Isotropic PCA and Affine-Invariant Clustering
(Extended Abstract)

S. Charles Brubaker Santosh S. Vempala *
Georgia Institute of Technology
Atlanta, GA 30332
{brubaker,vempala}@cc.gatech.edu

Abstract

We present an extension of Principal Component Analysis (PCA) and a new algorithm for clustering points in R^n *based on it. The key property of the algorithm is that it is affine-invariant. When the input is a sample from a mixture of two arbitrary Gaussians, the algorithm correctly classifies the sample assuming only that the two components are separable by a hyperplane, i.e., there exists a halfspace that contains most of one Gaussian and almost none of the other in probability mass. This is nearly the best possible, improving known results substantially [12, 10, 1]. For* $k > 2$ *components, the algorithm requires only that there be some* $(k-1)$*-dimensional subspace in which the "overlap" in every direction is small. Our main tools are isotropic transformation, spectral projection and a simple reweighting technique. We call this combination* isotropic PCA.

1. Introduction

We present an extension to Principal Component Analysis (PCA), which is able to go beyond standard PCA in identifying "important" directions. When the covariance matrix of the input (distribution or point set in R^n) is a multiple of the identity, then PCA reveals no information; the second moment along any direction is the same. Such inputs are called isotropic. Our extension, which we call *isotropic PCA*, can reveal interesting information in such settings. We use this technique to give an affine-invariant clustering algorithm for points in R^n.

Our main result is that applying isotropic PCA to points from a mixture of arbitrary Gaussians in $\mathbb{R}^n$ reveals a set of directions along which the Gaussians are well-separated. These directions span the Fisher subspace of the mixture, a classical concept in Pattern Recognition. Once these directions are identified, points can be classified according to which component of the distribution generated them, and hence all parameters of the mixture can be learned. Section 2.1 contains an examples that illustrates our method.

What separates this paper from previous work on learning mixtures is that our algorithm is affine-invariant. Indeed, for every mixture distribution that can be learned using a previously known algorithm, there is a linear transformation of bounded condition number that causes the algorithm to fail. For $k = 2$ components our algorithm has nearly the best possible guarantees (and subsumes all previous results) for clustering Gaussian mixtures. For $k > 2$, it requires that there be a $(k-1)$-dimensional subspace where the *overlap* of the components is small in every direction (See section 1.2). This condition can be stated in terms of the Fisher discriminant, a quantity commonly used in the field of Pattern Recognition with labeled data (see the full version for details). Because our algorithm is affine invariant, it makes it possible to unravel a much larger set of Gaussian mixtures than had been possible previously.

1.1 Previous Work

A mixture model is a convex combination of distributions of known type. In the most commonly studied version, a distribution F in R^n is composed of k unknown Gaussians. That is,

$$F = \mathrm{w}_1 N(\mu_1, \Sigma_1) + \ldots + \mathrm{w}_k N(\mu_k, \Sigma_k),$$

where the mixing weights w_i, means μ_i, and covariance matrices Σ_i are all unknown. Typically, $k \ll n$, so

*Supported in part by NSF award CCF-07 and a Raytheon fellowship.

that a concise model explains a high dimensional phenomenon. A random sample is generated from F by first choosing a component with probability equal to its mixing weight and then picking a random point from that component distribution. In this paper, we study the classical problem of unraveling a sample from a mixture, i.e., labeling each point in the sample according to its component of origin.

Heuristics for classifying samples include "expectation maximization" [6] and "k-means clustering" [11]. These methods can take a long time and can get stuck with suboptimal classifications. Over the past decade, there has been much progress on finding polynomial-time algorithms with rigorous guarantees for classifying mixtures, especially mixtures of Gaussians [5, 13, 12, 14, 10, 1]. Starting with Dasgupta's paper [5], one line of work uses the concentration of pairwise distances and assumes that the components' means are so far apart that distances between points from the same component are likely to be smaller than distances from points in different components. Arora and Kannan [12] establish nearly optimal results for such distance-based algorithms. Unfortunately their results inherently require separation that grows with the dimension of the ambient space and the largest variance of each component Gaussian.

To see why this is unnatural, consider k well-separated Gaussians in $\mathbb{R}^k$ with means $e_1, \ldots, e_k$, i.e. each mean is 1 unit away from the origin along a unique coordinate axis. Adding extra dimensions with arbitrary variance does not affect the separability of these Gaussians, but these algorithms are no longer guaranteed to work. For example, suppose that each Gaussian has a maximum variance of $\epsilon \ll 1$. Then, adding $O^*(k\epsilon^{-2})$ extra dimensions with variance ϵ will violate the necessary separation conditions.

To improve on this, a subsequent line of work uses spectral projection (PCA). Vempala and Wang [14] showed that for a mixture of *spherical* Gaussians, the subspace spanned by the top k principal components of the mixture contains the means of the components. Thus, projecting to this subspace has the effect of shrinking the components while maintaining the separation between their means. This leads to a nearly optimal separation requirement of

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

where μ_i is the mean of component i and σ_i^2 is the variance of component i along any direction. Note that there is no dependence on the dimension of the distribution. Kannan et al. [10] applied the spectral approach to arbitrary mixtures of Gaussians (and more generally, logconcave distributions) and obtained a separation that grows with a polynomial in k and the largest variance of each component:

$$\|\mu_i - \mu_j\| \geq \text{ poly}(k) \max\{\sigma_{i,\max}, \sigma_{j,\max}\}$$

where $\sigma_{i,\max}^2$ is the maximum variance of the ith component in any direction. The polynomial in k was improved in [1] along with matching lower bounds for this approach, suggesting this to be the limit of spectral methods. Going beyond this "spectral threshold" for arbitrary Gaussians has been a major open problem.

