Towards Clustering High-dimensional Gaussian Mixture Clouds in Linear Running Time

Dan Kushnir
Nokia Bell Laboratories

Shirin Jalali
Nokia Bell Laboratories

Iraj Saniee
Nokia Bell Laboratories

Abstract

Clustering mixtures of Gaussian distributions is a fundamental and challenging problem. State-of-the-art theoretical work on learning Gaussian mixture models has mostly focused on estimating the mixture parameters, where clustering is given as a byproduct. These methods have focused mostly on improving separation bounds for different mixture classes, and doing so in polynomial time and sample complexity. Less emphasis has been given to aligning these algorithms to the challenges of big data. In this paper, we focus on clustering n samples from an arbitrary mixture of c-separated Gaussians in $\mathbb{R}^p$ in time that is linear in p and n, and sample complexity that is independent of p. Our analysis suggests that for sufficiently separated Gaussians after $o(\log p)$ random projections a good direction is found that yields a small clustering error. Specifically, for a user-specified error e, the expected number of such projections is small and bounded by $o(\ln p)$ when $\gamma \leq c\sqrt{\ln \ln p}$ and $\gamma = Q^{-1}(e)$ is the separation of the Gaussians with Q as the tail distribution function of the normal distribution. Consequently, the expected overall running time of the algorithm is linear in n and quasi-linear in p at $o(\ln p)O(np)$, and the sample complexity is independent of p. Unlike the methods that are based on k-means, our analysis is applicable to any mixture class (spherical or non-spherical). Finally, an extension to $k > 2$ components is also provided.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

1 Introduction

Clustering Gaussian mixture models (GMMs) is a fundamental problem in machine learning that has been the subject of extensive research by statisticians and computer scientists. The goal is to minimize the clustering error probability, which is defined as the probability that the label of the Gaussian that has generated a point disagrees with its assigned label, up to a fixed permutation of the labels.

However, despite all the research in this area, in practice, the well-known and relatively old expectation maximization (EM) and k-means algorithms [Dempster et al.(1977), Lloyd(1982)] and their variants, while lacking required convergence guarantees, remain the most popular clustering methods. One reason is that modern data samples are typically high-dimensional and while the computational complexities of other proposed theoretical methods are polynomial in the ambient dimension and number of samples, they are still prohibitively large for practical purposes. For instance, there has been an extensive body of research on learning (estimating) the parameters of a GMM based on its samples with a running time polynomial in ambient dimension and number of samples (refer to [Huggins(2011)] for an overview of early methods in this line of work). Of course, once the parameters are learned with sufficient accuracy, then as a byproduct, one can cluster the points by assigning each point to the Gaussian cloud with highest posterior probability. Another practical challenge for methods developed in this area is that accurately learning the parameters in a high-dimensional Gaussian entails having a sample complexity which is again too large for practical purposes.

Given the mentioned challenges in clustering high-dimensional Gaussian distributions in $\mathbb{R}^p$ with a number of samples that is small compared to p, it is desirable to have computationally-efficient algorithms with a sample complexity not growing with the ambient dimension. Clearly, developing such methods is infeasible using techniques that are based on learning

the parameters of the Gaussians in $\mathbb{R}^p$. As a trivial example, learning the covariance matrix of an arbitrary Gaussian distribution in $\mathbb{R}^p$ requires at least p samples. Therefore, in this paper, we focus on developing efficient algorithms for model-based clustering of GMMs that do not require learning the parameters of the high-dimensional Gaussians.

Dimensionality reduction is a well-known and powerful tool that is employed to solve various machine learning problems including clustering and also GMM parameter estimation. Classical dimension reduction techniques aim at preserving certain metrics (e.g. Principal Component analysis [Pearson(1901)], LLE [Roweis and Saul(2000)], and more, see review in [Van Der Maaten et al.(2009)]). In the context of Gaussian parameters estimation, projection to lower subspaces has been key (see review in [Huggins(2011)]): Spectral projection methods [Kannan et al.(2005), Vempala and Wang(2004), Belkin and Sinha(2010)] involve quadratic complexity in n or p, while random-projection-based techniques have been proposed early by [Dasgupta(1999)] for the class of shared covariance mixture, it had running time that is $O(dn^2 + ndp)$, where $d = O(\log \frac{k}{\epsilon\delta})$ is the number of projections. More recent methods cope with arbitrary separation, for example, [Belkin and Sinha(2015)] use deterministic projections with $\binom{p}{2k^2}$ time complexity. As mentioned above these methods have polynomial sample-complexity as they aim at parameter estimation (see a brief comparison in table 1). In our context of clustering, we note the projection-based methods of [Boutsidis et al.(2010)] for solving k-means with convergence guarantees and running time $O(np\lceil\varepsilon^{-2}k(\log k)^{-1}\rceil)$, and EM-based method of [Fern and Brodley(2003)]. However, [Boutsidis et al.(2010)] does not generalize to non-spherical clusters without degradation, and in [Fern and Brodley(2003)] running time as well as performance analysis are yet an open question. In this paper, we analyze and explore using dimensionality reduction to efficiently cluster points of an arbitrary mixture without attempting to learn the parameters of the GMM in the ambient dimension p.

We focus our analysis on a mixture of two c-separated Gaussian distributions, and study the distribution of their separation under a random projection. Consider two c-separated Gaussian distributions, $\mathcal{N}(\mathbf{m}_1, \Sigma_1)$ and $\mathcal{N}(\mathbf{m}_2, \Sigma_2)$ in $\mathbb{R}^p$, where

$$c = \frac{\|\mathbf{m}_1 - \mathbf{m}_2\|}{\sqrt{p}\left(\sqrt{\lambda_{\max}(\Sigma_1)} + \sqrt{\lambda_{\max}(\Sigma_2)}\right)}, \qquad (1)$$

