Isotropic PCA and Affine-Invariant Clustering

S. Charles Brubaker* Santosh S. Vempala*

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 [14, 9, 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. Here we define overlap to be the ratio of the following two quantities: 1) the average squared distance between a point and the mean of its component, and 2) the average squared distance between a point and the mean of the mixture. The main result may also be stated in the language of linear discriminant analysis: if the standard Fisher discriminant [8] is small enough, labels are not needed to estimate the optimal subspace for projection. Our main tools are isotropic transformation, spectral projection and a simple reweighting technique. We call this combination *isotropic PCA*.

*College of Computing, Georgia Tech. Email: {brubaker,vempala}@cc.gatech.edu

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. When applied to the problem of unraveling mixtures of arbitrary Gaussians from unlabeled samples, the algorithm yields substantial improvements of known results.

To illustrate the technique, consider the uniform distribution on the set $X = \{(x, y) \in \mathbb{R}^2 : x \in \{-1, 1\}, y \in [-\sqrt{3}, \sqrt{3}]\}$, which is isotropic. Suppose this distribution is rotated in an unknown way and that we would like to recover the original x and y axes. For each point in a sample, we may project it to the unit circle and compute the covariance matrix of the resulting point set. The x direction will correspond to the greater eigenvector, the y direction to the other. See Figure 1 for an illustration. Instead of projection onto the unit circle, this process may also be thought of as importance weighting, a technique which allows one to simulate one distribution with another. In this case, we are simulating a distribution over the set X, where the density function is proportional to $(1 + y^2)^{-1}$, so that points near $(1, 0)$ or $(-1, 0)$ are more probable.

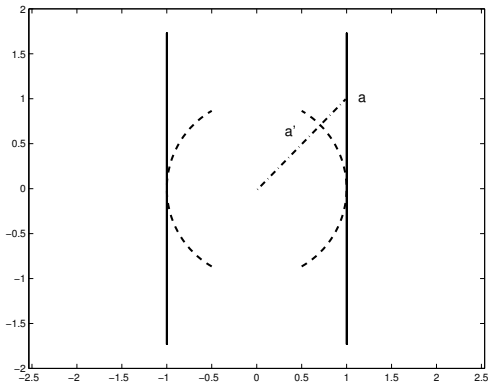


Figure 1: Mapping points to the unit circle and then finding the direction of maximum variance reveals the orientation of this isotropic distribution.

In this paper, we describe how to apply this method to mixtures of arbitrary Gaussians in $\mathbb{R}^n$ in order to find 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.

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. Because our algorithm is affine invariant, it makes it possible to unravel a much larger set of Gaussian mixtures than had

been possible previously.

The first step of our algorithm is to place the mixture in isotropic position (see Section 1.2) via an affine transformation. This has the effect of making the $(k-1)$-dimensional Fisher subspace, i.e., the one that minimizes the Fisher discriminant, the same as the subspace spanned by the means of the components (they only coincide in general in isotropic position), for *any* mixture. The rest of the algorithm identifies directions close to this subspace and uses them to cluster, without access to labels. Intuitively this is hard since after isotropy, standard PCA reveals no additional information. Before presenting the ideas and guarantees in more detail, we describe relevant related work.

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 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.

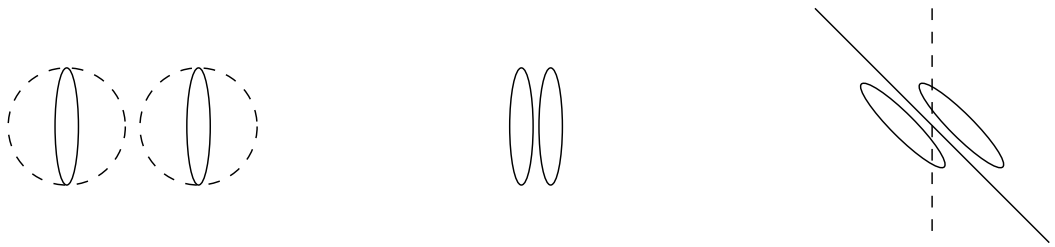
Heuristics for classifying samples include "expectation maximization" [5] and "k-means clustering" [11]. These methods can take a long time and can get stuck with suboptimal classifications. Over the past decade, there has been much progress on finding polynomial-time algorithms with rigorous guarantees for classifying mixtures, especially mixtures of Gaussians [4, 15, 14, 17, 9, 1]. Starting with Dasgupta's paper [4], 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 [14] 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 [17] 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

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

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. [9] applied the spectral approach to arbitrary mixtures of Gaussians (and more generally, logconcave distributions)



(a) Distance Concentration Separability

(b) Hyperplane Separability

(c) Intermean Hyperplane and Fisher Hyperplane.