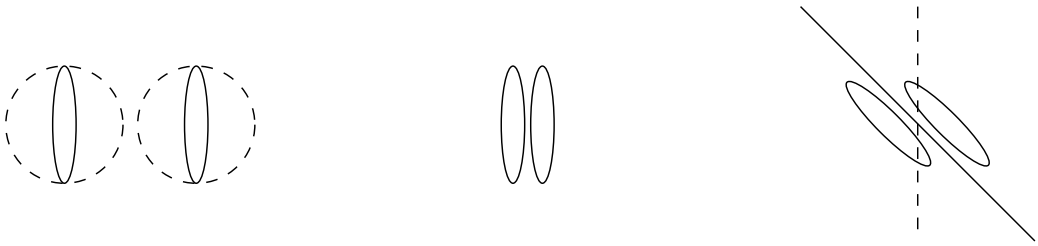
The representative hard case is the special case of two parallel "pancakes", i.e., two Gaussians that are spherical in $n-1$ directions and narrow in the last direction, so that a hyperplane orthogonal to the last direction separates the two. The spectral approach requires a separation that grows with their largest standard deviation which is unrelated to the distance between the pancakes (their means). Other examples can be generated by starting with Gaussians in k dimensions that are separable and then adding other dimensions, one of which has large variance. Because there is a subspace where the Gaussians are separable, the separation requirement should depend only on the dimension of this subspace and the components' variances in it.

A related line of work considers learning symmetric product distributions, where the coordinates are independent. Feldman et al [8] have shown that mixtures of axis-aligned Gaussians can be approximated without any separation assumption at all in time exponential in k. A. Dasgupta et al [4] consider heavy-tailed distributions as opposed to Gaussians or log-concave ones and give conditions under which they can be clustered using an algorithm that is exponential in the number of samples. Chaudhuri and Rao [3] have recently given a polynomial time algorithm for clustering such heavy tailed product distributions.

The full version of this paper [2] gives proofs for all lemmas and discusses the relationship between our notion of overlap and the Fisher discriminant.

1.2 Results

We assume we are given a lower bound w on the minimum mixing weight and k, the number of components. With high probability, our algorithm UNRAVEL returns a partition of space by hyperplanes so that each part (a polyhedron) encloses almost all of the probability mass of a single component and almost none of the other components. The error of such a set of polyhedra is the to-



(a) Distance Concentration Separability

(b) Hyperplane Separability

(c) Intermean Hyperplane and Fisher Hyperplane.

Figure 1. Previous work requires distance concentration separability which depends on the maximum directional variance (a). Our results require only hyperplane separability, which depends only on the variance in the separating direction(b). For non-isotropic mixtures the best separating direction may not be between the means of the components(c).

tal probability mass that falls outside the correct polyhedron.

We first state our result for two Gaussians in a way that makes clear the relationship to previous work that relies on separation.

Theorem 1. *Let* $\mathrm{w}_1, \mu_1, \Sigma_1$ *and* $\mathrm{w}_2, \mu_2, \Sigma_2$ *define a mixture of two Gaussians. There is an absolute constant* C *such that, if there exists a direction* v *such that*

$$|\mathrm{proj}_v(\mu_1 - \mu_2)| \geq$$
$$C\left(\sqrt{v^T\Sigma_1 v} + \sqrt{v^T\Sigma_2 v}\right)\mathrm{w}^{-2}\log^{1/2}\left(\frac{1}{\mathrm{w}\delta} + \frac{1}{\eta}\right),$$

then with probability $1-\delta$ *algorithm* UNRAVEL *returns two complementary halfspaces that have error at most* η *using time and a number of samples that is polynomial in* $n, \mathrm{w}^{-1}, \log(1/\delta)$.

The requirement is that in *some direction* the separation between the means must be comparable to the standard deviation. This separation condition of Theorem 1 is affine-invariant and much weaker than conditions of the form $\|\mu_1 - \mu_2\| \gtrsim \max\{\sigma_{1,\max}, \sigma_{2,\max}\}$ used in previous work. See Figure 1(a). The dotted line shows how previous work effectively treats every component as spherical. We require only hyperplane separability (Figure 1(b)), which is a weaker condition. We also note that the separating direction does not need to be the intermean direction as illustrated in Figure 1(c). The dotted line illustrates the hyperplane induced by the intermean direction, which may be far from the optimal separating hyperplane shown by the solid line.

For $k > 2$, instead of a single line, we seek a $(k-1)$-dimensional subspace in which to separate the components. Intuitively, we would like this subspace to minimize the distance between points and their component means relative to the distance between the means. This notion is captured in the classical Pattern Recognition concept of the Fisher discriminant [7, 9] and its optimizing subspace, which we call the Fisher subspace.

