

Robust PCA and Clustering on Noisy Mixtures

S. Charles Brubaker

Abstract

This paper presents a polynomial algorithm for learning mixtures of logconcave distributions in \mathbb{R}^n in the presence of malicious noise. That is, each sample is corrupted with some small probability, being replaced by a point about which we can make no assumptions. A key element of the algorithm is Robust Principle Components Analysis (PCA), which is less susceptible to corruption by noisy points. While noise may cause standard PCA to collapse well-separated mixture components so that they are indistinguishable, Robust PCA preserves the distance between some of the components, making a partition possible. It then recurses on each half of the mixture until every component is isolated. The success of this algorithm requires only a $O^*(\log n)$ factor increase in the required separation between components of the mixture compared to the noiseless case.

1 Introduction

Clustering points in \mathbb{R}^n is common practice in a large variety of applied fields such as computer vision, robotics, speech recognition, web search, and spam filtering. Heuristics such as “k-means clustering” [17, 11] and “expectation maximization” [7] have been used in these fields since such computation became practical. Over the past few years, however, algorithms with theoretical guarantees have been discovered, where it is assumed that the points come from a mixture distribution, i.e. a convex combination of distributions of a known type. Mixtures of Gaussians or logconcave distributions and mixtures of product distributions have received the most attention. The goal of clustering is to group points generated by the same component together and separate points generated by different components.

This paper explores the case where the data includes some small miscellaneous component in addition to a well-behaved mixture. Equivalently, we may say that the sampling process for the well-behaved mixture has some noise, whereby with some small probability a point is replaced by a noise point about which can make no assumptions. The practical importance of robustness to the presence of noise should be apparent to anyone who has tried to file his bills, organize a closet, or set up a directory structure on his hard disk. Some things just don’t belong to any large category. The presence of these “noisy” objects does not usually impede our ability to cluster or classify objects, suggesting that we should hold our algorithms to this standard as well. More concretely, in tasks such as document or web-page clustering it is unreasonable to assume that components will be well-separated with *absolutely nothing* in-between.

Previous work has addressed noiseless version of the clustering problem for mixtures of logconcave distributions [13, 1]. Recall that a mixture density is a convex combination of distributions of a known type. It has the functional form

$$F(x) = w_1 f_1(x) + \dots + w_k f_k(x),$$

where f_i is from a special class of distributions, e.g. Gaussians. The coefficients w_i are called the mixing weights and sum to one. Sampling from a mixture distribution can be thought of as a two step process. First, a component index i is chosen according to the mixing weights. Second, a point x is sampled according to the distribution f_i . Given a collection of points sampled in this fashion, clustering is the task of partitioning the points according to the component that generated them. The set of points drawn from a particular component is called a cluster. The case where the dimension of the points n is much larger than the number of components k is the most interesting for both theoreticians and practitioners.

What distinguishes this paper from previous work is that it assumes that some of the sample points may have been corrupted in an arbitrary way. That is, with probability at most ϵ the sample source outputs a point about which we can make no assumptions. We call such a sample source ϵ -noisy. This is the natural

analog to the malicious error models of [14, 21] for the clustering problem. Because the noise component is arbitrary, it may not be possible to cluster in the traditional sense. Indeed all noise points could be identical to one of the non-noise points, making them indistinguishable. Therefore, we set a different goal. Suppose the data set can be written as the disjoint union $S_1 \cup \dots \cup S_k \cup N$, where S_i corresponds to the set of points from component i and N to the set of noise points. Then we seek a collection of disjoint sets $C_1 \dots C_k$ such that for every S_i , there is a unique C_i where

$$S_i \subseteq C_i \subseteq S_i \cup N. \quad (1)$$

Although the sets $\{C_i\}$ may include some noise points, they induce a correct partition of the non-noise points $\{S_i\}$.

A common approach for classifying points from mixture models is to project to the top k principal components [13, 1, 22]. This fails in the presence of noise. In fact, only k well-chosen malicious noise points are required to cause the intermean distances to become arbitrarily small after such a projection. For instance let v_{n-k+1}, \dots, v_n be the k smallest spectral components for a set of samples S . To this set add k noise points, $x_1 = cv_n, \dots, x_k = cv_{n-k+1}$. For large values of c , the largest k spectral components of $S \cup \{x_i\}_{i=1}^k$ will converge to $v_{n-k+1} \dots v_n$.

Methods for clustering points after projection fare little better amidst malicious noise. For instance, the single link method (used for example in Achlioptas and McSherry [1]) fails, as a large links in the graph can be subdivided through the insertion of noise points.

1.1 Results

We present a polynomial time algorithm that given a noisy mixture of well-separated, logconcave distributions in \mathbb{R}^n , learns to separate the components of the mixture. That is, the algorithm finds a partition of \mathbb{R}^n into k sets with disjoint interiors each of which contains almost all of the probability mass of a unique component of the mixture. The error of such a partition is the total mass that falls outside of the correct set. As a corollary, this algorithm makes it possible to cluster points from a noisy source in the sense of Eqn. 1. The separation between the means necessary for the algorithm's success is only an $O^*(\log n)$ factor larger than the best analogous results without noise, treating k and w_{\min} as constants.

The input to our algorithm is a source of samples LM, a natural number k and a scalar w_{\min} . The quantity k is the number of non-noise components in the mixture and w_{\min} is a lower bound on the minimum mixing weight. For simplicity of the exposition, we state the results in terms of learning a partition of \mathbb{R}^n (i.e. a classifier). This can easily be turned into a statement strictly about clustering a set of points through a slight modification of the algorithm.¹

For a component i of the mixture, we define the following quantities : $\mu_i = E[x]$, $R_i^2 = E[\|x - \mu_i\|^2]$, and $\sigma_i^2 = \max_{v \in \mathbb{R}^n} E[(v \cdot (x - \mu_i))^2] / \|v\|^2$. We define the mixing weights w_i to be the probability that LM outputs a sample from component i . Thus, if LM outputs noise with probability ξ , then $\sum_{i=1}^k w_i = 1 - \xi$, effectively treating ξ as a mixing weight itself. We let w_{\min} be the minimum mixing weight.

Our main result is summarized in the following theorem.

Theorem 1. *Let \mathcal{F} be a mixture of k logconcave distributions with means $\{\mu_i\}$ and maximum variances $\{\sigma_i\}$. Let $\delta, \eta > 0$. There exist $\epsilon = \Omega(w_{\min} \log^{-2}(nk/(w_{\min}\delta\eta)))$ and $\alpha = O(k^{3/2}w_{\min}^{-1} \log(nk/(w_{\min}\delta\eta)))$ such that if LM is an ϵ -noisy sample source for \mathcal{F} and if for every pair of components i, j*

$$\|\mu_i - \mu_j\| \geq \alpha(\sigma_i + \sigma_j), \quad (2)$$

then the following holds. There is a polynomial algorithm that given access to at least $O(nkw_{\min}^{-1} \log^6(nk/\delta))$ samples from LM with probability $1 - \delta$ returns a partition of \mathbb{R}^n that correctly classifies a point from \mathcal{F} with probability $1 - \eta$.

¹For instance, we might divide the given points into overlapping blocks and cluster each block using the remaining points to simulate the sample source. The overlap of the blocks can be used to calculate to appropriate permutation of component indices for each block and thus obtain a clustering of the whole set.

1.2 Previous Work

The problem of learning mixture models was popularized by S. Dasgupta in [6], beginning with mixtures of spherical Gaussians. This result was later improved upon by S. Dasgupta and Shulman [19] and by Arora and Kannan [18] to require a separation of

$$\|\mu_i - \mu_j\| \geq O(n^{1/4})(\sigma_i + \sigma_j).$$

The latter work also handles a wider variety of cases, including the case where one Gaussian is “inside” another because of a large difference in their variances.

Vempala and Wang [22] showed that spectral projection preserves the intermean distances while dramatically reducing the intra-cluster distances. Their algorithm requires only

$$\|\mu_i - \mu_j\| \geq O^*(k^{1/4})(\sigma_i + \sigma_j)$$

separation for k components. This work was later extended in by Kannan, Salmasian, and Vempala [13] to mixtures of logconcave densities that are not necessarily isotropic. Achlioptas and McSherry [1] show similar results and explore the minimum necessary separation for clustering to be possible. In these works, the required separation is

$$\|\mu_i - \mu_j\| \geq O^*(k^{3/2} + w_{\min}^{-1/2})(\sigma_i + \sigma_j).$$

All of the above separation bounds are given in terms of σ_i^2 , the maximum directional variance. This is far stronger than what is required for separability, as the components may be separated along a direction where the variance are much smaller than σ_i^2 . Brubaker and Vempala [2] address this issue via an affine-invariant clustering algorithm. For two Gaussian components, the separation condition is nearly optimal for hyperplane separability, requiring only that there be *some* direction along which the distance between the means is large compared to the variance in this direction. For more than two components, the separation condition is stated in terms of the Fisher discriminant.