Figure 2: 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).

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 $\sigma^2_{i,\max}$ 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 [6] 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 [3] 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 [2] have recently given a polynomial time algorithm for clustering such heavy tailed product distributions.

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 total 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 2(a). The dotted line shows how previous work effectively treats every component as spherical. Hyperplane separability (Figure 2(b)) is a weaker condition. We also note that the separating direction does not need to be the intermean direction as illustrated in Figure 2(c). The dotted line illustrates hyperplane induced by the intermean direction, which may be far from the optimal separating hyperplane shown by the solid line.

It will be insightful to state this result in terms of the Fisher discriminant, a standard notion from Pattern Recognition [8, 7] that is used with labeled data. In words, the Fisher discriminant along direction p is

$$J(p) = \frac{\text{the intra-component variance in direction } p}{\text{the total variance in direction } p}$$

Mathematically, this is expressed as

$$J(p) = \frac{E\left[\|\mathrm{proj}_p(x - \mu_{\ell(x)})\|^2\right]}{E\left[\|\mathrm{proj}_p(x)\|^2\right]} = \frac{p^T(\mathrm{w}_1\Sigma_1 + \mathrm{w}_2\Sigma_2)p}{p^T(\mathrm{w}_1(\Sigma_1 + \mu_1\mu_1^T) + \mathrm{w}_2(\Sigma_2 + \mu_2\mu_2^T))p}$$

for x distributed according to a mixture distribution with means μ_i and covariance matrices Σ_i. We use $\ell(x)$ to indicate the component from which x was drawn.

Theorem 2. *There is an absolute constant* C *for which the following holds. Suppose that* $\mathcal{F}$ *is a mixture of two Gaussians such that there exists a direction* p *for which*

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

With probability $1-\delta$*, algorithm* UNRAVEL *returns a halfspace with error at most* η *using time and sample complexity polynomial in* $n, \mathrm{w}^{-1}, \log(1/\delta)$.

There are several ways of generalizing the Fisher discriminant for $k = 2$ components to greater k [7]. These generalizations are most easily understood when the distribution is isotropic. An isotropic distribution has the identity matrix as its covariance and the origin as its mean. An isotropic mixture therefore has

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

It is well known that any distribution with bounded covariance matrix (and therefore any mixture) can be made isotropic by an affine transformation. As we will see shortly, for $k = 2$, for an isotropic mixture, the line joining the means is the direction that minimizes the Fisher discriminant.

Under isotropy, the denominator of the Fisher discriminant is always 1. Thus, the discriminant is just the expected squared distance between the projection of a point and the projection of its mean, where projection is onto some direction p. The generalization to $k > 2$ is natural, as we may simply replace projection onto direction p with projection onto a $(k-1)$-dimensional subspace S. For convenience, let

$$\Sigma = \sum_{i=1}^{k} \mathrm{w}_i \Sigma_i.$$

Let the vector $p_1, \ldots, p_{k-1}$ be an orthonormal basis of S and let $\ell(x)$ be the component from which x was drawn. We then have under isotropy

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

for x distributed according to a mixture distribution with means μ_i and covariance matrices Σ_i. As Σ is symmetric positive definite, it follows that the smallest $k-1$ eigenvectors of the matrix are optimal choices of p_j and S is the span of these eigenvectors.

This motivates our definition of the Fisher subspace for *any* mixture with bounded second moments (not necessarily Gaussians).

Definition 1. *Let* $\{\mathrm{w}_i, \mu_i, \Sigma_i\}$ *be the weights, means, and covariance matrices for an isotropic* [1] *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). The proof is given in Section 3.2.

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 attempts to find the Fisher subspace (or one close to it) and succeeds in doing so, provided the discriminant is small enough.

The next definition will be useful in stating our main theorem precisely.