For simplicity, we adapt the definition of the Fisher subspace to the isotropic case. Recall that an isotropic distribution has the identity matrix as its covariance and the origin as its mean. Therefore,

$$\sum_{i=1}^{k}\mathrm{w}_i\mu_i = 0 \text{ and } \sum_{i=1}^{k}\mathrm{w}_i(\Sigma_i + \mu_i\mu_i^T) = I.$$

It is well known that any distribution with bounded covariance matrix (and therefore any mixture) can be made isotropic by an affine transformation.

Definition 1. Let $\{\mathrm{w}_i, \mu_i, \Sigma_i\}$ be the weights, means, and covariance matrices for an isotropic mixture distribution with mean at the origin and where $\dim(\mathrm{span}\{\mu_1, \ldots, \mu_k\}) = k-1$. Let $\ell(x)$ be the component from which x was drawn. The *Fisher subspace* F is defined as the $(k-1)$-dimensional subspace that minimizes

$$J(S) = E[\|\mathrm{proj}_S(x - \mu_{\ell(x)})\|^2].$$

over subspaces S of dimension $k-1$.

Note that $\dim(\mathrm{span}\{\mu_1, \ldots, \mu_k\})$ is only $k-1$ because isotropy implies $\sum_{i=1}^{k}\mathrm{w}_i\mu_i = 0$.

The next lemma provides a simple alternative characterization of the Fisher subspace as the span of the means of the components (after transforming to isotropic position).

Lemma 1. *Suppose $\{\mathrm{w}_i, \mu_i, \Sigma_i\}_{i=1}^k$ defines an isotropic mixture in $\mathbb{R}^n$. Let $\lambda_1 \geq \ldots \geq \lambda_n$ be the eigenvalues of the matrix $\Sigma = \sum_{i=1}^k \mathrm{w}_i \Sigma_i$ and let $v_1, \ldots, v_n$ be the corresponding eigenvectors. If the dimension of the span of the means of the components is $k-1$, then the Fisher subspace*

$$F = \mathrm{span}\{v_{n-k+2}, \ldots, v_n\} = \mathrm{span}\{\mu_1, \ldots, \mu_k\}.$$

Our algorithm attemps to find the Fisher subspace (or one close to it) and succeeds in doing so, provided that the components do not "overlap" much in the following sense.

Definition 2. The *overlap* of a mixture given as in Definition 1 is

$$\phi = \min_{S:\dim(S)=k-1} \max_{p \in S} p^T \Sigma p. \tag{1}$$

It is a direct consequence of the Courant-Fisher min-max theorem that ϕ is the $(k-1)$th smallest eigenvalue of the matrix Σ and the subspace achieving ϕ is the Fisher subspace, i.e.,

$$\phi = \left\| E[\mathrm{proj}_F(x - \mu_{\ell(x)})\mathrm{proj}_F(x - \mu_{\ell(x)})^T] \right\|_2.$$

We can now state our main theorem for $k > 2$.

Theorem 2. *There is an absolute constant C for which that following holds. Suppose that $\mathcal{F}$ is a mixture of k Gaussian components where the overlap satisfies*

$$\phi \leq C\mathrm{w}^3 k^{-3} \log^{-1}\left(\frac{nk}{\delta \mathrm{w}} + \frac{1}{\eta}\right)$$

With probability $1-\delta$, algorithm UNRAVEL *returns a set of k polyhedra that have error at most η using time and a number of samples that is polynomial in $n, \mathrm{w}^{-1}, \log(1/\delta)$.*

In words, the algorithm successfully unravels arbitrary Gaussians provided there exists a $(k-1)$-dimensional subspace in which along every direction, the expected squared distance of a point to its component mean is smaller than the expected squared distance to the overall mean by roughly a $\mathrm{poly}(k, 1/\mathrm{w})$ factor. There is no dependence on the largest variances of the individual components, and the dependence on the ambient dimension is logarithmic. This means that the addition of extra dimensions (even where the distribution has large variance) as discussed in Section 1.1 has little impact on the success of our algorithm.

2 Algorithm

The algorithm has three major components: an initial affine transformation, a reweighting step, and identification of a direction close to the Fisher subspace and a hyperplane orthogonal to this direction which leaves each component's probability mass almost entirely in one of the halfspaces induced by the hyperplane. The key insight is that the reweighting technique will either cause the mean of the mixture to shift in the intermean subspace, or cause the top $k-1$ principal components of the second moment matrix to approximate the intermean subspace. In either case, we obtain a direction along which we can partition the components.

We first find an affine transformation W which when applied to $\mathcal{F}$ results in an isotropic distribution. That is, we move the mean to the origin and apply a linear transformation to make the covariance matrix the identity. We apply this transformation to a new set of m_1 points $\{x_i\}$ from $\mathcal{F}$ and then reweight according to a spherically symmetric Gaussian $\exp(-\|x\|^2/(2\alpha))$ for $\alpha = \Theta(n/\mathrm{w})$. We then compute the mean $\hat{u}$ and second moment matrix $\hat{M}$ of the resulting set.

After the reweighting, the algorithm chooses either the new mean or the direction of maximum second moment and projects the data onto this direction h. By bisecting the largest gap between points, we obtain a threshold t, which along with h defines a hyperplane that separates the components. Using the notation $H_{h,t} = \{x \in \mathbb{R}^n : h^T x \geq t\}$, to indicate a halfspace, we then recurse on each half of the mixture. Thus, every node in the recursion tree represents an intersection of half-spaces. To make our analysis easier, we assume that we use different samples for each step of the algorithm. The reader might find it useful to read Section 2.1, which gives an intuitive explaination for how the algorithm works on parallel pancakes, before reviewing the details of the algorithm.