A related area of work is on learning product distributions, where the coordinates are independent (e.g. a Gaussian would be axis-aligned). Here the goal is not necessarily to cluster data but to approximate the density of the mixture. Freund and Mansour [10] first solved this problem for a mixture of two distributions of binary vectors, finding a model that approximates the true distribution in terms of Kullback-Leibler distance. Feldman and O’Donnell [8] extended this result to mixtures of any constant number of components and to discrete domains instead of binary vectors, i.e. $\{0, \dots, b-1\}^n$ instead of $\{0, 1\}^n$. Joined by Servedio in [9], they applied their technique to mixtures of a constant number of axis-aligned Gaussians, showing that they can be approximated without any separation assumption at all.

Another class of results on learning product distributions uses separation conditions which assume that the component centers be separated along many directions. Chaudhuri and Rao [4] note that results such as [13],[1] and [2] have a polynomial dependence on the inverse of the minimum mixing weight and reduce this to a logarithmic dependence by exploiting the independence of the coordinates. Beyond logconcave distributions, A. Dasgupta et al [5] consider a class of heavy-tailed product distributions and give separation conditions under which such distributions can be learned using an algorithm that is exponential in the number of samples. Chaudhuri and Rao [3] have recently given a polynomial algorithm to learn a related class of heavy-tailed product distributions.

The term “Robust PCA” has been used in previous work to describe several proposed algorithms in the fields of robust statistics, computer vision [15, 23] and bioinformatics [12]. Like our algorithm, these works attempt to find the PCA subspace in the presence of outliers. None, however, use the idea of alternating removing outliers and projecting to a lower dimensional subspace, as our algorithm does.

2 Algorithm

In previous work [13], the approach is to first project the data onto its top k spectral components, then extract a single cluster, and repeat. This strategy succeeds because the projection onto the top k components

preserves much of the intermean distances, while reducing the pairwise distance between points of the same component. The concentration of the pairwise distances is then exploited to remove a component.

In the presence of noise, however, this approach breaks down. In fact, only k well-chosen noise points are required to cause the intermean distances to become arbitrarily small after projection. To cope with this problem we first remove outliers. That is, we reduce the maximum distance between any two points to be $O(R_{\max} + \mu_{\max})$. Projection to the top k components still may not preserve the necessary intermean distances, but projection to the top $\lfloor (n - k)/2 \rfloor + k$ components will. By repeating this procedure of first removing outliers and then projecting, we reduce the dimension to k . We call this procedure Robust PCA.

Robust PCA will preserve enough of the distance between the components whose means are furthest apart so that the direction between their means can be approximated by a pair of samples, one coming from each component. Imagine projecting the entire mixture density onto this line. The concentration of the individual components implies that this density will be multimodal with large peaks and long flat valleys. By setting a threshold in the middle of a valley, we define a hyperplane that separates the components of the mixture.

We then recurse on the two half-spaces defined by this hyperplane. At lower levels of the recursion tree, we are recursing on the intersection of these hyperplanes, i.e. a polyhedron. Ideally, we would like to recurse on a submixture, i.e. a subset of the original mixture's components. Fortunately, each component is far enough away from the the support hyperplanes that the probability that a sample will appear on one side of a hyperplane while its component mean is on the other is vanishingly small. This enables us to simulate the desired submixture by rejecting samples until one from the correct part of \mathbb{R}^n is obtained.

2.1 Robust PCA

We now describe our algorithm in more detail, considering Robust PCA first. We initialize the subspace W to be \mathbb{R}^n and then reduce the dimension of this subspace in each iteration until it has only k dimensions. Each iteration considers a fresh set of samples in step 2a to avoid dependency issues in the analysis. These points are projected to the current subspace W . In steps 2b and 2c, we compute

$$t(Z) = s_r(\{s_r(\{\|p - q\| : q \in Z\}) : p \in Z\}) \quad (3)$$

where s_r is the function which selects the r th largest element of a set. This computation is accomplished by arranging all pairwise distances into a matrix indexed by p and q . For every row, we select the r th largest element (step 2b), and from the resulting set we then select the r th largest element (step 2c). This gives a pair p_0, q_0 such that $\|p_0 - q_0\| = t(Z)$.

Next, we remove all points that are further than $\xi t = 16\beta t$ from p_0 to form the set Z' and compute the matrix

$$\sum_{p \in Z'} (p - p_0)(p - p_0)^T.$$

We then compute the SVD of this matrix and let the top $\ell = \lfloor (\dim(W) - k)/2 \rfloor + k$ singular vectors be the span of W . This process is repeated with fresh samples until $\dim(W) = k$.

2.2 Clustering Noisy Mixtures

We now describe the algorithm for clustering noisy mixtures. For an input polyhedron P the goal is to separate the components whose means are contained in the polyhedron. We call this subset of components a submixture and to obtain samples from it we get points from the full mixture source LM and then intersect these points with the polyhedron P . This is done in steps 2 and 3. The analysis will show that this effectively simulates sampling from the submixture.

The set of samples is used to call Robust PCA to find a subspace W of dimension k . Another set of points X is sampled in step 3 and used to compute the resolution d in step 4, which serves as the bucket diameter as we search for a partition which does not cut the mixture components. We then obtain an independent sample Y that includes at least one point from every component with high probability. For every pair of

Algorithm 1 Robust PCA

Input:

- 1) Collection $\{Z_i\}$ of $\lceil \log_2 n \rceil$ sets of points in \mathbb{R}^n .
- 2) Integers k, r , scalar ξ .

Output: A subspace W of dimension k .

1. Let $W = \mathbb{R}^n$.
 2. While $\dim(W) > k$,
 - (a) Let $Z = \text{proj}_W(Z_i)$, where Z_i is the next set of samples.
 - (b) For every $p \in Z$ find the point $q(p)$, defined to be r th furthest away point.
 - (c) Find the point p_0 such that the distance $\|p_0 - q(p_0)\|$ is the r th largest distance in the set $\{\|p - q(p)\| : p \in Z\}$. Let $q_0 = q(p_0)$ and let $t(Z) = \|p_0 - q_0\|$.
 - (d) Let $Z' = Z \cap B(p_0, \xi t(Z))$.
 - (e) Let \bar{W} be the span of the top $\lfloor (\dim(W) - k)/2 \rfloor + k$ eigenvectors of the matrix $\sum_{p \in Z'} (p - p_0)(p - p_0)^T$.
-

points (a, b) , we then attempt to separate the components along the direction $v = (a - b)/\|a - b\|$. To separate along a direction we project a set of points to v and place them in buckets of size d , which divide the real line. If there are two full buckets with at least $w_{\min} m_X/4$ and a nearly empty bucket with at most $2\epsilon m_X$ points between, then we call this set of buckets a valley. We place the dividing threshold γ in the middle of the nearly empty bucket in the valley. We then recurse on the two polyhedra $P \cap H_{v, \gamma}$ and $P \cap H_{-v, -\gamma}$, where $H_{v, \gamma}$ denotes the halfspace $\{x \in \mathbb{R}^n : v \cdot x \geq \gamma\}$.

If no valley is found in all of $Y \times Y$, then we conclude that a single component has already been isolated and simply return the current polyhedron. The algorithm finally returns all such polyhedra.

3 Preliminaries

In our analysis, we will decompose a set Z obtained from LM into $S \cup N$ where S consists of the points drawn from \mathcal{F} and N consists of the noise points. Further, we decompose the set S into $S_1 \cup \dots \cup S_k$, where S_i consists of the points drawn from component i . For a point $p \in S$, we use $\ell(p)$ to denote the component from which p was drawn. We also use $\hat{\mu}_i$ to indicate the average of points from component i in a set. For a subspace W and polyhedron P , it will be convenient to define the following quantities. Let $I_P = \{i : \mu_i \in P\}$ and let \mathcal{F}_P be the submixture consisting of the components in I_P . Let

$$\begin{aligned} R_i^{(W)} &= E_i [\|\text{proj}_W(x - \mu_i)\|^2]^{1/2} \\ R_{\max}^{(W, P)} &= \max_{i \in I_P} R_i^{(W)}. \\ \mu_{\max}^{(W, P)} &= \max_{i, j \in I_P} \|\text{proj}_W(\mu_i - \mu_j)\| \\ \sigma_{\max}^{(P)} &= \max_{i \in I_P} \sigma_i. \end{aligned}$$

Note that E_i denotes an expectation with respect to the i th component of the mixture. When the superscript W is omitted, it may be assumed that \mathbb{R}^n is meant. The polyhedron P is often clear from context and may be omitted as well.

Throughout the analysis we use the fact that the lower bound on the separation $\alpha = \Theta(k^{3/2} w_{\min}^{-1} \log(nk/(w_{\min} \delta \eta)))$ and the upper bound on the noise $\epsilon = \Theta(w_{\min} \log^{-2}(nk/(w_{\min} \delta \eta)))$.