and $\lambda_{\max}(\Sigma_i)$ denotes the the maximum eigenvalue of Σ_i, $i = 1, 2$. Our theoretical analysis sheds light on the separability of Gaussian clouds under a random 1-dimensional projection and suggests that if the two clouds are sufficiently-separated (e.g. $c \geq 0.5$), then after a handful of projections, a "proper" direction with small clustering error can be found. To illustrate the implication of these new results consider the projection-based method proposed in [Dasgupta(1999)], for learning the parameters of GMMs. The method is based on random projections, and for two c-separated clouds, it is shown that if the two Gaussians are projected into a random d-dimensional space, such that $d \geq \frac{C_1}{\epsilon^2} \ln \frac{2}{\delta}$, then with probability exceeding $1 - \delta$, the projected d-dimensional Gaussians are $c\sqrt{1-\epsilon}$-separated. Here C_1 is a universal parameter. Using our proposed dual approach, instead of seeking a d-dimensional projection, we perform multiple 1-dimensional projections, until two γ-separated Gaussians are found. While our analysis hold for all values of p, asymptotically, they imply that as $p \to \infty$, the expected number of projections required to achieve γ-separation is upper-bounded by $\frac{1}{2Q(\frac{\gamma}{c})}$, where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \mathrm{e}^{-\frac{u^2}{2}} du. \qquad (2)$$

This shows that, unlike [Dasgupta(1999)], where the achieved separation ($c\sqrt{1-\epsilon}$) is always smaller than c, in this case, it is possible to achieve γ-separation in 1-dimensional space, even if $\gamma > c$. Moreover, the computational cost, i.e., the required number of projections, is small, if γ is comparable with c. In particular, if $\gamma \leq c\sqrt{\ln \ln p}$ the number of projections required is proved to be sub-logarithmic in p. As an example, we note the real data set of the USPS digits which has a minimal separation of $c = 0.63$ (see Table 10 in [Dasgupta(2000)]).

Under the above condition we propose a $o(\ln p)O(np)$-time recipe for clustering arbitrary Gaussian distributions based on 1-dimensional random projections. Given a user-prescribed error e (s.t. $\gamma = Q^{-1}(e)$), after each random projection, i) the parameters of the projected Gaussians are learned using for instance method of moments (MoM) [Pearson(1894)] that runs in $O(n)$-time, ii) if the desired separation γ is achieved the process stops. Otherwise, a new random projection is performed until a theoretically-derived budget for e is exhausted. Since parameter learning is done in 1-dimension, the sample complexity of our algorithm is independent of the ambient dimension p: $O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$, where ϵ is an error in the mixture parameter estimation, and δ is the confidence.

We provide our main results in Section 2. For space-limit reasons the proofs are provided in supplemental material. Our algorithm is presented in Section 3, and its sample complexity is analyzed in section 4. Experimental validation is provided in section 5.

Table 1: Complexity comparison with parameter learning methods

Author	Method	GMM Class	Running-time Complexity	Sample Complexity	Sep.	Parameters\comments
Dasgupta	Random projection	Shared Spherical	$O(dn^2 + ndp)$	$k^{O(\log^2(1/\varepsilon\delta))}$	$\sqrt{p}$	d- num. projections, $\{\|\hat{\mu}_i - \mu_i\|\} \leq \varepsilon\sigma\sqrt{p}$
Arora et. al.	Distance based	Arbitrary GMM	$O(p^2\text{poly}(k)\log^2\frac{p}{\delta})$, $O(pn^2)$ distance computation	$O(p \cdot \text{poly}(k)\log\frac{p}{\delta})$	$\Omega(p^{\frac{1}{4}})$	k - num. Gaussians
Vempala et. al.	Spectral and distance-based	Spherical GMM	$poly(p,k)$, p^3 for SVD, $O(pn^2)$ distances	$poly(p,k)$	$\Omega(k^{\frac{1}{4}})$	
Kalai et. al.	Random projection and MoM	Arbitrary 2-GMM	$\text{poly}(p, \frac{1}{\epsilon}, \frac{1}{\delta}, \frac{1}{w}, \frac{1}{D_{1,2}})$, p^2 projections	same as running time	≥ 0	$D_{1,2}$ - distributions distance, ϵ its accuracy param.
Sinha et. al.	Determin. projection MoM	Arbitrary GMM	$\text{poly}(p, \frac{1}{\epsilon}, \frac{1}{\delta}, B)$, algorithm uses $\binom{p}{2k^2}$ projections	$\text{poly}(p, \frac{1}{\epsilon}, \frac{1}{\delta}, B)$	≥ 0	ϵ - L_2 error in params. B - radius of params. ball
This paper	**Random 1D Projections and MoM**	**Arbitrary 2-mixture**	**expected $o(\log p)O(np)$ for $\frac{\gamma}{c} \leq \sqrt{\ln\ln p}$**	$O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$	$\sqrt{p}$	γ - separation in 1D, for clustering error $e \leq Q(\gamma)$

2 Main Results

In this section we consider data that is generated according to a mixture of two Gaussian distributions $\mathcal{N}(\mathbf{m}_1, \Sigma_1)$ and $\mathcal{N}(\mathbf{m}_2, \Sigma_2)$, which are c-separated. We study the probability that a random projection achieves a 1-dimensional separation γ or higher, which can be directly related to a prescribed clustering error e as $e \leq Q(\gamma)$ for two Gaussian distributions. Moreover, we prove conditions for the number of 1-dimensional projections required to achieve separation γ to be sublogarithmic in p when γ (corresponding to a clustering error in 1-dimension) is similar to c. These results allow the construction of very efficient (and simple) clustering algorithms that run in $o(\ln p)O(np)$ for arbitrary mixtures, with sample complexity that is independent of p.

We divide the main results into two cases. The first case is when the two Gaussians are spherical balls. The second case is when Σ_1 and Σ_2 are arbitrary positive semi-definite matrices. We also demonstrate the extension of our theoretical analysis for a mixture of $k > 2$ Gaussians.