2.1 Parallel Pancakes

The following special case, which represents the open problem in previous work, will illuminate the intuition behind the new algorithm. Suppose $\mathcal{F}$ is a mixture of two spherical Gaussians that are well-separated, i.e. the intermean distance is large compared to the standard deviation along any direction. We consider two cases, one where the mixing weights are equal and another where they are imbalanced.

After isotropy is enforced, each component will become thin in the intermean direction, giving the density

Algorithm 1 Unravel

Input: Integer k, scalar w. Initialization: $P = \mathbb{R}^n$.

1. (Isotropy) Use samples lying in P to compute an affine transformation W that makes the distribution nearly isotropic (mean zero, identity covariance matrix).
2. (Reweighting) Use m_1 samples in P and for each compute a weight $e^{-\|x\|^2/(\alpha)}$ (where $\alpha > n/\mathrm{w}$).
3. (Separating Direction) Find the mean of the reweighted data $\hat{\mu}$. If $\|\hat{\mu}\| > \sqrt{\mathrm{w}}/(32\alpha)$, let $h = \hat{\mu}$. Otherwise, find the second moment matrix $\hat{M}$ of the reweighted points and let h be its top principal component.
4. (Recursion) Project m_2 sample points to h and find the largest gap between points in the interval $[-1/2, 1/2]$. If this gap is less than $1/4(k-1)$, then return P. Otherwise, set t to be the midpoint of the largest gap, recurse on $P \cap H_{h,t}$ and $P \cap H_{-h,-t}$, and return the union of the polyhedra produces by these recursive calls.

the appearance of two parallel pancakes. When the mixing weights are equal, the means of the components will be equally spaced at a distance of $1 - \phi$ on opposite sides of the origin. For imbalanced weights, the origin will still lie on the intermean direction but will be much closer to the heavier component, while the lighter component will be much further away. In both cases, this transformation makes the variance of the mixture 1 in every direction, so the principal components give us no insight into the inter-mean direction.

Consider next the effect of the reweighting on the mean of the mixture. For the case of equal mixing weights, symmetry assures that the mean does not shift at all. For imbalanced weights, however, the heavier component, which lies closer to the origin will become heavier still. Thus, the reweighted mean shifts toward the mean of the heavier component, allowing us to detect the intermean direction.

Finally, consider the effect of reweighting on the second moments of the mixture with equal mixing weights. Because points closer to the origin are weighted more, the second moment in every direction is reduced. However, in the intermean direction, where part of the moment is due to the displacement of the component means from the origin, it shrinks less. Thus, the direction of maximum second moment is the intermean direction.

3 Preliminaries

For a matrix Z, we will denote the ith largest eigenvalue of Z by $\lambda_i(Z)$ or just λ_i if the matrix is clear from context. Unless specified otherwise all norms are the 2-norm. For symmetric matrices, this is $\|Z\|_2 = \lambda_1(Z) = \max_{x \in \mathbb{R}^n} \|Zx\|_2/\|x\|_2$.

The following two facts from linear algebra will be useful in our analysis.

Fact 2. *Let $\lambda_1 \geq \ldots \geq \lambda_n$ be the eigenvalues for an n-by-n symmetric positive definite matrix Z and let $v_1, \ldots v_n$ be the corresponding eigenvectors. Then*

$$\lambda_n + \ldots + \lambda_{n-k+1} = \min_{S:\dim(S)=k} \sum_{j=1}^{k} p_j^T Z p_j,$$

where $\{p_j\}$ is any orthonormal basis for S. If $\lambda_{n-k} > \lambda_{n-k+1}$, then $\mathrm{span}\{v_n, \ldots, v_{n-k+1}\}$ is the unique minimizing subspace.

Recall that a matrix Z is positive semi-definite if $x^T Z x \geq 0$ for all non-zero x.

Fact 3. *Suppose that the matrix*

$$Z = \begin{bmatrix} A & B^T \\ B & D \end{bmatrix}$$

is symmetric positive semi-definite and that A and D are square submatrices. Then $\|B\| \leq \sqrt{\|A\|\|D\|}$.

We now give the proof of Lemma 1, which shows that for an isotropic distribution, the Fisher subspace is the intermean subspace.

Proof of Lemma 1. By Definition 1 for an isotropic distribution, the Fisher subspace minimizes

$$J(S) = E[\|\mathrm{proj}_S(x - \mu_{\ell(x)})\|^2] = \sum_{j=1}^{k-1} p_j^T \Sigma p_j,$$

where $\{p_j\}$ is an orthonormal basis for S.