Algorithm 2 Cluster Noisy logconcave Mixture

Input:

- 1) Sampling source LM which generates point in \mathbb{R}^n .
- 2) Integer k , reals ϵ, w_{\min} .
- 3) Polyhedron P . (Note $P = \mathbb{R}^n$ in the initial call.)

Output: A collection of k polyhedra.

1. For $i = 1$ to $\lceil \log n \rceil$ let Z_i a set m_Z points from LM.
 2. Let W be the subspace returned by Robust PCA for the collection $\{Z_i \cap P\}$, $\xi = 16\beta$ and $r = \lfloor 2\epsilon m_Z \rfloor$.
 3. Let $X = \text{proj}_W(X_0 \cap P)$, where X_0 is a set of m_X samples obtained from LM.
 4. Let $d = t(X)/10k$.
 5. Let $Y = \text{proj}_W(Y_0)$, where Y_0 is a set of m_Y samples from LM.
 6. For every $(a, b) \in Y \times Y$
 - (a) Let $v = (a - b)/\|a - b\|$.
 - (b) Let $b_i = |\{x \in X : \text{proj}_v(x) \in [id, (i + 1)d]\}|$.
 - (c) If there is a triple $i_1 < i_2 < i_3$ where $b_{i_1}, b_{i_3} > w_{\min} m_X/4$ and $b_{i_2} \leq 2\epsilon m_X$, then let $\gamma = (i_2 + 1/2)d$ and recurse with $P = P \cap H_{v, \gamma}$ and $P = P \cap H_{-v, -\gamma}$. Return the collection of polyhedra produced by these calls.
 7. Return P .
-

3.1 Safe Polyhedra

The success of the algorithm depends on the fact that intersecting the sample set from LM with the polyhedron P in steps 2 and 3, of Algorithm 2 effectively simulates sampling from the submixture \mathcal{F}_P . That is, this intersection has the effect of including all points from components in I_P and excluding all points from other components. This motivates the following definition.

Definition 1. A polyhedron P is η -safe for a mixture \mathcal{F} if

1. For every $i \in I_P$, we have $\Pr[x \notin P] \leq \eta$, where x is a random point from component i .
2. For every $i \notin I_P$, we have $\Pr[x \in P] \leq \eta$, where x is a random point from component i .

The concentration of logconcave distributions yields a simple criterion for showing that a halfspace is safe. We use the following theorem from [16].

Theorem 2. Let $R^2 = \max_{\|v\|=1} E[(v \cdot (x - \mu))^2]$ for a random variable x from a logconcave distribution. Then

$$\Pr(\|x - \mu\| > tR) < e^{-t+1}.$$

Restricting this to a single dimension gives the following corollary.

Corollary 3. Let $H_{v, \gamma} = \{x \in \mathbb{R}^n : v \cdot x \geq \gamma\}$ be a halfspace in \mathbb{R}^n . For every $\eta > 0$, there is a factor $\beta_{\text{safe}} = O(\log 1/\eta)$ such that if for every component i in a logconcave mixture \mathcal{F} ,

$$|v \cdot \mu_i - \gamma| > \beta_{\text{safe}} \sigma_i,$$

then $H_{v, \gamma}$ is η -safe for \mathcal{F} .

Halfspaces are then easily combined into polyhedra.

Proposition 4. *If P_1 is η_1 -safe for \mathcal{F} and P_2 is η_2 -safe for \mathcal{F}_{P_1} , then $P_1 \cap P_2$ is $(\eta_1 + \eta_2)$ -safe for \mathcal{F} .*

Proof. Suppose component $i \in I_{P_1 \cap P_2}$ and let x be distributed according to component i . Then

$$\Pr[x \notin P_1 \cap P_2] \leq \Pr[x \notin P_1] + \Pr[x \notin P_2] \leq \eta_1 + \eta_2.$$

Now, suppose component $i \notin I_{P_1 \cap P_2}$. We distinguish two cases. If $i \notin I_{P_1}$, then

$$\Pr[x \in P_1 \cap P_2] \leq \Pr[x \in P_1] \leq \eta_1.$$

On the other hand, if $i \in I_{P_1 \setminus P_2}$, then

$$\Pr[x \in P_1 \cap P_2] \leq \Pr[x \in P_2] \leq \eta_2.$$

□

3.2 Properties of Sample Sets

As we will argue, the polyhedra obtained by the algorithm will be safe. Therefore, we expect that the polyhedra will contain the points from the components whose means are contained in the polyhedra. We also expect that no set chosen in steps 1 or 3 of Algorithm 2 will contain much more than its share of noise points and that the empirical means and variances will be close to those of the component distributions themselves. Our analysis rests on these sets obtained from LM in steps 1 and 3 having these and other key properties that are summarized in the following definition.

Definition 2. A set $S_1 \cup \dots \cup S_n \cup N$ of m points from LM is *good* for subspace W , polyhedron P , and scalar β if the following conditions hold.

1. For every component i , if $\mu_i \in P$, then $S_i \subseteq P$, and if $\mu_i \notin P$, then $S_i \cap P = \emptyset$.
2. $|S_i| \geq w_i m / 2$ for all components i , and $|N| \leq 2\epsilon m$.
3. For every component $i \in I_P$ and every $p \in S_i$, $\|\text{proj}_W(p - \mu_i)\| \leq \beta R_i^{(W)}$
4. For every component $i \in I_P$ $\|\text{proj}_W(\mu_i - \hat{\mu}_i)\| \leq \frac{\sigma_i}{4}$.
5. For every component $i \in I_P$

$$\frac{7}{8} R_i^{(W)} \leq \frac{1}{|S_i|} \sum_{p \in S_i} \|\text{proj}_W(p - \hat{\mu}_i)\|^2 \leq \frac{8}{7} R_i^{(W)}.$$

6. For some pair $i, j \in I_P$ such that $\|\text{proj}_W(\mu_i - \mu_j)\| = \mu_{\max}^{(W,P)}$, it holds for all $p \in S_i \cup S_j$, that $\|\text{proj}_u(p - \mu_{\ell(p)})\| \leq \beta \sigma_{\max}^{(P)}$, where u is the unit vector along the direction $\text{proj}_W(\mu_i - \mu_j)$.

For convenience, we will sometimes say the set $Z = \text{proj}_W(Z_0 \cap P)$ is “generated by a good set for W, P , and β .” This is not really a property of the set Z itself, but rather of W, P, β and an implicit Z_0 (drawn from LM) such that $Z = \text{proj}_W(Z_0 \cap P)$.

It is important to note that a set is only good *for a particular subspace*. In our analysis on Robust PCA, we will require that Z_i from step 1 of Algorithm 2 be good for polyhedron P and the current subspace W in step 2a of Robust PCA, where $Z_i \cap P$ is used. Thus, Z_1 must be good for \mathbb{R}^n and Z_2 must be good for the subspace obtained after one iteration in Robust PCA, etc. Finally, the set X_0 used in step 3 of Algorithm 2 must be good for W_k the subspace returned by Robust PCA. The following lemma shows that this happens with high probability.

Lemma 5. Let Z_0 be a set of m points generated by ϵ -noisy sample source LM for a logconcave mixture. Let P be polyhedron that is $(\delta/2m)$ -safe. Let W be any subspace of \mathbb{R}^n . There exist $M_{good} = O(n/w_{\min} \log^5 nk/\delta)$ and $\beta_{good} = O(\log(mk/\delta))$ such that with probability $1 - \delta$ if $m \geq M_{good}$, then Z_0 is good for W , P , and any $\beta \geq \beta_{good}$.

Proof. We consider the goodness properties in order. From the definition of η -safe, item 1 holds with probability $1 - \delta/2$. Item 2 follows from a Chernoff bound (recall that ϵ is an upper bound on the noise of LM and not the noise itself). The remaining items are standard results for logconcave distributions. See [13]. \square

3.3 Bounds on t

It is important that we be able to approximate the greatest distance between two non-noise points. This enables us to put a ball around the non-noise data in Robust PCA so as to remove noise points that are far away from the non-noise points (step 2d of Robust PCA). It is also critical in determining d the resolution at which we look for valleys in the partitioning phase (step 4 of the clustering algorithm).

Lemma 6. Suppose that $Z = S \cup N$ was generated by a good set from LM for W , P , and β . Then $t = t(Z)$ has the bounds

$$\max\{\mu_{\max}^{(W,P)} - 2\beta\sigma_{\max}^{(P)}, R_{\max}^{(W,P)}/2\} \leq t \leq \max_{p,q \in S} \|p - q\| \leq \mu_{\max}^{(W,P)} + 2\beta R_{\max}^{(W,P)}.$$

Proof. By definition $Z = \text{proj}_W(Z_0 \cap P)$, where Z_0 is good for P , W and β . For convenience, we partition Z into the non-noise points S and the noise points N , so that $Z = S \cup N$. To avoid cumbersome notation, we will drop the superscript W and P for the quantities μ_{\max} and R_{\max} .