2.1 Mixture of spherical Gaussians

Consider the special case where $\Sigma_i = \sigma_i^2 I_p$, for $i = 1, 2$. We examine projecting points generated according to $w_1\mathcal{N}(\mathbf{m}_1, \Sigma_1) + w_2\mathcal{N}(\mathbf{m}_2, \Sigma_2)$ using a random vector $\mathbf{A} = (A_1, \ldots, A_p)$, where $A_1, \ldots, A_p$ are independent and identically distributed (i.i.d.) as $\mathcal{N}(0, 1)$. Using this projection, we derive a mixture of two Gaussians in $\mathbb{R}$. Conditioned on $\mathbf{A} = \mathbf{a}$, the two Gaussians $\mathcal{N}(\mathbf{m}_1, \Sigma_1)$ and $\mathcal{N}(\mathbf{m}_1, \Sigma_2)$ in $\mathbb{R}^p$ are mapped to $\mathcal{N}(\langle\mathbf{m}_1, \mathbf{a}\rangle, \sigma_1^2\|\mathbf{a}\|^2)$, and $\mathcal{N}(\langle\mathbf{m}_2, \mathbf{a}\rangle, \sigma_2^2\|\mathbf{a}\|^2)$, respectively. Therefore, the two projected distributions are γ-separation, if $|\langle\mathbf{m}_1, \mathbf{a}\rangle - \langle\mathbf{m}_2, \mathbf{a}\rangle| > \gamma(\sigma_1 + \sigma_2)\|\mathbf{a}\|$, or

$$|\langle\mathbf{m}_1 - \mathbf{m}_2, \mathbf{a}\rangle| > \gamma(\sigma_1 + \sigma_2)\|\mathbf{a}\|. \tag{3}$$

Since $\mathbf{A}$ is not a fixed vector, the question is that given the randomness in the generation of the projection vector $\mathbf{A}$, what is the probability that condition (3) holds. In other words, given $\mathbf{m}_1$, $\mathbf{m}_2$, σ_1 and σ_2, we are interested in $\text{P}(|\langle\mathbf{m}_1 - \mathbf{m}_2, \mathbf{A}\rangle| > \gamma(\sigma_1 + \sigma_2)\|\mathbf{A}\|)$, or

$$\text{P}\left(\left|\langle\frac{\mathbf{m}_1 - \mathbf{m}_2}{\|\mathbf{m}_1 - \mathbf{m}_2\|}, \frac{\mathbf{A}}{\|\mathbf{A}\|}\rangle\right| > \frac{\gamma(\sigma_1 + \sigma_2)}{\|\mathbf{m}_1 - \mathbf{m}_2\|}\right),$$

where $A_1, \ldots, A_p \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. The following key theorem derives a lower bound on this probability.

Theorem 1 *Consider two spherical Gaussian distributions $\mathcal{N}(\mathbf{m}_1, \sigma_1^2 I_p)$ and $\mathcal{N}(\mathbf{m}_2, \sigma_2^2 I_p)$ in $\mathbb{R}^p$. Consider projecting each point generated according to these Gaussian distributions using $\mathbf{A} = (A_1, \ldots, A_p)$, where $A_1, \ldots, A_p$ are i.i.d. $\mathcal{N}(0, 1)$. Given $\gamma > 0$, let*

$$c \triangleq \frac{\|\mathbf{m}_1 - \mathbf{m}_2\|}{(\sigma_1 + \sigma_2)\sqrt{p}}. \tag{4}$$

Then, the probability that the separation of the projected Gaussians is larger than γ is larger than

$$2\left(1 - \text{e}^{-\frac{p-1}{2}(\tau - \log(1+\tau))}\right) Q\left(\frac{\gamma}{c}\sqrt{\frac{(1 - \frac{1}{p})}{(1 - \frac{\gamma^2}{pc^2})}(1 + \tau)}\right), \tag{5}$$

where $\tau > 0$ is a free parameter.

In Lemma 1 below, we map the separation γ of the projected Gaussian distributions in $\mathbb{R}$ to the error probability of an optimal classifier that has access to the parameters of the two projected Gaussians. Intuitively, the higher the separation of the two projected Gaussians, the lower the associated classification error.

Lemma 1 *Consider points in $\mathbb{R}$ drawn from a mixture of two Gaussian distributions $w\mathcal{N}(m_1,\sigma_1)+(1-w)\mathcal{N}(m_2,\sigma_2)$. Assume that the two components of the mixture are c-separated. Then, the error probability of the optimal Bayesian classifier is smaller than $Q\left(\frac{c}{2}\right)$. In the special case where $\sigma_1=\sigma_2=\sigma$, the error probability of the optimal Bayesian classifier is smaller than $Q(c)$.*

Note that based on Lemma 1, if the two high-dimensional Gaussians share a covariance matrix, then a separation of $\gamma=Q^{-1}(e)$ in $\mathbb{R}$ is sufficient for achieving error e. If $\Sigma_1\neq\Sigma_2$, then γ is set as $2Q^{-1}(e)$.

Next, Lemma 2 shows that the expected value of the squared separation of the randomly projected Gaussian distributions is equal to c^2. Lemma 2 is later utilized to approximate the unknown separation c from the empirical expectation $\hat{\mathrm{E}}[\gamma^2]$ in 1-dimensional projections.

Lemma 2 *Consider $\mathbf{m}_1,\mathbf{m}_2\in\mathbb{R}^p$ and $\sigma_1,\sigma_2\in\mathbb{R}^+$ and define c as in (4). Then, under a random 1-dimensional projection with $\mathbf{A}=(A_1,\ldots,A_p)$, where $A_1,\ldots,A_p\overset{\text{i.i.d.}}{\sim}\mathcal{N}(0,1)$,*

$$\mathrm{E}\left[\frac{|\langle\mathbf{A},\mathbf{m}_1-\mathbf{m}_2\rangle|^2}{(\sigma_1+\sigma_2)^2\|\mathbf{A}\|^2}\right]=c^2. \tag{6}$$

In practice, given a desired accuracy e, Lemma 1 is used to derive γ, the desired separation in $\mathbb{R}$, that guarantees accuracy e. Then, using Lemma 2 to estimate the unknown separation c, one can compute the probability of attaining γ, or the expected number of projections, via Theorem 1.

Next, we derive bounds on the expected number of projections needed to achieve a separation γ. Let $d(\gamma)$ denote the expected number of projections required to achieve γ-separation in $\mathbb{R}$. The following corollaries use Theorem 1 to analyze and bound $d(\gamma)$ for different separation values of the two high-dimensional Gaussians.