By Fact 2, one minimizing subspace is the span of the smallest $k-1$ eigenvectors of the matrix Σ, i.e. $v_{n-k+2}, \ldots, v_n$. Because the distribution is isotropic,

$$\Sigma = I - \sum_{i=1}^{k} \mathrm{w}_i \mu_i \mu_i^T.$$

and these vectors become the largest eigenvectors of $\sum_{i=1}^{k} \mathrm{w}_i \mu_i \mu_i^T$. Clearly, $\mathrm{span}\{v_{n-k+2}, \ldots, v_n\} \subseteq \mathrm{span}\{\mu_1, \ldots, \mu_k\}$, but both spans have dimension $k-1$ making them equal. This also implies that

$$1 - \lambda_{n-k+2}(\Sigma) = v_{n-k+2}^T \sum_{i=1}^{k} \mathrm{w}_i \mu_i \mu_i^T v_{n-k+2} > 0.$$

Thus, $\lambda_{n-k+2}(\Sigma) < 1$. On the other hand v_{n-k+1}, must be orthogonal every μ_i, so $\lambda_{n-k+1}(\Sigma) = 1$. Therefore, $\lambda_{n-k+1}(\Sigma) > \lambda_{n-k+2}(\Sigma)$ and by Fact 2 $\mathrm{span}\{v_{n-k+2}, \ldots, v_n\} = \mathrm{span}\{\mu_1, \ldots, \mu_k\}$ is the unique minimizing subspace. □

Under the conditions of Lemma 1, the overlap may be characterized as

$$\phi = \lambda_{n-k+2}(\Sigma) = 1 - \lambda_{k-1}\left(\sum_{i=1}^{k} \mathrm{w}_i \mu_i \mu_i^T\right).$$

For clarity of the analysis, we will assume that Step 1 of the algorithm produces a perfectly isotropic mixture. The required number of samples to make the distribution nearly isotropic is no larger than for other aspects of the algorithm, and as our analysis shows, the algorithm is robust to small estimation errors.

We will also assume for convenience of notation that the the unit vectors along the first $k-1$ coordinate axes $e_1, \ldots e_{k-1}$ span the intermean (i.e. Fisher) subspace. That is, $F = \mathrm{span}\{e_1, \ldots, e_{k-1}\}$. When considering this subspace it will be convenient to be able to refer to projection of the mean vectors to this subspace. Thus, we define $\tilde{\mu_i} \in R^{k-1}$ to be the first $k-1$ coordinates of μ_i; the remaining coordinates are all zero. In other terms,

$$\tilde{\mu_i} = \begin{bmatrix} I_{k-1} & 0 \end{bmatrix} \mu_i \ .$$

In this coordinate system the covariance matrix of each component has a particular structure, which will be useful for our analysis. For the rest of this paper we fix the following notation: an isotropic mixture is defined by $\{\mathrm{w}_i, \mu_i, \Sigma_i\}$. We assume that $\mathrm{span}\{e_1, \ldots, e_{k-1}\}$ is the intermean subspace and A_i,B_i, and D_i are defined such that

$$\mathrm{w}_i \Sigma_i = \begin{bmatrix} A_i & B_i^T \\ B_i & D_i \end{bmatrix} \tag{2}$$

where A_i is a $(k-1) \times (k-1)$ submatrix and D_i is a $(n-k+1) \times (n-k+1)$ submatrix.

Lemma 4 (Covariance Structure). *Using the above notation,*

$$\|A_i\| \leq \phi \ , \|D_i\| \leq 1 \ , \|B_i\| \leq \sqrt{\phi}$$

for all components i.

For small ϕ, the covariance between intermean and non-intermean directions, i.e. B_i, is small. For $k = 2$, this means that all densities will have a "nearly parallel pancake" shape. In general, it means that $k-1$ of the principal axes of the Gaussians will lie close to the intermean subspace.

4 Finding a Vector near the Fisher Subspace

In this section, show that the direction h chosen by step 3 of the algorithm is close to the intermean subspace. Section 5 argues that this direction can be used to partition the components. Finding the separating direction is the most challenging part of the classification task and represents the main contribution of this work.

The direction h is chosen either based on the reweighted sample mean $\hat{u}$ or the reweighted sample second moment matrix $\hat{M}$. In expectation, these quantities are

$$u \equiv E\left[x \exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right] = \sum_{i=1}^{k} \mathrm{w}_i \rho_i \mu_i - \frac{1}{\alpha} \sum_{i=1}^{k} \mathrm{w}_i \rho_i \Sigma_i \mu_i + f \tag{3}$$

and

$$M \equiv E\left[x x^T \exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right] = \sum_{i=1}^{k} \mathrm{w}_i \rho_i (\Sigma_i + \mu_i \mu_i^T - \frac{1}{\alpha}(\Sigma_i \Sigma_i + \mu_i \mu_i^T \Sigma_i + \Sigma_i \mu_i \mu_i^T)) + F \tag{4}$$

respectively, where $\rho_i = E_i[\exp\left(-\frac{\|x\|^2}{2\alpha}\right)]$ (expectation being taken with respect to the ith component) and $\|f\|$ and $\|F\|$ are $O(\alpha^{-2})$. Note that to ensure the average of the ρ_i is at least a constant requires $\alpha = \Omega(n/\mathrm{w})$.

To see how these are useful, we first assume zero overlap and that the sample reweighted moments behave exactly according to expectation. Zero overlap implies that the B_i submatrix of Lemma 4 is zero, and therefore terms involving $\mu_i \Sigma_i$ vanish. In this case, the mean shift u becomes

$$v \equiv \sum_{i=1}^{k} \mathrm{w}_i \rho_i \mu_i.$$

We can intuitively think of the components that have greater ρ_i as gaining mixing weight and those with

smaller ρ_i as losing mixing weight. As long as the ρ_i are not all equal, we will observe some shift of the mean in the intermean subspace, i.e. Fisher subspace. Therefore, we may use this direction to partition the components.