To obtain the upper bound, observe that for any $p \in S \cup N$,

$$s\{\|p - q\|\}_{q \in S \cup N} \leq \max_{q \in S} \|p - q\|,$$

since there are at most $2\epsilon m$ elements in N . Similarly,

$$t = s\{s\{\|p - q\|\}_{q \in S \cup N}\}_{p \in S \cup N} \leq s\{\max_{q \in S} \|p - q\|\}_{p \in S \cup N} \leq \max_{p,q \in S} \|p - q\|.$$

But for any pair of points $p, q \in S$,

$$\|p - q\| \leq \|p - \text{proj}_W(\mu_{\ell(p)})\| + \|\text{proj}_W(\mu_{\ell(p)} - \mu_{\ell(q)})\| + \|\text{proj}_W(\mu_{\ell(q)} - q)\|$$

by the triangle inequality, where $\ell(p)$ is the index of the component from which p was drawn. Using the definition of “good” (Definition 2, item 3), we have that the first and last terms are bounded by βR_{\max} . Combining this with the definition of μ_{\max} , we have

$$t \leq \max_{p,q \in S} \|p - q\| \leq \mu_{\max} + 2\beta R_{\max}.$$

Next we give a lower bound in terms of μ_{\max} . For a pair of components i and j

$$\begin{aligned} t &= s\{s\{\|p - q\|\}_{q \in S \cup N}\}_{p \in S \cup N} \\ &\geq s\{s\{\|p - q\|\}_{q \in S_j}\}_{p \in S_i}. \end{aligned}$$

Note that these quantities are well defined, since $|S_j|, |S_i| > w_{\min}m/2 > \lfloor 2\epsilon m \rfloor$ by item 2 of Definition 2 and our choice of ϵ . We continue

$$\begin{aligned} s\{s\{\|p - q\|\}_{q \in S_j}\}_{p \in S_i} &\geq s\{\min_{q \in S_j} \|p - q\|\}_{p \in S_i} \\ &\geq \min_{p \in S_i, q \in S_j} \|p - q\|. \end{aligned}$$

Now suppose that i and j are the two components such that $\|\text{proj}_W(\mu_i - \mu_j)\| = \mu_{\max}$ and $\mu_i, \mu_j \in P$. Let u be the unit vector along the direction $\text{proj}_W(u_i - u_j)$. Then for any $p \in S_i, q \in S_j$ that are closest we have from the triangle inequality that

$$\begin{aligned} \|p - q\| &\geq |\text{proj}_u(p - q)| \\ &\geq |\text{proj}_u(\mu_i - \mu_j)| - |\text{proj}_u(\mu_i - p)| - |\text{proj}_u(\mu_j - q)|. \end{aligned}$$

Using the definition of ‘‘good’’ again (Definition 2, item 6), have that $|\text{proj}_u(\mu_i - p)|$ and $|\text{proj}_u(\mu_j - q)|$ are at most $\beta\sigma_{\max}$. At the same time, by construction $|\text{proj}_u(\mu_i - \mu_j)| = \mu_{\max}$. Thus,

$$t \geq \min_{p \in S_i, q \in S_j} \|p - q\| \geq \mu_{\max} - 2\beta\sigma_{\max}.$$

To give a lower bound in terms of R_{\max} , let i be a component such that $R_{\max} = S_i$. Then for any $p \in R_i$, by item 5 of Definition 2

$$\begin{aligned} \frac{7}{8}R_{\max} &\leq \frac{1}{|S_i|} \sum_{q \in S_i} \|\text{proj}_W(\hat{\mu}_i) - q\|^2 \\ &\leq \frac{1}{|S_i|} \sum_{q \in S_i} \|p - q\|^2 \\ &\leq s\{\|p - q\|^2\}_{q \in S_i} + \frac{2\epsilon m}{|S_i|} \max_{q \in S_i} \|p - q\|^2. \end{aligned}$$

Applying items 2 and 3 of Definition 2 to the last term, we have that $2\epsilon m/|S_i| \leq 4\epsilon m/w_{\min}$ and $\max_{q \in S_i} \|p - q\|^2 \leq 4\beta^2 R_{\max}^2$. Thus,

$$\frac{7}{8}R_{\max} \leq s\{\|p - q\|^2\}_{q \in S_i} + \frac{4\epsilon}{w_{\min}} (4\beta^2 R_{\max}^2)$$

Rearranging this yields,

$$s\{\|p - q\|^2\}_{q \in S_i} \geq \frac{7}{8}R_{\max} - \frac{4\epsilon}{w_{\min}} (4\beta^2 R_{\max}^2).$$

For an appropriate choice of $\epsilon = C_\epsilon w_{\min} \beta^{-2}$, we have the lower bound

$$s\{\|p - q\|^2\}_{q \in S_i} \geq R_{\max}^2/4,$$

which holds for every $p \in S_i$. Thus,

$$t^2 = s\{s\{\|p - q\|^2\}_{q \in S \cup N}\}_{p \in S \cup N} \geq s\{s\{\|p - q\|^2\}_{q \in S_i}\}_{p \in S_i} \geq R_{\max}^2/4.$$

□

3.4 A Spectral Lemma

For a matrix A , let $\lambda_j(A)$ be the j th largest eigenvalue of the matrix. When the matrix is clear from context, we may simply write λ_j . The following lemma will be useful in our analysis of Robust PCA.

Lemma 7. *Let $A = M + C$ where M and C are symmetric positive semi-definite $n \times n$ matrices and $\text{rank}(M) = k$. Then for $j > k$,*

$$\lambda_j(A) \leq \frac{1}{j - k} \sum_{i=1}^j \lambda_i(C).$$

Proof of Lemma 7. We use the following well-known theorem (see Theorem 4.8 of [20] for example).

Theorem 8. *Let $A = M + C$ where M and C are symmetric n -by- n matrices. Then*

$$\begin{aligned} \sum_{i=1}^j \lambda_i(M) + \lambda_{n-j+i}(E) &\leq \sum_{i=1}^j \lambda_i(M + E) \\ &\leq \sum_{i=1}^j \lambda_i(M) + \lambda_i(E). \end{aligned}$$

Thus,

$$\sum_{i=1}^k \lambda_i(M) \leq \sum_{i=1}^k \lambda_i(A)$$

and

$$\sum_{i=1}^j \lambda_i(A) \leq \sum_{i=1}^j \lambda_i(M) + \lambda_i(C).$$

Using the first of these inequalities shows that

$$\begin{aligned} (j - k)\lambda_j(A) &\leq \sum_{i=k+1}^j \lambda_i(A) \\ &\leq \sum_{i=k+1}^j \lambda_i(A) + \sum_{i=1}^k \lambda_i(A) - \sum_{i=1}^k \lambda_i(M) \\ &= \sum_{i=1}^j \lambda_i(A) - \sum_{i=1}^k \lambda_i(M). \end{aligned}$$

The second then shows

$$\sum_{i=1}^j \lambda_i(A) - \sum_{i=1}^k \lambda_i(M) \leq \sum_{i=1}^j \lambda_i(C),$$

since $\lambda_i(M) = 0$ for $i > k$. □

4 Analysis

We now turn to the major portion of our analysis. In Section 4.1, we analyze the effect of Robust PCA, showing that it preserves much of the distance between at least two means. In Section 4.3, we show the correctness of the partitioning step. Finally, we synthesize the whole argument in Section 4.2 to give the main theorem.

The essential parameters of the algorithm and analysis are $m_Z, m_X, m_Y, \beta, \epsilon$, and α . In terms of the quantities $n, w_{\min}, \delta, \eta$, and k , these are

$$\begin{aligned} m_Z &= C_Z n w_{\min}^{-1} \log^5(nk/\delta) \\ m_X &= C_X n w_{\min}^{-1} \log^5(nk/\delta) \\ m_Y &= C_Y w_{\min}^{-1} \log(k/\delta) \\ \beta &= C_\beta \log((m_X + m_Y + m_Z)k \log(n)/(\delta\eta)) \\ &= O(\log(nk/(w_{\min}\delta\eta))) \\ \epsilon &= C_\epsilon w_{\min}/\beta^2 \\ &= \Omega(w_{\min} \log^{-2}(nk/(w_{\min}\delta\eta))) \\ \alpha &= C_\alpha k^{3/2} w_{\min}^{-1} \beta \log n \\ &= O(k^{3/2} w_{\min}^{-1} \log(nk/(w_{\min}\delta\eta))) \end{aligned}$$

where the leading factor is an appropriate constant. We will exercise the choice of these constants in the course of the analysis. The reader will find it useful to refer to these equations in following the proof. Without loss of generality, we may assume that η is a polynomial factor smaller than δ .