[1] For non-isotropic mixtures, the Fisher discriminant generalizes to $\sum_{j=1}^{k-1} p_j^T \left(\sum_{i=1}^{k} \mathrm{w}_i(\Sigma_i + \mu_i \mu_i^T)\right)^{-1} \Sigma p_j$ and the overlap to $p^T \left(\sum_{i=1}^{k} \mathrm{w}_i(\Sigma_i + \mu_i \mu_i^T)\right)^{-1} \Sigma p$

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

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

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. [2]

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 $H_{h,t} = \{x \in \mathbb{R}^n : h^T x \ge t\}$, to indicate a halfspace, we then recurse

[2] This practice of transforming the points and then looking at the second moment matrix can be viewed as a form of kernel PCA; however the connection between our algorithm and kernel PCA is superficial. Our transformation does not result in any standard kernel. Moreover, it is dimension-preserving (it is just a reweighting), and hence the "kernel trick" has no computational advantage.

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.

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 covariance 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.

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 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.

2.2 Overview of Analysis

To analyze the algorithm, in the general case, we will proceed as follows. Section 3 shows that under isotropy the Fisher subspace coincides with the intermean subspace (Lemma 1), gives the necessary sampling convergence and perturbation lemmas and relates overlap to a more conventional notion of separation (Prop. 5). Section 3.3 gives approximations to the first and second moments. Section 4 then combines these approximations with the perturbation lemmas to show that the vector h (either the mean shift or the largest principal component) lies close to the intermean subspace. Finally, Section 5 shows the correctness of the recursive aspects of the algorithm.

3 Preliminaries

3.1 Matrix Properties

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 $\|B\| \leq \sqrt{\|A\|\|D\|}$.

Proof. Let y and x be the top left and right singular vectors of B, so that $y^T B x = \|B\|$. Because Z is positive semi-definite, we have that for any real γ,

$$0 \leq [\gamma x^T \; y^T] Z [\gamma x^T \; y^T]^T = \gamma^2 x^T A x + 2\gamma y^T B x + y^T D y.$$

This is a quadratic polynomial in γ that can have only one real root. Therefore the discriminant must be non-positive:

$$0 \geq 4(y^T B x)^2 - 4(x^T A x)(y^T D y).$$

We conclude that

$$\|B\| = y^T B x \leq \sqrt{(x^T A x)(y^T D y)} \leq \sqrt{\|A\|\|D\|}.$$

□

3.2 The Fisher Criterion and Isotropy

We begin with the proof of the lemma that for an isotropic mixture the Fisher subspace is the same as the intermean subspace.

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

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

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, $\text{span}\{v_{n-k+2}, \ldots, v_n\} \subseteq \text{span}\{\mu_1, \ldots, \mu_k\}$, 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 $\text{span}\{v_{n-k+2}, \ldots, v_n\} = \text{span}\{\mu_1, \ldots, \mu_k\}$ is the unique minimizing subspace. □

It follows directly that 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. Theorem 4 gives a bound on the required number of samples to make the distribution nearly isotropic, and as our analysis shows, our 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, $F = \text{span}\{e_1, \ldots, e_{k-1}\}$. 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 $\text{span}\{e_1, \ldots, e_{k-1}\}$ is the intermean subspace and A_i,B_i, and D_i are defined such that

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

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.

Proof of Lemma 4. Because $\text{span}\{e_1, \ldots, e_{k-1}\}$ is the Fisher subspace

$$\phi = \max_{v \in \mathbb{R}^{k-1}} \frac{1}{\|v\|^2} \sum_{i=1}^{k} v^T A_i v = \left\| \sum_{i=1}^{k} A_i \right\|_2 .$$

Also $\sum_{i=1}^{k} D_i = I$, so $\| \sum_{i=1}^{k} D_i \| = 1$. Each matrix $\mathrm{w}_i \Sigma_i$ is positive definite, so the principal minors A_i,D_i must be positive definite as well. Therefore, $\|A_i\| \leq \phi$, $\|D_i\| \leq 1$, and $\|B_i\| \leq \sqrt{\|A_i\|\|D_i\|} = \sqrt{\phi}$ using Fact 3. □

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.

We conclude this section with a proposition connecting, for $k = 2$, the overlap to a standard notion of separation between two distributions, so that Theorem 1 becomes an immediate corollary of Theorem 2.

Proposition 5. *If there exists a unit vector p such that*

$$|p^T(\mu_1 - \mu_2)| > t(\sqrt{p^T \mathrm{w}_1 \Sigma_1 p} + \sqrt{p^T \mathrm{w}_2 \Sigma_2 p}),$$

then the overlap $\phi \leq J(p) \leq (1 + \mathrm{w}_1 \mathrm{w}_2 t^2)^{-1}$.