On the other hand, if all of the ρ_i are equal and the overlap is zero, then M becomes

$$\begin{aligned}\Gamma &\equiv \sum_{i=1}^{k} \rho_i \begin{bmatrix} \mathrm{w}_i \tilde{\mu}_i \tilde{\mu}_i^T + A_i & 0 \\ 0 & D_i - \frac{\rho_i}{\mathrm{w}_i \alpha} D_i^2 \end{bmatrix} \quad (5) \\ &= \bar{\rho} \begin{bmatrix} I & 0 \\ 0 & I - \frac{1}{\alpha} \sum_{i=1}^{k} \frac{1}{\mathrm{w}_i} D_i^2 \end{bmatrix}.\end{aligned}$$

Notice that the second moments in the subspace $\mathrm{span}\{e_1, \ldots, e_{k-1}\}$ are maintained while those in the complementary subspace are reduced by $\mathrm{poly}(1/\alpha)$. Therefore, the top eigenvector will be in the intermean subspace, which is the Fisher subspace.

We now argue that this same strategy can be adapted to work in general, i.e., with nonzero overlap and sampling errors, with high probability. A critical aspect of this argument is that the norm of the error term $\hat{M} - \Gamma$ depends only on ϕ and k and not the dimension of the data. See Lemma 9 and the supporting Lemma 4 and Fact 3.

Since we cannot know directly how imbalanced the ρ_i are, we choose the method of finding a separating direction according the norm of the vector $\|\hat{u}\|$. Recall that when $\|\hat{u}\| > \sqrt{\mathrm{w}}/(32\alpha)$ the algorithm uses $\hat{u}$ to determine the separating direction h. Lemma 5 guarantees that this vector is close to the Fisher subspace. When $\|\hat{u}\| \leq \sqrt{\mathrm{w}}/(32\alpha)$, the algorithm uses the top eigenvector of the covariance matrix $\hat{M}$. Lemma 6 guarantees that this vector is close to the Fisher subspace.

Lemma 5 (Mean Shift Method). *Let $\epsilon > 0$. There exists a constant C such that if $m_1 \geq Cn^4 \mathrm{poly}(k, \mathrm{w}^{-1}, \log n/\delta)$, then the following holds with probability $1 - \delta$. If $\|\hat{u}\| > \sqrt{\mathrm{w}}/(32\alpha)$ and*

$$\phi \leq \frac{\mathrm{w}^2 \epsilon}{2^{14} k^2},$$

then

$$\frac{\|\hat{u}^T v\|}{\|\hat{u}\| \|v\|} \geq 1 - \epsilon.$$

Lemma 6 (Spectral Method). *Let $\epsilon > 0$. There exists a constant C such that if $m_1 \geq Cn^4 \mathrm{poly}(k, \mathrm{w}^{-1}, \log n/\delta)$, then the following holds with probability $1 - \delta$. Let $v_1, \ldots, v_{k-1}$ be the top $k-1$ eigenvectors of $\hat{M}$. If $\|\hat{u}\| \leq \sqrt{\mathrm{w}}/(32\alpha)$ and*

$$\phi \leq \frac{\mathrm{w}^2 \epsilon}{640^2 k^2}$$

then

$$\min_{v \in \mathrm{span}\{v_1, \ldots, v_{k-1}\}, \|v\|=1} \|\mathrm{proj}_F(v)\| \geq 1 - \epsilon.$$

The proof of Lemma 5 consists of bounding the difference between u and v using the smallness of ϕ and between u and $\hat{u}$ using sample convergence. Details can be found in the full version of this paper. Here we focus on Lemma 6.

4.1 Spectral Method

The following lemmas are used to prove Lemma 6. We first show that the smallness of the mean shift $\hat{u}$ implies that the coefficients ρ_i are sufficiently uniform to allow us to apply the spectral method.

Claim 7. *If $\|\hat{u}| \leq \sqrt{\mathrm{w}}/(32\alpha)$ and*

$$\sqrt{\phi} \leq \frac{\mathrm{w}}{64k},$$

then

$$\|\rho - 1\bar{\rho}\|_2 \leq \frac{1}{8\alpha}.$$

Next, we establish important properties of the matrix Γ defined in Eqn. 5.

Lemma 8 (Ideal Case). *If $\|\rho - 1\bar{\rho}\|_\infty \leq 1/(8\alpha)$, then*

$$\lambda_{k-1}(\Gamma) - \lambda_k(\Gamma) \geq \frac{1}{4\alpha},$$

and the top $k-1$ eigenvectors of Γ span the means of the components.

Next, we show that Γ is close to $\hat{M}$ be the following two lemmas.