Corollary 1 *Consider the same setup as in Theorem 1.Then,*

$$\lim_{p\to\infty} d(\gamma)\leq\frac{1}{2Q(\frac{\gamma}{c})}.$$

In the following corollary we establish the conditions on γ and c so that with a number of projections that is sub-logarithmic in p γ can be achieved.

Corollary 2 *Consider the same setup as in Theorem 1. If γ is such that $\gamma\leq c(\ln\ln p)^{\frac{1-\eta}{2}}$, where $\eta>0$ is a free parameter, then $d(\gamma)=o(\ln p)$.*

In a similar manner Corollary 3 captures the tradeoff between the number of projections and the resulting 1-dimensional separation for $\gamma=(\ln\ln p)^{\frac{1-\eta}{2}}$ with $d=o(\ln p)$ projections. This result provides a substantially higher running-time but for a tradeoff in the accuracy. The proof follows similarly to the proof of Corollary (2).

Corollary 3 *Consider the same setup as in Theorem 1. If γ is such that $\gamma\leq c(\ln p)^{\frac{1-\eta}{2}}$, where $\eta>0$ is a free parameter, then $d(\gamma)=o(p)$.*

To exemplify the tradeoff implications, consider $\frac{\gamma}{c}=\sqrt{\ln\ln p}=1.49$, $p=10^4$, and $c=1$. According to an optimal Bayes classifier this yields 5% clustering error in 1-dimension. To achieve that error $d(\gamma)\leq 9.24$ projections are sufficient to be examined, on average. On the other hand, for $\frac{\gamma}{c}=\sqrt{\ln p}=3.03$ the clustering error is essentially 0, however, the average number of projections required to achieve this error rate is $d(\gamma)\leq 10^4$.

The conditions provided in corollary 2 address the similarity between γ and c and enable us to construct novel and efficient algorithms employing remarkably small number of projections if γ is close to c up to a log-logarithmic factor in p.

2.2 The case of k-GMM ($k>2$)

We extend Theorem 1 via a union bound to the case of k Gaussians:

Theorem 2 *Consider $\mathbf{m}_1,...,\mathbf{m}_k\in\mathbb{R}^p$ and $\sigma_1,...,\sigma_k\in\mathbb{R}^+$. Assume that $\mathbf{A}=(A_1,\ldots,A_p)$ are generated i.i.d. according to $\mathcal{N}(0,1)$. Given $\gamma_{\min}>0$, and $i,j\in\{1,\ldots,k\}$ let*

$$c_{(i,j)}=\frac{\|\mathbf{m}_i-\mathbf{m}_j\|}{\sqrt{p}(\sigma_i+\sigma_j)}.$$

Let $c_{\min}\triangleq\min_{i,j}c_{(i,j)}$. Define event $\mathcal{B}$ as having separation larger than $\gamma_{\min}$ by all pairs of projected Gaussians. Thats is,

$$\mathcal{B}\triangleq\Big\{\Big|\langle\mathbf{m}_i-\mathbf{m}_j,\frac{\mathbf{A}}{\|\mathbf{A}\|}\rangle\Big|\geq\gamma_{\min}(\sigma_i+\sigma_j):\ \forall(i,j)\in\{1,\ldots,k\}^2,i\neq j\Big\}. \tag{7}$$

Then,

$$\mathrm{P}(\mathcal{B}^c)\leq \frac{k^2}{2}\Big(1-2Q\Big(\frac{\gamma_{\min}}{c_{\min}}\sqrt{\frac{1.1}{1-\frac{\gamma_{\min}^2}{c_{\min}^2 p}}}\Big)(1-\mathrm{e}^{-0.002p})\Big). \tag{8}$$

The following corollary of Theorem 2 provides a better understanding of the running time dependency of this method on the number of components k.

Corollary 4 *Consider the same setup as Theorem 2. Let $d(\gamma_{\min})$ denote the expected number of projections required to obtain separation $\gamma_{\min}$ between each pair of projected Gaussians. Then, if*

$$\gamma_{\min} \leq (1-\alpha)\sqrt{\frac{2\pi}{1.1}}\frac{c_{\min}}{k^2},$$

for some $\alpha \in (0,1)$, then $\limsup_{p\to\infty} d(\gamma_{\min}) \leq \frac{1}{\alpha}$.

2.3 Mixture of two arbitrary Gaussians

At this stage we are read to generalize the results of the previous section to arbitrary Gaussians with covariance matrices Σ_1 and Σ_2. Conditioned on $\mathbf{A}=\mathbf{a}$, projecting points $\mathbf{X}$ drawn from Gaussian distribution $\mathcal{N}(\mathbf{m}_i, \Sigma_i)$ as $\mathbf{X}^T\mathbf{a}$ are distributed as a Gaussian distribution with mean $\mathrm{E}[\langle \mathbf{X}, \mathbf{a}\rangle] = \langle \mathbf{m}_i, \mathbf{a}\rangle$, and variance $\mathrm{var}(\langle \mathbf{X}, \mathbf{a}\rangle) = \mathbf{a}^T\Sigma_i\mathbf{a}$. As argued before, the two projected clusters are γ-separated, if $|\langle \mathbf{m}_1, \mathbf{a}\rangle - \langle \mathbf{m}_2, \mathbf{a}\rangle| > \gamma(\sqrt{\mathbf{a}^T\Sigma_1\mathbf{a}} + \sqrt{\mathbf{a}^T\Sigma_2\mathbf{a}})$, or

$$|\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{a}\rangle| > \gamma\Big(\sqrt{\mathbf{a}^T\Sigma_1\mathbf{a}} + \sqrt{\mathbf{a}^T\Sigma_2\mathbf{a}}\Big), \qquad (9)$$

for some appropriate $\gamma > 0$. Unlike the condition stated in (3), both sides of (9) depend on the direction of $\mathbf{a}$. Therefore, analyzing the following probability

$$\mathrm{P}\left(|\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{A}\rangle| > \gamma(\sqrt{\mathbf{A}^T\Sigma_1\mathbf{A}} + \sqrt{\mathbf{A}^T\Sigma_2\mathbf{A}}\,)\right),$$

is more complicated. The following theorems 3 and 4 provide lower bounds on this probability for the cases of $\Sigma_1+\Sigma_2$ having a full rank $r=p$, and for the case of partial rank $r<p$, respectively.