4.1 Robust PCA

Here we show that Robust PCA preserves most of the distance between the two components that are furthest part (Lemma 9). We accomplish this by showing that only a small fraction of this distance is lost as the dimension of the data is halved in each iteration (Lemma 10).

This result rests on two key claims. Claim 11 shows that the diameter of the non-noise data can be approximated in the presence of noise and that this permits the algorithm to place a relatively tight ball around the non-noise data, excluding noise points that are far away. The estimated diameter (roughly the parameter t) can neither be too small (or non-noise points will be excluded), nor too large (or noise points at the edge of the ball may have too large of an effect on the eigenvectors).

The other key claim is Claim 12 which bounds the maximum variance of the data in the subspace that is thrown out in an iteration. Recall that Robust PCA projects to $\lfloor (\dim(W) - k)/2 \rfloor + k$ dimensions, removes outliers outside the ball $B(p_0, 16\beta t)$, and repeats until a k dimensional subspace is found.

To illustrate why simply removing outliers and using standard PCA to project to a k dimensional subspace is inadequate, we define the following matrices. Let N' be the set of noise points after outliers are removed and assume that no non-noise points are removed. The remaining points are therefore $S \cup N'$. Assume p_0 be the origin, that $W = \mathbb{R}^n$ and consider the matrix computed in step 2e of Robust PCA

$$A = \frac{1}{m'} \sum_{p \in S \cup N'} pp^T,$$

where N' consists of the noise points that were not removed and $m' = |S \cup N'|$. Using the sample means \hat{u}_i and covariance $\hat{\Sigma}_i$, we can decompose this matrix as the sum of

$$\begin{aligned} M &= \frac{1}{m'} \sum_{i=1}^k |S_i| \hat{\mu}_i \hat{\mu}_i^T \\ C &= \frac{1}{m'} \sum_{i=1}^k |S_i| \hat{\Sigma}_i \\ E &= \frac{1}{m'} \sum_{p \in N'} pp^T. \end{aligned}$$

The matrix M is the mixture of the outer product of the means, C is the mixture of the covariances, and E is the noise contribution.

Without noise, the second moment matrix A is just $M + C$. The rank of M is k and its eigenvectors are the subspace that we would ideally like to find, i.e. the span of the means. The matrix C can be viewed as a perturbation, which may cause the eigenvectors of $M + C$ to differ from those of M . The 2-norm of C is bounded from above by σ_{\max}^2 , while the 2-norm of M is bounded from below in terms of μ_{\max}^2 . For an adequate separation of the component means the matrix M dominates so that applying PCA to $M + C$ gives a k dimension subspace that is close to the span of the means.

In the presence of noise, however, we must account not only for the perturbation caused by C but that caused by E (the noise component) as well. At first, it may seem that the noise component cannot have a large effect. As we will show in the proof of Claim 12, the sum of the eigenvalues of E is comparable to that of C . Recall that the 2-norm is the largest eigenvalue. While the sum of the eigenvalues of C may be on the order of $n\sigma_{\max}^2$, this is spread out over all n eigenvectors, so that no one eigenvalue is larger than σ_{\max}^2 . We have no such guarantee for E ; the sum may be concentrated in a single eigenvalue and therefore a single eigenvector. Even worse, it could be spread out over a constant fraction of the eigenvectors, each challenging

the dominance of the eigenvectors of M . Hence, some constant fraction of the dimension must be preserved in order to avoid removing the eigenvectors of M , i.e. the span of the means. Claim 12 shows that half of the dimension is adequate to preserve most of the distance between the means.

In our analysis, we often will identify a subspace by giving its dimension as a subscript. For instance, we will use W_k to denote the subspace returned by Robust PCA and W_ℓ for intermediate subspaces within Robust PCA. The main result of this section is the following lemma.

Lemma 9. *Suppose that every set Z obtained in step 2a of Robust PCA was generated by a good set for the current subspace W , P , and β . Then, letting W_k be the final subspace,*

$$\mu_{\max}^{(W_k, P)^2} \geq \frac{1}{2} \mu_{\max}^{(P)^2}.$$

Proof. This lemma is proved by applying the following lemma to each successive projection, until $n = k$.

Lemma 10. *Let P be a polyhedron in \mathbb{R}^n and let W be a subspace in \mathbb{R}^n with dimension greater than k , where $\mu_{\max}^{(W, P)} \geq \mu_{\max}^{(P)}/2$. Suppose that the set Z in step 2a of Robust PCA is generated by a good set for W, P and β . Let W_ℓ be the subspace of dimension $\lfloor (\dim(W) - k)/2 \rfloor + k$ obtained in step 2e. Then for all pairs of components $i, j \in I_P$*

$$\|\text{proj}_{W_\ell}(\mu_i - \mu_j)\|^2 \geq \|\text{proj}_W(\mu_i - \mu_j)\|^2 - (\lfloor (\dim(W) - k)/2 \rfloor + 1)^{-1} \frac{1}{4} \mu_{\max}^{(P)^2} - \frac{32k}{w_{\min}} \sigma_{\max}^{(P)^2}.$$

Continuing with the proof of Lemma 9, let μ_i and μ_j be the means of two components such that $\|\mu_i - \mu_j\| = \mu_{\max}^{(P)}$. We observe that the quantity $\mu_{\max}^{(W, P)}$ can only decrease as the dimension of W is reduced. Therefore, when we unravel the recurrence relation implied by Lemma 10, we may simplify the bound to

$$\|\text{proj}_{W_k}(\mu_i - \mu_j)\|^2 \geq \|\mu_i - \mu_j\|^2 - \frac{\mu_{\max}^{(P)^2}}{8} \sum_{j=0}^{\lceil \log_2(n-k) \rceil} \frac{1}{2^j} - \sum_{j=0}^{\lceil \log_2(n-k) \rceil} \frac{32}{w_{\min}} k \sigma_{\max}^{(P)^2},$$

By our choice of the pair μ_i, μ_j , we have $\|\mu_i - \mu_j\|^2 = \mu_{\max}^{(P)^2}$. Clearly, the first sum is bounded by $\mu_{\max}^{(P)^2}/4$. The second sum becomes $32k w_{\min}^{-1} \sigma_{\max}^2 \lceil \log_2(n-k) \rceil$. By the choice of α in Theorem 1, however, we may assume that this is no larger than $\mu_{\max}^{(P)^2}/4$ either. Thus,

$$\mu_{\max}^{(W, P)^2} \geq \|\text{proj}_{W_k}(\mu_i - \mu_j)\|^2 \geq \frac{1}{2} \mu_{\max}^{(P)^2}.$$

□

Proof of Lemma 10. Let $Z = S \cup N$ be a set generated by a good set of m_Z samples from LM. Let p_0, q_0 be a pair of points in $S \cup N$ satisfying $\|p_0 - q_0\| = t$. Because the denoising step removes all points outside of the ball $B(p_0, 16\beta t)$, we define the sets of remaining points $S' = S \cap B(p_0, 16\beta t)$ and $N' = N \cap B(p_0, 16\beta t)$. For convenience, we define $m' = |S' \cup N'|$.

We first claim that no non-noise points are eliminated (i.e. $S = S'$) and give a bound on the radius of the ball $B(p_0, 16\beta t)$.

Claim 11. *Suppose $p_0, q_0 \in S \cup N$ satisfy $\|p_0 - q_0\| = t$. Then*

$$S \subseteq B(p_0, 16\beta t) \subseteq B(p_0, 32\beta^2(\mu_{\max} + R_{\max})).$$

Thus, the second moment matrix used for the spectral analysis becomes

$$A = \frac{1}{m'} \sum_{p \in S \cup N'} (p - p_0)(p - p_0)^T.$$

This matrix has the following critical property.

Claim 12. Suppose that $p_0, q_0 \in S \cup N$ satisfy $\|p_0 - q_0\| = t$. Then for $\ell = \lfloor (\dim(W) - k)/2 \rfloor + k$,

$$\lambda_{\ell+1}(A) \leq \frac{w_{\min}}{64} (\lfloor (\dim(W) - k)/2 \rfloor + 1)^{-1} \mu_{\max}^{(W,P)^2} + 4k\sigma_{\max}^{(P)^2}.$$

By definition W_ℓ is the span of the top ℓ components of the matrix A . Let \bar{W}_ℓ be the complementary subspace in W . Consider a pair of means μ_i, μ_j . We will establish an upper bound on $\|\text{proj}_{\bar{W}_\ell}(\mu_i - \mu_j)\|^2$ and thus a lower bound on

$$\|\text{proj}_{W_\ell}(\mu_i - \mu_j)\|^2 = \|\text{proj}_W(\mu_i - \mu_j)\|^2 - \|\text{proj}_{\bar{W}_\ell}(\mu_i - \mu_j)\|^2 \quad (4)$$

to prove the lemma.

