeed applicable to this more general setting.

6 Discussion and Open questions

The primary goal of this paper was to demonstrate a method for high-dimensional clustering which, in contrast to existing work, provably identifies relevant features that are not distinguished by the marginal separation of component means alone. The method we present is computationally feasible and statistically efficient with sample complexity that primarily depends on the number of relevant dimensions, and only logarithmically on the total number of features. However, this goal was achieved by considering a very simple model - a mixture of two non-spherical Gaussians with same covariance and mixture weights. While we believe that it will be straightforward to adapt our results to allow uneven mixture weights and different covariance matrices, extending our approach to handle more than two components is a difficult problem and is the topic of ongoing work. Theoretically, the bounds we have demonstrated can be tightened in a few places, particularly for support recovery using a primal-dual witness argument. Additionally, it will be interesting to demonstrate matching lower bounds for this problem to establish optimality of the sample complexity.

Acknowledgements

This research is supported in part by NSF CAREER grant IIS-1252412.

References

[1] Wei Sun, Junhui Wang, and Yixin Fang. Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, 6:148–167, 2012.

[2] Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490), 2010.

[3] Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, 66(3):793–804, 2010.

[4] Jeffrey L Andrews and Paul D McNicholas. Variable selection for clustering and classification. *ArXiv e-prints 1303.5294*, March 2013.

[5] Xiangyu Chang, Yu Wang, Rongjian Li, and Zongben Xu. Sparse K-Means with ℓ_∞/ℓ_0 Penalty for High-Dimensional Data Clustering. *ArXiv e-prints 1403.7890*, March 2014.

[6] Yao-ban Chan and Peter Hall. Using evidence of mixed populations to select variables for clustering very high-dimensional data. *Journal of the American Statistical Association*, 105(490), 2010.

[7] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *COLT*, volume 18, page 458. Springer, 2005.

[8] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *COLT*, pages 444–457. Springer-Verlag, 2005.

[9] Kamalika Chaudhuri and Satish Rao. Learning mixtures of product distributions using correlations and independence. In *COLT*, pages 9–20, 2008.

[10] S Charles Brubaker and Santosh S Vempala. Isotropic pca and affine-invariant clustering. In *Building Bridges*, pages 241–281. Springer, 2008.

[11] Moritz Hardt and Eric Price. Sharp bounds for learning a mixture of two gaussians. *ArXiv e-prints 1404.4997*, April 2014.

[12] Cathy Maugis, Gilles Celeux, and Marie-Laure Martin-Magniette. Variable selection for clustering with gaussian mixture models. *Biometrics*, 65(3):701–709, 2009.

[13] Hyangmin Lee and Jia Li. Variable selection for clustering by separability based on ridgelines. *Journal of Computational and Graphical Statistics*, 21(2):315–337, 2012.

[14] Adrian E Raftery and Nema Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

[15] Wei Pan and Xiaotong Shen. Penalized model-based clustering with application to variable selection. *The Journal of Machine Learning Research*, 8:1145–1164, 2007.

[16] Stephane Gaiffas and Bertrand Michel. Sparse Bayesian Unsupervised Learning. *ArXiv e-prints 1401.8017*, January 2014.

[17] Cathy Maugis and Bertrand Michel. A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM: Probability and Statistics*, 15:41–68, January 2011.

[18] Cathy Maugis and Bertrand Michel. Slope heuristics for variable selection and clustering via gaussian mixtures. *Technical Report 6550, INRIA*, 2008.

[19] Akshay Krishnamurthy. High-dimensional clustering with sparse gaussian mixture models. Available at `http://www.cs.cmu.edu/~akshaykr/files/sgmm_paper.pdf`.

[20] Anani Lotsi and Ernst Wit. High dimensional sparse gaussian graphical mixture model. http://arxiv.org/abs/1308.3381.

[21] Ruan L., Yuan M., and Zou H. Regularized parameter estimation in high-dimensional gaussian mixture models. *Neural Computation*, 23:1605–1622, 2011.

[22] Martin H Law, Anil K Jain, and Mário Figueiredo. Feature selection in mixture-based clustering. In *Advances in Neural Information Processing Systems*, pages 625–632, 2002.

[23] Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *The Journal of Machine Learning Research*, 5:845–889, 2004.

[24] Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems*, pages 2139–2147, 2013.

[25] Jinfeng Yi, Lijun Zhang, Jun Wang, Rong Jin, and Anil Jain. A single-pass algorithm for efficiently recovering sparse cluster centers of high-dimensional data. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 658–666, 2014.

[26] Ery Arias-Castro and Nicolas Verzelen. Detection and Feature Selection in Sparse Mixture Models. *ArXiv e-prints 1405.1478*, May 2014.

[27] Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496), 2011.

[28] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[29] Wei-Chen Chen, Ranjan Maitra, and Volodymyr Melnykov. EMCluster: EM algorithm for model-based clustering of finite mixture gaussian distribution, 2012. R Package, URL http://cran.r-project.org/package=EMCluster.