Theorem 3 *Consider $\mathbf{m}_1, \mathbf{m}_2 \in \mathbb{R}^p$ and semi-positive definite matrices Σ_1 and Σ_2. Assume that the entries of $\mathbf{A} = (A_1, \ldots, A_p)$ are generated i.i.d. according to $\mathcal{N}(0,1)$. Let $\lambda_{\max}$ denote the maximum eigenvalue of $\Sigma_1+\Sigma_2$. Also, given $\gamma > 0$, let*

$$\beta \triangleq \frac{2\gamma^2\lambda_{\max}p}{\|\mathbf{m}_1 - \mathbf{m}_2\|^2}.$$

Then, for any $\tau > 0$, the probability that the 1-dimensional projected Gaussians using a uniformly random direction are γ-separated, i.e., $\mathrm{P}\left(|\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{A}\rangle| \geq \gamma(\sqrt{\mathbf{A}^T\Sigma_1\mathbf{A}} + \sqrt{\mathbf{A}^T\Sigma_2\mathbf{A}}\,)\right)$, can be lower-bounded by

$$Q\Big(\sqrt{\beta\frac{(1-\frac{1}{p})}{(1-\frac{\beta}{p})}(1+\tau)}\Big)(1-\mathrm{e}^{-\frac{p-1}{2}(\tau-\log(1+\tau))}). \quad (10)$$

In Theorem 4 we consider the case where the covariance matrices are not full-rank. In this case the expected number of required projections significantly decreases if the rank of $\Sigma_1+\Sigma_2$ is much smaller than p:

Theorem 4 *Consider $\mathbf{m}_1, \mathbf{m}_2 \in \mathbb{R}^p$ and semi-positive definite matrices Σ_1 and Σ_2. Assume that the entries of $\mathbf{A} = (A_1, \ldots, A_p)$ are generated i.i.d. according to $\mathcal{N}(0,1)$. Let r and $\lambda_{\max}$ denote the rank and the maximum eigenvalue of $\Sigma_1+\Sigma_2$, respectively. Also, given $\gamma > 0$, $\tau_1 \in (0,1)$ and $\tau_2 > 0$, let*

$$\beta \triangleq \frac{2(1+\tau_2)\gamma^2\lambda_{\max}r}{(1-\tau_1)\|\mathbf{m}_1 - \mathbf{m}_2\|^2}.$$

Then, for any $\tau > 0$, the probability that the 1-dimensional projected Gaussians using a uniformly random direction are γ-separated, i.e., $\mathrm{P}\left(|\langle \mathbf{m}_1 - \mathbf{m}_2, \mathbf{A}\rangle| \geq \gamma(\sqrt{\mathbf{A}^T\Sigma_1\mathbf{A}} + \sqrt{\mathbf{A}^T\Sigma_2\mathbf{A}}\,)\right)$, can be lower-bounded by

$$2Q\left(\sqrt{\beta\frac{(1-\frac{1}{p})}{(1-\frac{\beta}{p})}(1+\tau)}\right)(1-\mathrm{e}^{-\frac{p-1}{2}(\tau-\log(1+\tau))}) - \mathrm{e}^{\frac{p}{2}(\tau_1+\log(1-\tau_1))} - \mathrm{e}^{-\frac{r}{2}(\tau_2-\log(1+\tau_2))}.$$

As before, let $d(\gamma)$ denote the expected number of 1-dimensional random projections required to attain γ-separation in 1-dimension. Similar to the case of spherical Gaussians, Corollaries 5 and 6 study the number of projections required for attaining a separation γ. The proofs follow closely the proofs of Corollaries 1 and 2.

Corollary 5 *Consider two c-separated Gaussian distributions in $\mathbb{R}^p$ with means $\mathbf{m}_1, \mathbf{m}_2 \in \mathbb{R}^p$ and covariance matrices Σ_1 and Σ_2. Let $\beta \triangleq \frac{2\gamma^2\lambda_{\max}p}{\|\mathbf{m}_1-\mathbf{m}_2\|^2}$, where $\lambda_{\max}$ denotes the maximal eigenvalue of the matrix $\Sigma_1+\Sigma_2$. Then as $\lim_{p\to\infty} d(\gamma) \leq \frac{1}{2Q(\sqrt{\beta})}$.*

Corollary 6 *Consider the same setup as in Corollary 5. If γ is such that $\sqrt{\beta} = (\ln\ln p)^{\frac{1-\eta}{2}}$, where $\eta > 0$ is a free parameter, then $d(\gamma) = o(\ln p)$.*

3 Algorithm

In this section, we propose Algorithm 1 for clustering arbitrary mixtures of Gaussian distributions. The algorithm receives as input the $n \times p$ data matrix X, a prescribed error - e, and a maximum number of 1-dimensional projections - M. M can be estimated, for example, as $o(\ln p)$ before execution based on corollaries 2, or 6. The algorithm sequentially performs 1-dimensional projections, where each projection's direction is chosen uniformly at random. After each random projection, the parameters of the projected mixture of Gaussians in 1-dimension and its corresponding clustering error are estimated with the MoM algorithm of [Hardt(2015)] or EM [Dempster et al.(1977)]. This

Algorithm 1: ClusterGMM

Data: $X - n \times p$ data matrix, e - error, M - projection budget
Result: $\mathcal{C}^*$
initialization: $i = 1$ $\hat{e} = \infty$
while $i < M$ **do**
  Project to random direction: $\langle X, A^i \rangle$
  Learn 1-dimensional parameters: $(\hat{m}_1^i, \hat{m}_2^i, \hat{\sigma}_1^i, \hat{\sigma}_2^i, \hat{w}_1^i)$
  Learn a separator $\mathcal{C}^*$ and compute $\hat{e}$
  if $\hat{e} < e$ **then**
    return($\mathcal{C}^*$)
  if *Mixture is spherical* **then**
    Estimate the necessary number of projections - $\bar{M}$ (using Lemma 2 and Thm. 1)
    if $M < \bar{M}$ **then**
      print("Error not Achievable")
      EXIT;
print("Error not Achievable")

process is iterated until either the desired accuracy e is achieved by the current projection, or the maximum number of projections M is reached.