Let v denote unit vector in the direction of $\text{proj}_{\bar{W}_\ell}(\mu_i - \mu_j)$. Let $e = \mu_i - \mu_j - (\hat{\mu}_i - \hat{\mu}_j)$. Then

$$\begin{aligned} \|\text{proj}_{\bar{W}_\ell}(\mu_i - \mu_j)\|^2 &= (v^T(\mu_i - \mu_j))^2 \\ &\leq 2(v^T(\hat{\mu}_i - \hat{\mu}_j))^2 + 2\|e\|^2. \end{aligned}$$

By item 4 of Definition 2, the term $\|e\|^2$ is bounded from above by $\sigma_{\max}^{(P)^2}/4$. To bound the remaining term we use the fact that

$$1 \leq \frac{2}{w_{\min}} \frac{|S_i|}{m} \leq \frac{2}{w_{\min}} \frac{|S_i|}{m'}$$

from item 2 of Definition 2 to argue

$$\begin{aligned} (v^T(\hat{\mu}_i - \hat{\mu}_j))^2 &\leq 2((v^T(\hat{\mu}_i - p_0))^2 + (v^T(\hat{\mu}_j - p_0))^2) \\ &\leq 2 \sum_{i=1}^k (v^T(\hat{\mu}_i - p_0))^2 \\ &\leq \frac{4}{w_{\min}} \cdot \frac{1}{m'} \sum_{i=1}^k |S_i| (v^T(\hat{\mu}_i - p_0))^2. \end{aligned}$$

For each i ,

$$|S_i| (v^T(\hat{\mu}_i - p_0))^2 \leq \sum_{p \in S_i} (v^T(p - p_0))^2.$$

Including the points from N' , we then have

$$\frac{1}{m'} \sum_{i=1}^k |S_i| (v^T(\hat{\mu}_i - p_0))^2 \leq v^T \left(\frac{1}{m'} \sum_{p \in S \cup N'} (p - p_0)(p - p_0)^T \right) v \leq \lambda_{\ell+1}(A).$$

since v is in the subspace \bar{W}_ℓ .

From Claim 12, we have

$$\begin{aligned} \|\text{proj}_{\bar{W}_\ell}(\mu_i - \mu_j)\|^2 &\leq \frac{4}{w_{\min}} \lambda_{\ell+1}(A) + \frac{\sigma_{\max}^{(P)^2}}{4} \\ &\leq \lfloor (\dim(W) - k)/2 \rfloor^{-1} \frac{1}{8} \mu_{\max}^{(W,P)^2} + \frac{32}{w_{\min}} k \sigma_{\max}^{(P)^2}. \end{aligned}$$

Combined with Eqn. 4, this proves the lemma. \square

Proof of Claim 11. Since Lemma 10 assumes that $\mu_{\max}^{(W,P)} \geq \mu_{\max}^{(P)}/2$, we have

$$2\beta\sigma_{\max}^{(P)} \leq \alpha\sigma_{\max}^{(P)}/4 \leq \mu_{\max}^{(P)}/4 \leq \mu_{\max}^{(W,P)}/2,$$

using a suitable α and the separation assumption of Eqn. 2.

By Lemma 6 then

$$t \geq \mu_{\max}^{(W,P)} - 2\beta\sigma_{\max}^{(P)} \geq \mu_{\max}^{(W,P)}/2.$$

Also, by the same lemma $t \geq R_{\max}^{(W,P)}/2$. Without loss of generality assume $\beta \geq 1$. Then

$$\max_{p,q \in S} \|p - q\| \leq \mu_{\max}^{(W,P)} + 2\beta R_{\max}^{(W,P)} \leq 2t + 4\beta t < 8\beta t.$$

Thus, no two points in S can be further than $8\beta t$ apart.

Now let p be an arbitrary point in S and let $q \in S \cap B(p_0, t)$. Note that such a point q exists because by definition of t , $B(p_0, t)$ contains $(1 - 2\epsilon m) > |N|$ points. We have by the triangle inequality and Lemma 6 that

$$\|p_0 - p\| \leq \|p_0 - q\| + \|q - p\| \leq t + 8\beta t < 16\beta t \leq 32\beta^2(\mu_{\max}^{(W,P)} + R_{\max}^{(W,P)}).$$

Thus, $S \subseteq B(p_0, 16\beta t) \subseteq B(p_0, 32\beta^2(\mu_{\max} + R_{\max}))$. \square

Proof of Claim 12. Without loss of generality, let us assume that p_0 is the origin. Thus, the matrix from line 6 of the algorithm becomes $A = \frac{1}{m'} \sum_{p \in S \cup N'} pp^T$. Using the sample means $\hat{\mu}_i$, we can decompose this matrix as the sum of

$$\begin{aligned} M &= \frac{1}{m'} \sum_{i=1}^k |S_i| \text{proj}_W(\hat{\mu}_i) \text{proj}_W(\hat{\mu}_i)^T \\ C &= \frac{1}{m'} \sum_{i=1}^k \sum_{p \in S_i} \text{proj}_W(p - \hat{\mu}_i) \text{proj}_W(p - \hat{\mu}_i)^T \\ E &= \frac{1}{m'} \sum_{p \in N'} pp^T. \end{aligned}$$

Our strategy will be to bound $\sum_{i=1}^{\ell+1} \lambda_i(C + E)$ and apply Lemma 7 to bound $\lambda_{\ell+1}(A)$.

$$\begin{aligned} \sum_{i=1}^n \lambda_i(C) &\leq \frac{1}{m'} \sum_{i=1}^k \sum_{p \in S_i} \|\text{proj}_W(p - \hat{\mu}_i)\|^2 \\ &= \sum_{i=1}^k \frac{|S_i|}{m'} \frac{1}{|S_i|} \sum_{p \in S_i} \|\text{proj}_W(p - \hat{\mu}_i)\|^2 \\ &\leq \max_i \frac{1}{|S_i|} \sum_{p \in S_i} \|\text{proj}_W(p - \hat{\mu}_i)\|^2 \\ &\leq \frac{8}{7} R_{\max}^{(W,P)2}. \end{aligned}$$

By Claim 11, $N' \subseteq B(0, 16\beta t) \subseteq B(32\beta^2(\mu_{\max}^{(W,P)} + R_{\max}^{(W,P)}))$, so

$$\sum_{i=1}^n \lambda_i(E) \leq 2\epsilon \max_{p \in N'} \|p\|^2 \leq \epsilon 64\beta^2(\mu_{\max}^{(W,P)2} + R_{\max}^{(W,P)2}) \leq \frac{w_{\min}}{64}(\mu_{\max}^{(W,P)2} + R_{\max}^{(W,P)2}),$$

for an appropriate choice of $\epsilon = C_\epsilon w_{\min} \beta^{-2}$ from Theorem 1. Combining these bounds

$$\begin{aligned} \sum_{i=1}^n \lambda_i(C) + \lambda_i(E) &\leq \frac{w_{\min}}{64} \mu_{\max}^{(W,P)2} + \left(\frac{w_{\min}}{64} + \frac{8}{7} \right) R_{\max}^{(W,P)2} \\ &\leq \frac{w_{\min}}{64} \mu_{\max}^{(W,P)2} + 2 \dim(W) \sigma_{\max}^2. \end{aligned}$$

Lemma 7 then gives the bound

$$\begin{aligned}
\lambda_{\ell+1}(A) &\leq \frac{1}{\ell+1-k} \sum_{i=1}^{\ell+1} \lambda_i(C+E) \\
&\leq \frac{1}{\ell+1-k} \sum_{i=1}^n \lambda_i(C) + \lambda_i(E) \\
&\leq (\lfloor (\dim(W) - k)/2 \rfloor + 1)^{-1} \\
&\quad \left(\frac{w_{\min}}{64} \mu_{\max}^{(W,P)^2} + 2 \dim(W) \sigma_{\max}^2 \right). \tag{5}
\end{aligned}$$

We note that $2 \dim(W) / (\lfloor (\dim(W) - k)/2 \rfloor + 1) \leq 4k$. If $\dim(W) < 2k$, then this is trivial. On the other hand, if $\dim(W) \geq 2k$, then

$$\frac{2 \dim(W)}{\lfloor (\dim(W) - k)/2 \rfloor + 1} \leq \frac{4 \dim(W)}{\dim(W) - k} \leq 8 \leq 4k.$$

The bound of Eqn. 5 then becomes

$$\lambda_{\ell+1}(A) \leq (\lfloor (\dim(W) - k)/2 \rfloor + 1)^{-1} \frac{w_{\min}}{64} \mu_{\max}^{(W,P)^2} + 4k \sigma_{\max}^2.$$

□

4.2 Partitioning Components

We show that Algorithm 2 successfully partitions the components. The algorithm tries many directions in the subspace W until it finds a one with a “valley” corresponding to the intuition given in Section 2. We capture this notion formally in the following definition.