Proof of Proposition 5. Since the mean of the distribution is at the origin, we have $\mathrm{w}_1 p^T \mu_1 = -\mathrm{w}_2 p^T \mu_2$. Thus,

$$\begin{aligned} |p^T\mu_1 - p^T\mu_2|^2 &= (p^T\mu_1)^2 + (p^T\mu_2)^2 + 2|p^T\mu_1||p^T\mu_2| \\ &= (\mathrm{w}_1 p^T\mu_1)^2 \left(\frac{1}{\mathrm{w}_1^2} + \frac{1}{\mathrm{w}_2^2} + \frac{2}{\mathrm{w}_1\mathrm{w}_2} \right), \end{aligned}$$

using $\mathrm{w}_1 + \mathrm{w}_2 = 1$. We rewrite the last factor as

$$\frac{1}{\mathrm{w}_1^2} + \frac{1}{\mathrm{w}_2^2} + \frac{2}{\mathrm{w}_1\mathrm{w}_2} = \frac{\mathrm{w}_1^2 + \mathrm{w}_2^2 + 2\mathrm{w}_1\mathrm{w}_2}{\mathrm{w}_1^2\mathrm{w}_2^2} = \frac{1}{\mathrm{w}_1^2\mathrm{w}_2^2} = \frac{1}{\mathrm{w}_1\mathrm{w}_2} \left(\frac{1}{\mathrm{w}_1} + \frac{1}{\mathrm{w}_2} \right).$$

Again, using the fact that $\mathrm{w}_1 p^T \mu_1 = -\mathrm{w}_2 p^T \mu_2$, we have that

$$\begin{aligned} |p^T\mu_1 - p^T\mu_2|^2 &= \frac{(\mathrm{w}_1 p^T\mu_1)^2}{\mathrm{w}_1\mathrm{w}_2} \left(\frac{1}{\mathrm{w}_1} + \frac{1}{\mathrm{w}_2} \right) \\ &= \frac{\mathrm{w}_1 (p^T\mu_1)^2 + \mathrm{w}_2 (p^T\mu_2)^2}{\mathrm{w}_1\mathrm{w}_2}. \end{aligned}$$

Thus, by the separation condition

$$\mathrm{w}_1 (p^T\mu_1)^2 + \mathrm{w}_2 (p^T\mu_2)^2 = \mathrm{w}_1\mathrm{w}_2 |p^T\mu_1 - p^T\mu_2|^2 \geq \mathrm{w}_1\mathrm{w}_2 t^2 (p^T \mathrm{w}_1 \Sigma_1 p + p^T \mathrm{w}_2 \Sigma_2 p).$$

To bound $J(p)$, we then argue

$$\begin{aligned} J(p) &= \frac{p^T \mathrm{w}_1 \Sigma_1 p + p^T \mathrm{w}_2 \Sigma_2 p}{\mathrm{w}_1(p^T\Sigma_1 p + (p^T\mu_1)^2) + \mathrm{w}_2(p^T\Sigma_2 p + (p^T\mu_2)^2)} \\ &= 1 - \frac{\mathrm{w}_1(p^T\mu_1)^2 + \mathrm{w}_2(p^T\mu_2)^2}{\mathrm{w}_1(p^T\Sigma_1 p + (p^T\mu_1)^2) + \mathrm{w}_2(p^T\Sigma_2 p + (p^T\mu_2)^2)} \\ &\leq 1 - \frac{\mathrm{w}_1\mathrm{w}_2 t^2(\mathrm{w}_1 p^T\Sigma_1 p + \mathrm{w}_2 p^T\Sigma_2 p)}{\mathrm{w}_1(p^T\Sigma_1 p + (p^T\mu_1)^2) + \mathrm{w}_2(p^T\Sigma_2 p + (p^T\mu_2)^2)} \\ &\leq 1 - \mathrm{w}_1\mathrm{w}_2 t^2 J(p), \end{aligned}$$

and $J(p) \leq 1/(1 + \mathrm{w}_1\mathrm{w}_2 t^2)$. □

3.3 Approximation of the Reweighted Moments

Our algorithm works by computing the first and second reweighted moments of a point set from $\mathcal{F}$. In this section, we examine how the reweighting affects the second moments of a single component and then give some approximations for the first and second moments of the entire mixture.