For the spherical case, one can use on-the-fly Lemma 2 to estimate c at the iteration i as

$$\bar{c} = \sqrt{\frac{1}{i}\sum_{j=1}^{i} \hat{\gamma}_j^2}, \tag{11}$$

where $\hat{\gamma}_j$ is the estimated 1-dimensional separation from projection $\boldsymbol{A^j}$. Once c is estimated via (11) one can update the number of projections to achieve $\gamma = Q^{-1}(e)$ via Theorem 1 and its Corollary 2. If the required number of projections (compute based on $\bar{c}$) is larger than the budget M the algorithm can be stopped.

Note that Alg. 3.3 of [Hardt(2015)] involves computing the 6 moments of the 1-dimensional projected sample and finding the roots of a low degree polynomial. Hence, the parameters estimation step comprises of linear running time complexity in the sample size. Alternatively, using EM [Dempster et al.(1977)] with its linear running time for each step comprises overall linear running time for a bounded number of iterations.

We provide numerical experiments in section 5.

4 Sample complexity

In this section, we study the sample complexity of our proposed algorithm. Note that in our algorithm, parameter estimation is only done after 1-dimensional projections, and hence in $\mathbb{R}$. After each random projection, we use Algorithm 3.3 of [Hardt(2015)] to estimate the parameters of the projected mixture of two Gaussian distributions. Algorithm 3.3 is a variation of the well-known method of moments proposed by Pearson in [Pearson(1894)]. The following result from [Hardt(2015)] summarizes the performance of Algorithm 3.3 in estimating the parameters of a mixture of two general Gaussians in 1-dimension.

Theorem 5 (Theorem 3.10 in [Hardt(2015)]) *Consider a mixture of two Gaussian distribution* $w\mathcal{N}(\mu_1, \sigma_1) + (1-w)\mathcal{N}(\mu_2, \sigma_2)$. *Let* $\sigma^2 = w(1-w)(\mu_1 - \mu_2)^2 + w\sigma_1^2 + (1-w)\sigma_2^2$ *denote the variance of this distribution. Then, given* $n = O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$ *samples, Algorithm 3.3, with probability* $1-\delta$, *returns estimates of the parameters as* $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_1, \hat{w})$, *which under the right permutation of the indices, satisfy the following guarantees, for* $i = 1, 2$,

- *If* $n \geq \left(\frac{\sigma^2}{|\mu_1 - \mu_2|^2}\right)^6$, *then* $|\mu_i - \hat{\mu}_i| \leq \epsilon|\mu_1 - \mu_2|$, $|\sigma_i^2 - \hat{\sigma}_i^2| \leq \epsilon|\mu_1 - \mu_2|^2$, *and* $|w - \hat{w}| \leq \epsilon$.
- *If* $n \geq \left(\frac{\sigma^2}{|\sigma_1^2 - \sigma_2^2|}\right)^6$, *then* $|\sigma_i^2 - \hat{\sigma}_i^2| \leq \epsilon|\sigma_1^2 - \sigma_2^2| + |\mu_1 - \mu_2|^2$, *and* $|w - \hat{w}| \leq \epsilon + \frac{|\mu_1 - \mu_2|^2}{|\sigma_1^2 - \sigma_2^2|}$.
- *For any* $n \geq 1$, *the algorithm performs as well as assuming the mixture is a single Gaussian, and* $|\mu_i - \hat{\mu}_i| \leq |\mu_1 - \mu_2| + \epsilon\sigma$, *and* $|\sigma_i^2 - \hat{\sigma}_i^2| \leq |\mu_1 - \mu_2|^2 + |\sigma_1^2 - \sigma_2^2| + \epsilon\sigma^2$.

Consider a mixture of two c-separated Gaussians in $\mathbb{R}^p$, and assume that γ denotes the separation required for the projected Gaussians in $\mathbb{R}$ to achieve the desired error e. In such a setting, to analyze the sample complexity of our proposed methods, we need to show that, there exists γ' smaller than γ, such that

i) if after a random projection, the two projected Gaussian are γ''-separated, where $\gamma'' > \gamma'$, then the number of samples is such that, with high probability, we are able to estimate the separation γ'' accurately,

ii) if the two projected Gaussians are γ''-separated, where $\gamma'' \leq \gamma'$, then, with high probability, we are able to reject that direction.

Condition i) guarantees that the direction with separation γ, if it exists, will be detected, and Condition ii) ensures that there are no false detections, where a direction with low-separation is misidentified as a good direction. In the following, we derive the sample complexity required to satisfy each condition.

First, to address Condition i) the following corollary 7, a direct result of Theorem 5, shows that, if the two components of a Gaussian mixture model are separated enough in 1-dimension, given sufficient number of samples, Algorithm 3.3 of [Hardt(2015)] returns accurate estimates of *all* parameters. Then, Theorem 6 connects the error in estimating the parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, w)$ to the error in estimating the clustering error. Since the ultimate goal of our algorithm is to find a direction which yields a desired clustering error, it is important to establish such a connection, which, given the desired clustering error, characterizes some sufficient accuracy in estimating the parameters.

Corollary 7 *Let $(X_1, \ldots, X_n)$ denote n i.i.d. samples of a mixture of two c-separated Gaussians $w\mathcal{N}(\mu_1, \sigma_1) + (1-w)\mathcal{N}(\mu_2, \sigma_2)$, where $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$. Further assume that the separation $c = |\mu_1 - \mu_2|/(\sigma_1 + \sigma_2)$ in 1-dimension is larger than $\gamma_{\min}$. Let $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{w})$ denote the estimates of $(\mu_1, \mu_2, \sigma_1, \sigma_2, w)$ returned by Algorithm 3.3 of [Hardt(2015)]. Then, if $n = O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$ and $n \geq \frac{1}{(2\gamma_{\min})^{12}}$, then $|\mu_i - \hat{\mu}_i| \leq \epsilon|\mu_1 - \mu_2|$, $|\sigma_i^2 - \hat{\sigma}_i^2| \leq \epsilon|\mu_1 - \mu_2|^2$, and $|w - \hat{w}| \leq \epsilon$.*