Definition 3. Let X be a set of m_X points in \mathbb{R} . For $i \in \mathbb{Z}$ let $b_i = |\{x \in X : x \in (id, (i+1)d)\}|$. We say that X has a *valley* if there is a triple $i_1 < i_2 < i_3$ such that $b_{i_1}, b_{i_3} > w_{\min} m_X / 8$ and $b_{i_2} \leq 2\epsilon m_X$. We define the point $d(i_2 + 1/2)$ to be the *middle* of the valley.

Assuming that d is well-chosen, the existence of a valley is ensured by the fact that the means are well-separated compared to the width of the widest component. If d is too small, then we are likely to find valleys within the point set of a single component. If d is too large, then the whole mixture might fit into a single unit of resolution or “bucket.” When d is chosen correctly, non-noise points from two components fill the outer buckets, while only noise points fill the middle one.

The following two Lemmas show that with high probability the algorithm succeeds in a given node in the recursion tree. The first applies for internal nodes in the tree where components need to be separated, the second applies to the leaves where the division of space terminates.

Lemma 13. Let $\delta > 0$. Suppose that P is η -safe for \mathcal{F} and that $|I_P| > 1$. Then with probability $1 - \delta$, the halfspaces $H_{v,\gamma}$ and $H_{-v,-\gamma}$ are (η/k) -safe and each contains at least one component mean of I_P .

Lemma 14. Let $\delta > 0$. Suppose that P is η -safe for \mathcal{F} and that $|I_P| = 1$. Then with probability $1 - \delta$, no valley will be found by Algorithm 2 in step 6c.

Proof of Lemma 13. We first consider the set Y_0 obtained in line 6. Let $Y_0 = S_1 \cup \dots \cup S_k \cup N$ disjointly, where N is the set of noise points and S_i is the set of points from component i . We show that with probability $1 - \delta/2$ the set Y_0 has the following two properties.

1. Y_0 contains at least one point from every component.
2. For every component i and every $p \in S_i$, $\|\text{proj}_W(p - \mu_i)\| \leq \beta R_i^{(W)}$.

The probability that no point from component i is in a set of m_Y points is at most $(1 - w_{\min})^m \leq \exp(-w_{\min} m_Y) \leq \frac{\delta}{4k}$, where we have used the fact that $m_Y \geq w_{\min}^{-1} \log(4k/\delta)$ in the last step. Taking a union bound over all k components shows that the probability that no samples are taken from some component is at most $\delta/4$.

To show the second property, we consider projection of the mixture distribution onto the subspace W . Component i of this distribution has mean $\text{proj}_W(\mu_i)$ and the component is logconcave, being the projection of a logconcave distribution. Applying theorem 2 gives the result, since β is larger than the requirement of $\beta \geq O(\log(m_Y/\delta))$.

Without loss of generality, we assume that $\eta \leq \delta(4(\lceil \log n \rceil + 1)(m_Z + m_X))^{-1}$. Thus, by Lemma 5 we may argue that with probability $1 - \delta/2$ every set used by Robust PCA is generated by a good set (Definition 2) as required by Lemma 9 and the set X_0 is good as well. Overall, the collection of sample sets has the desired properties with probability $1 - \delta$.

Assuming that all sets given to Robust PCA are good, Lemma 9 guarantees that after projection to W

$$\mu_{\max}^{(W,P)} \geq \frac{\mu_{\max}^{(P)}}{2}.$$

With an appropriate choice of α , this implies that in the subspace W

$$\mu_{\max}^{(W,P)} \geq \mu_{\max}^{(P)}/2 \geq \frac{\alpha}{2} \sigma_{\max}^{(P)} \geq 41k^{3/2} \beta \sigma_{\max}^{(P)} \geq 41k \beta R_{\max}^{(P)}.$$

This fact enables us to use the following claim.

Claim 15. *If $|I_P| > 1$ and*

$$\mu_{\max}^{(W,P)} \geq 41k \beta R_{\max}^{(W,P)},$$

then the quantity $d = t/10k$ satisfies the bounds

$$4\beta R_{\max}^{(W,P)} \leq d \leq \frac{\mu_{\max}^{(W,P)}}{5k}.$$

Proof of Claim 15. We first derive a lower bound. Since $t \geq \mu_{\max}^{(W,P)} - 2\beta R_{\max}^{(W,P)}$ by Lemma 6 and $\mu_{\max}^{(W,P)} > 41k\beta R_{\max}^{(W,P)}$, we have

$$d = \frac{t}{10k} \geq \frac{\mu_{\max}^{(W,P)} - 2\beta R_{\max}^{(W,P)}}{10k} \geq \frac{41k\beta R_{\max}^{(W,P)} - 2\beta R_{\max}^{(W,P)}}{10k} \geq 4\beta R_{\max}^{(W,P)}.$$

To show the upper bound, we observe that since $\mu_{\max}^{(W,P)} \geq 41k\beta R_{\max}^{(W,P)}$, we have that $2\mu_{\max}^{(W,P)} \geq \mu_{\max}^{(W,P)} + 2\beta R_{\max}^{(W,P)}$. Thus, by Lemma 6

$$d = \frac{t}{10k} \leq \frac{\mu_{\max}^{(W,P)} + 2\beta R_{\max}^{(W,P)}}{10k} \leq \frac{2\mu_{\max}^{(W,P)}}{10k} = \frac{\mu_{\max}^{(W,P)}}{5k}.$$

□

The remainder of the proof rests on the following two claims. The first shows that any valley that is found produces a (η/k) -safe halfspace. The second claim shows that a valley will indeed be found.

Claim 16. *For any direction $v \in W_k$, if $\text{proj}_v(X)$ has a valley with midpoint γ , then $H_{v,\gamma}$ and $H_{-v,-\gamma}$ are (η/k) -safe for \mathcal{F}_P .*

Proof. Because the set X is generated by a good set, the points from a single component j must be contained in an interval that is centered about the component mean of size $2\beta R_{\max}^{(W,P)}$. By Claim 15 this is at most $d/2$, half of the width of a bucket.

Suppose that a point $d(i + 1/2)$ falls into one of these intervals for some $i \in \mathbb{Z}$. Then the entire interval must be contained in the bucket i , i.e.

$$[\mu_j - \beta R_{\max}^{(W,P)}, \mu_j + \beta R_{\max}^{(W,P)}] \subseteq [di, (d+1)).$$

But then, $b_i \geq w_{\min} m_Z / 2$. For an appropriate choice of ϵ , however, $w_{\min} m_Z / 2 > 2\epsilon m_Z$, and hence i cannot be the middle part of the valley. We conclude that if $\text{proj}_v(W)$ has valley with middle γ , then for all $i \in I_P$

$$|v \cdot \mu_i - \gamma| > \beta R_{\max}^{(W,P)} \geq \beta \sigma_i.$$

Hence by Proposition 3, the halfspaces $H_{v,\gamma}$ and $H_{-v,-\gamma}$ are (η/k) -safe for \mathcal{F}_P , since we may choose $\beta \geq \beta_3 = O(\log(k/\eta))$. \square

Claim 17. *There is a pair $(a, b) \in Y \times Y$ such that the unit vector v in the direction $a - b$ has a valley.*

Proof. Let μ_i and μ_j be two components such that $\|\text{proj}_W(\mu_i - \mu_j)\| = \mu_{\max}^{(W,P)}$. Let $a \in S_i$ and let $b \in S_j$, where $Y = S_1 \cup \dots \cup S_k \cup N$. Define v to be the unit vector along $a - b$. We will show that μ_i and μ_j are far apart v .

Because we have assumed Y is generated by a good set, $\|(a - b) - \text{proj}_W(\mu_i - \mu_j)\| \leq 2\beta R_{\max}^{(W,P)}$. Thus,

$$\begin{aligned} |\text{proj}_v(\mu_i - \mu_j)| &= \frac{|(a - b) \cdot \text{proj}_W(\mu_i - \mu_j)|}{\|(a - b)\|} \\ &\geq \|\text{proj}_W(\mu_i - \mu_j)\| \left(1 - \frac{\|(a - b) - \text{proj}_W(\mu_i - \mu_j)\|^2}{\|\text{proj}_W(\mu_i - \mu_j)\|^2}\right)^{1/2} \\ &\geq \mu_{\max}^{(W,P)} \left(1 - \frac{4\beta^2 R_{\max}^{(W,P)2}}{\mu_{\max}^{(W,P)2}}\right)^{1/2} \end{aligned}$$

Because W has only k dimensions $R_{\max}^{(W,P)2} \leq k\sigma_{\max}^{(P)2}$. As argued above by Lemma 9, $\mu_{\max}^{(W,P)} \geq \mu_{\max}^{(P)}/2 \geq \alpha\sigma_{\max}^{(P)}/2$. Therefore,

$$\begin{aligned} |\text{proj}_v(\mu_i - \mu_j)| &\geq \mu_{\max}^{(W,P)} \left(1 - \frac{\beta^2 k \sigma_{\max}^{(P)2}}{\alpha^2 \sigma_{\max}^{(P)2}}\right)^{1/2} \\ &\geq \frac{\mu_{\max}^{(W,P)}}{2}, \end{aligned}$$

for $\alpha \geq \beta\sqrt{2k}$.