Lemma 9. *If $\|\rho - 1\bar{\rho}\|_\infty < 1/(2\alpha)$, then*

$$\|M - \Gamma\|_2^2 \leq \frac{16^2 k^2}{\mathrm{w}^2 \alpha^2} \phi.$$

Lemma 10. *Let $\epsilon, \delta > 0$ and let $\hat{M}$ be the reweighted sample matrix of second moments for a set of m points drawn from an isotropic mixture of k Gaussians in n dimensions, where*

$$m \geq C_1 \frac{n\alpha}{\epsilon^2} \log \frac{n\alpha}{\delta}.$$

and C_1 is an absolute constant. Then

$$\mathsf{P}\left[\left\|\hat{M} - M\right\| > \epsilon\right] < \delta.$$

Finally, we state Stewart's Lemma, which shows that if there is a large eigenvalue gap at the rth eigenvalue of a matrix, then the top r dimensional subspace is preserved under small perturbations to the matrix.

Lemma 11 (Stewart's Theorem). *Suppose A and $A + E$ are n-by-n symmetric matrices and that*

$$A = \begin{array}{c} \left[\begin{array}{cc} D_1 & 0 \\ 0 & D_2 \end{array} \right] \\ \begin{array}{cc} r & n-r \end{array} \end{array} \quad \begin{array}{c} r \\ n-r \end{array}$$

$$E = \begin{array}{c} \left[\begin{array}{cc} E_{11} & E_{21}^T \\ E_{21} & E_{22} \end{array} \right] \\ \begin{array}{cc} r & n-r \end{array} \end{array} \quad \begin{array}{c} r \\ n-r \end{array} .$$

Let the columns of V be the top r eigenvectors of the matrix $A + E$ and let P_2 be the matrix with columns $e_{r+1}, \ldots, e_n$. If $d = \lambda_r(D_1) - \lambda_1(D_2) > 0$ and

$$\|E\| \leq \frac{d}{5},$$

then

$$\|V^T P_2\| \leq \frac{4}{d}\|E_{21}\|_2.$$

Proof of Lemma 6. To bound the effect of overlap and sample errors on the eigenvectors, we apply Stewart's Lemma (Lemma 11). Define $d = \lambda_{k-1}(\Gamma) - \lambda_k(\Gamma)$ and $E = \hat{M} - \Gamma$.

We assume that the mean shift satisfies $\|\hat{u}\| \leq \sqrt{\mathrm{w}}/(32\alpha)$ and that ϕ is small. By Lemma 8, this implies that

$$d = \lambda_{k-1}(\Gamma) - \lambda_k(\Gamma) \geq \frac{1}{4\alpha}. \tag{6}$$

To bound $\|E\|$, we use the triangle inequality $\|E\| \leq \|\Gamma - M\| + \|M - \hat{M}\|$. Lemma 9 bounds the first term by

$$\|M - \Gamma\| \leq \sqrt{\frac{16^2k^2}{\mathrm{w}^2\alpha^2}\phi} = \sqrt{\frac{16^2k^2}{\mathrm{w}^2\alpha^2} \cdot \frac{\mathrm{w}^2\epsilon}{640^2k^2}} \leq \frac{1}{40\alpha}\sqrt{\epsilon}.$$

By Lemma 10, we obtain the same bound on $\|M - \hat{M}\|$ with probability $1 - \delta$ for large enough m_1. Thus,

$$\|E\| \leq \frac{1}{20\alpha}\sqrt{\epsilon}.$$

Combining the bounds of Eqn. 6 and 4.1, we have

$$\sqrt{1-(1-\epsilon)^2}d - 5\|E\| \geq \sqrt{1-(1-\epsilon)^2}\frac{1}{4\alpha} - 5\frac{1}{20\alpha}\sqrt{\epsilon} \geq 0, \tag{7}$$

as $\sqrt{1-(1-\epsilon)^2} \geq \sqrt{\epsilon}$. This implies both that $\|E\| \leq d/5$ and that $4\|E_{21}|/d < \sqrt{1-(1-\epsilon)^2}$, enabling us to apply Stewart's Lemma to the matrix pair Γ and $\hat{M}$.

By Lemma 8, the top $k-1$ eigenvectors of Γ, i.e. $e_1, \ldots, e_{k-1}$, span the means of the components. Let the columns of P_1 be these eigenvectors. Let the columns of P_2 be defined such that $[P_1, P_2]$ is an orthonormal matrix and let $v_1, \ldots, v_k$ be the top $k-1$ eigenvectors of $\hat{M}$. By Stewart's Lemma, letting the columns of V be $v_1, \ldots, v_{k-1}$, we have

$$\|V^T P_2\|_2 \leq \sqrt{1-(1-\epsilon)^2},$$

or equivalently,

$$\min_{v \in \mathrm{span}\{v_1,\ldots,v_{k-1}\}, \|v\|=1} \|\mathrm{proj}_F v\| = \sigma_{k-1}(V^T P_1) \geq 1 - \epsilon. \tag{8}$$

□

5 Recursion

In this section, we show the for every direction h that is close to the intermean subspace, the "largest gap clustering" step produces a pair of complementary halfspaces that partitions $\mathbb{R}^n$ while leaving only a small part of the probability mass on the wrong side of the partition, small enough that with high probability, it does not affect the samples used by the algorithm.