Theorem 6 *Consider $(X_1, \ldots, X_n)$ that are generated i.i.d. according to a mixture of two γ-separated Gaussians $w\mathcal{N}(\mu_1, \sigma_1) + (1-w)\mathcal{N}(\mu_2, \sigma_2)$, where $\sigma_1 = \sigma_2$, $w \in [w_{\min}, 0.5]$, $\mu_1 < \mu_2$ and $\gamma \in [\gamma_{\min}, \gamma_{\max}]$. Let $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{w})$ denote the estimate of the unknown parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, w)$. Let e_{opt} and $\hat{e}$ denote the minimum achievable clustering error and a clustering error based on the estimated parameters, respectively. Then, if $|\mu_i - \hat{\mu}_i| \leq \epsilon|\mu_1 - \mu_2|$, $|\sigma_i^2 - \hat{\sigma}_i^2| \leq \epsilon|\mu_1 - \mu_2|^2$, $|w - \hat{w}| \leq \epsilon$, and $(16\gamma_{\max}^2 + 8\gamma_{\max} \ln \frac{1-w_{\min}}{w_{\min}} + 2\gamma_{\max}\epsilon)\epsilon < \frac{1}{2}$, then*

$$|\hat{e} - e_{\text{opt}}| \leq \Big(2\gamma + \frac{1}{w_{\min}\gamma} + \left(\frac{1}{\gamma} + 2\gamma\right) \ln \frac{1-w_{\min}}{w_{\min}} + \frac{8\gamma_{\max}^2}{\gamma} + 2\gamma \left(4\gamma + 2 \ln \frac{1-w_{\min}}{w_{\min}}\right)^2\Big)\epsilon + Q\left(\frac{1}{4\gamma\epsilon} + \epsilon_1\right) + \epsilon_2,$$

where $\epsilon_1 = o(1/\epsilon)$ and $\epsilon_2 = o(\epsilon)$.

Note that, as expected, as $\gamma_{\min}$ converges to zero, by Corollary 7, the required number of samples for accurate estimation of the parameters grows to infinity. On the other hand, too small separation γ corresponds to large overlap of the two Gaussians. Hence, to establish condition ii) we later describe a procedure to discard directions with low separation. As confirmed in the following lemma 3, unless the weights of the two Gaussians are very non-uniform, i.e. $\min(w, 1-w)$ is far from 0.5, low separation corresponds to high clustering error.

Lemma 3 *Consider i.i.d. points generated as $w\mathcal{N}(\mu_1, \sigma) + (1-w)\mathcal{N}(\mu_2, \sigma)$. Without loss of generality, assume that $\mu_1 \leq \mu_2$ and $w < 0.5$. Let $\gamma = (\mu_2 - \mu_1)/(2\sigma)$. Also, let e_{opt} denote the error probability of an optimal Bayesian classifier. Then, if $w \leq 0.1$,*

$$e_{\text{opt}} \geq wQ\left(-\frac{1}{\gamma} + \gamma\right). \tag{12}$$

For $w \in (0.1, 0.5]$,

$$e_{\text{opt}} \geq wQ\left(\gamma\right). \tag{13}$$

Therefore, if the ultimate goal is to achieve a reasonable clustering error through multiple random projections, for those directions with too small separation, we only need to identify them and discard them. In other words, for such directions, it is not necessary to estimate all the parameters of the projected Gaussians accurately, as they ultimately are not going to be used for clustering. The following lemma provides a mechanism for identifying and discarding all directions that have a separation smaller than some threshold.

Lemma 4 *Let $(X_1, \ldots, X_n)$ denote n i.i.d. samples of a mixture of two γ-separation Gaussians $w\mathcal{N}(\mu_1, \sigma_1) + (1-w)\mathcal{N}(\mu_2, \sigma_2)$, where $\sigma_1 = \sigma_2$, $\gamma = (\mu_2 - \mu_1)/(\sigma_1 + \sigma_2) < 1/2$ and $\mu_1 < \mu_2$. Let $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{w})$ denote the estimates of $(\mu_1, \mu_2, \sigma, \sigma, w)$ returned by Algorithm 3.3 of [Hardt(2015)]. Then, if $n = O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$, with probability larger than $1 - \delta$,*

$$\frac{|\hat{\mu}_1 - \hat{\mu}_2|}{\hat{\sigma}_1 + \hat{\sigma}_2} \leq \frac{3\gamma + \epsilon}{1 - 2\sqrt{\gamma^2 + \epsilon}}.$$

To shed more light on the implications of Lemma 4, consider, for example, a mixture of two 1-dimensional Gaussians with equal variance and separation γ smaller than $\frac{1}{8}$. Then, given $n = O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$ i.i.d. samples, with probability larger than $1 - \delta$, the estimated separation (using parameters derived from Algorithm 3.3 of [Hardt(2015)]) is smaller than

$$\frac{\frac{3}{8} + \epsilon}{1 - 2\sqrt{(\frac{1}{8})^2 + \epsilon}} = \frac{1}{2} + o(\epsilon).$$

Therefore, if after performing each random projection, we estimate the parameters of the two Gaussians using Algorithm 3.3 of [Hardt(2015)] and then estimate their separation as $\frac{|\hat{\mu}_1 - \hat{\mu}_2|}{\hat{\sigma}_1 + \hat{\sigma}_2}$ and discard all those directions that have estimated separation smaller than $\frac{1}{2}$, we would, with high probability, discard all directions