By Claim 15, $d \leq \mu_{\max}^{(W,P)}/(5k)$, so we have $\text{proj}_v(\mu_i - \mu_j) \geq 5kd/2$.

We now turn our attention to the set $X = S_1 \cup \dots \cup S_k \cup N$ (the set Y will not be referred to again). Because X is good, every set $\text{proj}_v(S_i)$ must be contained in an interval centered around $\text{proj}_v(\mu_i)$ of length $2\beta R_{\max}$. By the lower bound on d from Claim 15, the width of this interval is at most $d/2$. Since there are k of these, this leaves $5kd/2 - kd/2 = 2kd$ of “empty” space between $\text{proj}_v(\mu_i)$ and $\text{proj}_v(\mu_j)$, in which only noise point can fall. This space can be cut into at most $k - 1$ pieces, meaning that at least one piece must have length $2d$. An interval $[d\ell, d(\ell + 1))$ must be contained in one of these pieces, and this will form the middle of a valley, with buckets containing $\text{proj}_v(S_i)$ and $\text{proj}_v(S_j)$ serving as the other buckets. \square

\square

Proof of Lemma 14. Without loss of generality, we may assume that $\eta \leq \delta/(4m_X)$. By Lemma 5, with probability $1 - \delta/2$ the set X_0 is good for P, W_k, β .

Assuming that X_0 is good we can derive a lower bound on d . Since $t \geq R_{\max}^{(W,P)}/2$ by Lemma 6, we have that

$$d \geq \frac{R_{\max}^{(W,P)}}{20k}.$$

Also assuming that X_0 is good, we have that $X = \text{proj}_{W_k}(X_0 \cap P)$ consists only of points from a single component S_j and a set of noise points N . The set N consists of no more than $2\epsilon m_X$ points. Thus, for any direction u generated by $Y \times Y$, we have

$$b_i = |\{x \in S_j \cup N : \text{proj}_u(x) \in [di, d(i+1)]\}|.$$

For purposes of analysis, We define

$$b'_i = |\{x \in S_j : \text{proj}_u(x) \in [di, d(i+1)]\}|.$$

Suppose that $i_1 < i_2 < i_3$ form a valley. This implies that $b_{i_1} \geq w_{\min} m_X / 4$ and that

$$b'_{i_1} \geq b_{i_1} - |N| \geq w_{\min} m_X / 4 - 2\epsilon m_X \geq w_{\min} m_Z / 8.$$

choosing ϵ appropriately. The same bound holds for b_{i_3} . On the other hand,

$$b'_{i_2} \leq b_{i_2} \leq 2\epsilon m_X \leq w_{\min} m_X / 32,$$

for an appropriate choice of ϵ . Since $m_X = C_X n w_{\min}^{-1} \log^5(nk/\delta)$, we argue that this event has probability less than $\delta/2$, using the following claim.

Claim 18. *Let $\xi, \delta > 0$. Consider a logconcave distribution \mathcal{F} in one dimension with variance σ^2 and let $d \geq C\sigma$. Let S be a sample set of m points drawn from \mathcal{F} and let $b_i = |\{p \in S : p \in [di, d(i+1)]\}|$. There is a constant C' such that if $m \geq C'\xi^{-1} \log(\log(m)/C\delta)$, then with probability $1 - \delta$ the following holds for every $i \in \mathbb{Z}$ and $\xi' \geq \xi$.*

1. If $b_i > 2\xi' m$, then $\Pr[x \in [di, d(i+1)]] > \xi'$.
2. If $b_i < \xi' m / 2$, then $\Pr[x \in [di, d(i+1)]] < \xi'$.

Proof. We first observe that with probability $1 - \delta/2$ no point will be further than $\sigma \log(6m/\delta)$ away from the mean, using a trivial application (1 dimension only) of Theorem 2. Therefore, all but $2C^{-1} \log 6m/\delta$ buckets will be empty. For a single bucket, a Chernoff bound shows $m > 12\xi^{-1} \log(1/\delta')$ ensures that the desired property holds with probability $1 - \delta'$. With $\delta' = \delta C / (4 \log(6m/\delta))$, we may apply a union bound to prove the lemma. \square

\square

4.3 Proof of the Main Theorem

Proof of Theorem 1. Consider one node in the recursion tree of Algorithm 2, and suppose that P has $j < k$ support hyperplanes and that P is $(\eta j/k)$ -safe for \mathcal{F} . Note that in the root of the tree this is true because \mathbb{R}^n is 0-safe. In the case where the polyhedron contains more than one component mean (i.e. $|I_P| > 1$), Lemma 13 shows that with probability $1 - \delta'$ the half-space $H_{v,\gamma}$ obtained in line 9 excludes at least one component mean and is (η/k) -safe for \mathcal{F}_P . Proposition 4 then shows that $P \cap H_{v,\gamma}$ is $(\eta(j+1)/k)$ -safe for \mathcal{F} .

On the other hand, if the polyhedron P contains only one mean (i.e. $|I_P| = 1$), then with probability $1 - \delta'$ the algorithm does not find a valley and returns the polyhedron by Lemma 14. Thus, with probability $1 - 2k\delta'$ the algorithm returns a set of k polyhedra, each containing exactly one component mean and each η -safe for \mathcal{F} . Thus, we have by definition of η -safe that the collection of polyhedra induce a classifier that is correct with probability $1 - \eta$, as the theorem claims. \square

Acknowledgments

This research began as a class project with Vladimir Urazov. The author is grateful to Santosh Vempala for his feedback.

References

- [1] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proc. of COLT*, 2005.
- [2] S. C. Brubaker and S. Vempala. Isotropic pca and affine-invariant clustering. In M. Grötschel and G. Katona, editors, *Building Bridges Between Mathematics and Computer Science*, volume 19 of *Bolyai Society Mathematical Studies*, 2008.
- [3] K. Chaudhuri and S. Rao. Beyond gaussians: Spectral methods for learning mixtures of heavy-tailed distributions. In *Proc. of COLT*, 2008.
- [4] K. Chaudhuri and S. Rao. Learning mixtures of product distributions using correlations and independence. In *Proc. of COLT*, 2008.
- [5] A. Dasgupta, J. Hopcroft, J. Kleinberg, and M. Sandler. On learning mixtures of heavy-tailed distributions. In *Proc. of FOCS*, 2005.
- [6] S. DasGupta. Learning mixtures of gaussians. In *Proc. of FOCS*, 1999.
- [7] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JRSS B*, 39:1–38, 1977.
- [8] J. Feldman and R. O’Donnell. Learning mixtures of product distributions over discrete domains. *Siam J. or Computing*, 37(5):1536–1564, 2008.
- [9] J. Feldman, R. A. Servedio, and R. O’Donnell. Pac learning axis-aligned mixtures of gaussians with no separation assumption. In *COLT*, pages 20–34, 2006.
- [10] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *COLT*, pages 53–62, 1999.
- [11] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *Applied Statistics*, 28(1):100–108, 1979.
- [12] M. Hubert and S. Engelen. Robust pca and classification in biosciences. *Bioinformatics*, 20(11):1728–1736, 2004.
- [13] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. *Siam J. Computing*, 38(3):1141–1156, 2008.
- [14] M. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM J. on Computing*, 22(4):807–837, 1993.
- [15] F. De la Torre and M.J. Black. A framework for robust subspace learning. *IJCV*, 54:117–142, 2004.
- [16] L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Struct. Algorithms*, 30(3):307–358, 2007.
- [17] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

- [18] R. Kannan S. Arora. Learning mixtures of arbitrary gaussians. *Ann. Appl. Probab.*, 15(1A):69–92, 2005.
- [19] L. Schulman S. DasGupta. A two-round variant of em for gaussian mixtures. In *Sixteenth Conference on Uncertainty in Artificial Intelligence*, 2000.
- [20] G.W. Stewart and Ji guang Sun. *Matrix Perturbation Theory*. Academic Press, Inc., 1990.
- [21] L. G. Valiant. Learning disjunction of conjunctions. In *IJCAI*, pages 560–566, 1985.
- [22] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. *Proc. of FOCS 2002; JCCS*, 68(4):841–860, 2004.
- [23] L. Xu and A.L. Yuille. Robust principal component analysis by self-organizing rules based on statistical physics approach. *Trans. on Neural Networks*, 6(1):131 – 143, 1995.