Lemma 12. *Let $\delta, \delta' > 0$, where $\delta' \leq \delta/(2m_2)$, and let m_2 satisfy $m_2 \geq n/k \log(2k/\delta)$. Suppose that h is a unit vector such that*

$$\|\mathrm{proj}_F(h)\| \geq 1 - \frac{\mathrm{w}}{2^{10}(k-1)^2 \log \frac{1}{\delta'}}.$$

Let $\mathcal{F}$ be a mixture of $k > 1$ Gaussians with overlap

$$\phi \leq \frac{\mathrm{w}}{2^9(k-1)^2} \log^{-1} \frac{1}{\delta'}.$$

Let X be a collection of m_2 points from $\mathcal{F}$ and let t be the midpoint of the largest gap in set $\{h^T x : x \in X\}$. With probability $1 - \delta$, the halfspace $H_{h,t}$ has the following property. For a random sample y from $\mathcal{F}$ either

$$y, \mu_{\ell(y)} \in H_{h,t} \text{ or } y, \mu_{\ell(y)} \notin H_{h,t}$$

with probability $1 - \delta'$.

The idea behind the proof is simple. We first show that two of the means are at least a constant distance

apart. We then bound the width of a component along the direction h, i.e. the maximum distance between two points belonging to the same component. If the width of each component is small, then clearly the largest gap must fall between components. Setting t to be the midpoint of the gap, we avoid cutting any components.

In the proof of the main theorem for large k, we will need to have every point sampled from $\mathcal{F}$ in the recursion subtree classified correctly by the halfspace, so we will assume δ' considerably smaller than m_2/δ.

The second lemma shows that all submixtures have smaller overlap to ensure that all the relevant lemmas apply in the recursive steps.

Lemma 13. *The removal of any subset of components cannot induce a mixture with greater overlap than the original.*

The proofs of the main theorems are now apparent. Consider the case of $k = 2$ Gaussians first. Using $m_1 = \omega(kn^4 \mathrm{w}^{-3} \log(n/\delta \mathrm{w}))$ samples to estimate $\hat{u}$ and $\hat{M}$ is sufficient to guarantee that the estimates are accurate. For a well-chosen constant C, the condition

$$\phi \leq J(p) \leq C\mathrm{w}^3 \log^{-1}\left(\frac{1}{\delta \mathrm{w}} + \frac{1}{\eta}\right)$$

of Theorem 2 implies that

$$\sqrt{\phi} \leq \frac{\mathrm{w}\sqrt{\epsilon}}{640 \cdot 2},$$

where

$$\epsilon = \frac{\mathrm{w}}{2^9} \log^{-1}\left(\frac{2m_2}{\delta} + \frac{1}{\eta}\right).$$

The arguments of Section 4 then show that the direction h selected in step 3 satisfies

$$\|P_1^T h\| \geq 1 - \epsilon = 1 - \frac{\mathrm{w}}{2^9} \log^{-1}\left(\frac{m_2}{\delta} + \frac{1}{\eta}\right).$$

Already, for the overlap we have

$$\sqrt{\phi} \leq \frac{\mathrm{w}\sqrt{\epsilon}}{640 \cdot 2} \leq \sqrt{\frac{\mathrm{w}}{2^9(k-1)^2}} \log^{-1/2} \frac{1}{\delta'}.$$

so we may apply Lemma 12 with $\delta' = (m_2/\delta + 1/\eta)^{-1}$. Thus, with probability $1-\delta$ the classifier $H_{h,t}$ is correct with probability $1 - \delta' \geq 1 - \eta$.

We follow the same outline for $k > 2$, with the quantity $1/\delta' = m_2/\delta + 1/\eta$ being replaced with $1/\delta' = m/\delta + 1/\eta$, where m is the total number of samples used. This is necessary because the half-space $H_{h,t}$ must classify every sample point taken below it in the recursion subtree correctly. This adds the n and k factors so that the required overlap becomes

$$\phi \leq C\mathrm{w}^3 k^{-3} \log^{-1}\left(\frac{nk}{\delta \mathrm{w}} + \frac{1}{\eta}\right)$$

for an appropriate constant C. The correctness in the recursive steps is guaranteed by Lemma 13. Assuming that all previous steps are correct, the termination condition of step 4 is clearly correct when a single component is isolated.

6 Conclusion

We have presented an affine-invariant extension of principal components. We expect that this technique should be applicable to a broader class of problems. For example, mixtures of distributions with some mild properties such as center symmetry and some bounds on the first few moments might be solvable using isotropic PCA. It would be nice to characterize the full scope of the technique for clustering and also to find other applications, given that standard PCA is widely used.

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. Learning mixtures of product distributions using correlations and independence. In *Proc. of COLT*, 2008.

[4] A. Dasgupta, J. Hopcroft, J. Kleinberg, and M. Sandler. On learning mixtures of heavy-tailed distributions. In *Proc. of FOCS*, 2005.

[5] S. DasGupta. Learning mixtures of gaussians. In *Proc. of FOCS*, 1999.

[6] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[7] R. O. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, 2001.

[8] 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.

[9] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[10] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proceedings of the 18th Conference on Learning Theory*. University of California Press, 2005.

[11] 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. University of California Press, 1967.

[12] R. K. S. Arora. Learning mixtures of arbitrary gaussians. *Ann. Appl. Probab.*, 15(1A):69–92, 2005.

[13] L. S. S. DasGupta. A two-round variant of em for gaussian mixtures. In *Sixteenth Conference on Uncertainty in Artificial Intelligence*, 2000.

[14] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. *Proc. of FOCS 2002; JCCS*, 68(4):841–860, 2004.