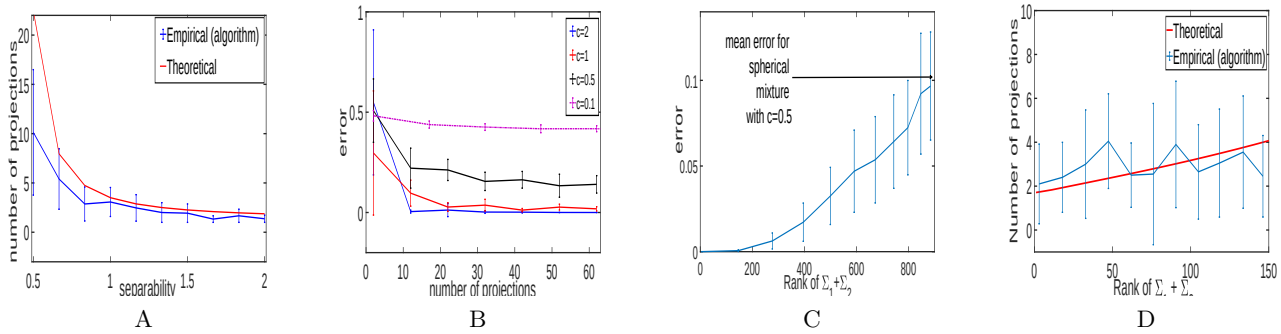

Figure 1: **Spherical Gaussians**: Data contains 10K points realization for a mixture of two Gaussians in $\mathbb{R}^p$, $p = 100$. A - Projection order vs. separation to reach 20% error. Algorithm performance compared with the theoretical upper-bound of Theorem 1. B - number of projections vs. accuracy for separation values 0.1, 0.5, 1, and 2. **Non-Spherical Gaussians**: Data contains 10K points realization for a mixture of two Gaussians in $\mathbb{R}^p$, $p = 1000$. C - Error vs. rank of $(\Sigma_1 + \Sigma_2)$. D - number of projections vs. rank compared with Theorem 4 bound.

with a separation smaller than 1/8. Therefore, if the desired clustering error is smaller than $Q(\frac{1}{2})$, then this procedure discards directions that have no chance of yielding the required performance. For directions with separation larger than 1/8, we need to have enough samples to estimate the parameters accurately. The required number of samples for achieving this goal is shown in Corollary 7, which follows directly from Theorem 3.10 of [Hardt(2015)]. Note that using this procedure, directions with estimated separation smaller than 0.5 include those directions with separation in $(\frac{1}{8}, \frac{1}{2})$, for which, with high probability, we have estimated the parameters accurately, and those directions with separation smaller than $\frac{1}{8}$, for which we have only a crude estimate of the parameters.

5 Experiments

Spherical Gaussians: Number of projections vs. separation. We generate 10K points in $\mathbb{R}^{100}$, with $w_1 = w_2$ and $\sigma_1 = \sigma_2$. The user's desired error is fixed at $e = 20\%$ as we measure the number of projections used until the error is achieved. Fig. 1-A plots the number of projections scanned until the prescribed accuracy is attained for various c values. We also use the lower bound provided by Theorem 1 to plot the inverse of the probability bound defined there corresponding to the expected number of projections to achieve e. The tightness of the bound is clearly observed.

Spherical Gaussians: Error vs. number of projections. Fig. 1-B reports the accuracy values vs. number of projections for varying c values. The experiment marks the necessary number of projections to achieve the minimal possible error. The curves demonstrate the high efficiency in which the algorithm can cluster the data to a prescribed error that corresponds to the high dimensional separation.

Non-spherical Gaussians: Accuracy vs. rank. Using Algorithm 1, we examine the error as function of the rank of the covariance summation matrix $(\Sigma_1 + \Sigma_2)$. Fig. 1-C demonstrates this error. We note that as the matrix approaches the full rank (with equal variance in the populated dimensions) the error of our algorithm approaches the error 0.1 attained for spherical Gaussians at $c = 0.5$.

Non-spherical Gaussians: Number of projections vs. rank. In Fig. 1-D we report results for a 4% error prescribed, as we examine the number of projections required by our algorithm vs. the rank of $(\Sigma_1 + \Sigma_2)$, and the bound provided by Theorem 4. The separation in this case is 0.5.

We note that our bounds are on the *expected* number of projections. Moreover, slight deviations in the number of projections from the bound may also occur due to precision in algorithm convergence (EM or MoM).

6 Conclusion

In this paper we study the problem of clustering arbitrary GMMs using independent and uniformly at random one-dimensional projections. To achieve this goal, we derived bounds on the number of independent random 1-dimensional projections required to achieve a desired clustering error e in $\mathbb{R}$, for both spherical and non-spherical GMMs. Our bounds show that for sufficiently-separated high-dimensional Gaussians, e can be achieved in linear running time in both the dimension of the data and its sample size. Moreover, our sample complexity is independent of the original data dimension p. Our analysis also provides a mechanism that allows discarding the directions in which the mixture parameters cannot be estimated accurately. Finally, we also studied the case of $k > 2$ Gaussians.

References

M. Belkin and K. Sinha. Toward learning gaussian mixtures with arbitrary separation. In *Proc. Ann. Conf. on Learn. Theory (COLT)*, pages 407–419. Citeseer, 2010.

M. Belkin and K. Sinha. Polynomial learning of distribution families. *SIAM J. on Comp.*, 44(4):889–911, 2015.

C. Boutsidis, A. Zouzias, and P. Drineas. Random projections for k-means clustering. In *Adv. in Neu. Inf. Proc. Sys.*, pages 298–306, 2010.

S. Dasgupta. Learning mixtures of Gaussians. In *40th Annual Symp. on Found. of Comp. Sci.*, pages 634–644. IEEE, 1999.

S. Dasgupta. Experiments with random projection. In *Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2000.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. of the Roy. Stat. Soc.. Series B (meth.)*, pages 1–38, 1977.

X. Z. Fern and C. E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proc. of the 20th Int. Conf. on Mach. Learn. (ICML)*, pages 186–193, 2003.

M. Hardt and E. Price. Tight bounds for learning a mixture of two gaussians. In *Proc. of the 47th Ann. ACM Sym. on Theory of Comp.*, pages 753–760. ACM, 2015.

J. Huggins. Provably learning mixtures of gaussians and more. Technical report, Technical report, Columbia Uni., 2011.

R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Int. Conf. on Comp. Learn. Theory*, pages 444–457. Springer, 2005.

S. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, 28(2):129–137, 1982.

K. Pearson. Contributions to the mathematical theory of evolution. *Phil. Trans. of the Roy. Soc. of London. A*, 185:71–110, 1894.

K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323–2326, 2000.

L. Van Der Maaten, E. Postma, and J. Van den Herik. Dimensionality reduction: a comparative. *J. Mach. Learn. Res*, 10:66–71, 2009.

S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *J. of Comp. and Sys. Sci.*, 68(4):841–860, 2004.