3.3.1 Single Component

The first step is to characterize how the reweighting affects the moments of a single component. Specifically, we will show for any function f (and therefore x and xx^T in particular) that for $\alpha > 0$,

$$E\left[f(x)\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right] = \sum_i \mathrm{w}_i \rho_i E_i\left[f(y_i)\right],$$

Here, $E_i[\cdot]$ denotes expectation taken with respect to the component i, the quantity $\rho_i = E_i\left[\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right]$, and y_i is a Gaussian variable with parameters slightly perturbed from the original ith component.

Claim 6. *If $\alpha = n/\mathrm{w}$, the quantity $\rho_i = E_i\left[\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right]$ is at least $1/2$.*

Proof. Because the distribution is isotropic, for any component i, $\mathrm{w}_i E_i[\|x\|^2] \leq n$. Therefore,

$$\rho_i = E_i\left[\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right] \geq E_i\left[1 - \frac{\|x\|^2}{2\alpha}\right] \geq 1 - \frac{1}{2\alpha}\frac{n}{\mathrm{w}_i} \geq \frac{1}{2}.$$

□

Lemma 7 (Reweighted Moments of a Single Component). *For any $\alpha > 0$, with respect to a single component i of the mixture*

$$E_i\left[x\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right] = \rho_i(\mu_i - \frac{1}{\alpha}\Sigma_i\mu_i + f)$$

and

$$E_i\left[xx^T\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right] = \rho(\Sigma_i + \mu_i\mu_i^T - \frac{1}{\alpha}(\Sigma_i\Sigma_i + \mu_i\mu_i^T\Sigma_i + \Sigma_i\mu_i\mu_i^T) + F)$$

where $\|f\|, \|F\| = O(\alpha^{-2})$.

We first establish the following claim.

Claim 8. *Let x be a random variable distributed according to the normal distribution $N(\mu, \Sigma)$ and let $\Sigma = Q\Lambda Q^T$ be the singular value decomposition of Σ with $\lambda_1, \ldots, \lambda_n$ being the diagonal elements of Λ. Let $W = \operatorname{diag}(\alpha/(\alpha+\lambda_1), \ldots, \alpha/(\alpha+\lambda_n))$. Finally, let y be a random variable distributed according to $N(QWQ^T\mu, QW\Lambda Q^T)$. Then for any function $f(x)$,*

$$E\left[f(x)\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right] = \det(W)^{1/2}\exp\left(-\frac{\mu^T QWQ^T\mu}{2\alpha}\right)E\left[f(y)\right].$$

Proof of Claim 8. We assume that $Q = I$ for the initial part of the proof. From the definition of a Gaussian distribution, we have

$$E\left[f(x)\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right] = \det(\Lambda)^{-1/2}(2\pi)^{-n/2}\int_{\mathbb{R}^n} f(x)\exp\left(-\frac{x^Tx}{2\alpha} - \frac{(x-\mu)^T\Lambda^{-1}(x-\mu)}{2}\right).$$

Because Λ is diagonal, we may write the exponents on the right hand side as

$$\sum_{i=1}^n x_i^2\alpha^{-1} + (x_i-\mu_i)^2\lambda_i^{-1} = \sum_{i=1}^n x_i^2(\lambda^{-1}+\alpha^{-1}) - 2x_i\mu_i\lambda_i^{-1} + \mu_i^2\lambda_i^{-1}.$$

Completing the square gives the expression

$$\sum_{i=1}^n \left(x_i - \mu_i\frac{\alpha}{\alpha+\lambda_i}\right)^2\left(\frac{\lambda_i\alpha}{\alpha+\lambda_i}\right)^{-1} + \mu_i^2\lambda_i^{-1} - \mu_i^2\lambda_i^{-1}\frac{\alpha}{\alpha+\lambda_i}.$$

The last two terms can be simplified to $\mu_i^2/(\alpha+\lambda_i)$. In matrix form the exponent becomes

$$(x - W\mu)^T(W\Lambda)^{-1}(x-W\mu) + \mu^TW\mu\alpha^{-1}.$$

For general Q, this becomes

$$\left(x - QWQ^T\mu\right)^T Q(W\Lambda)^{-1}Q^T\left(x - QWQ^T\mu\right) + \mu^TQWQ^T\mu\alpha^{-1}.$$

Now recalling the definition of the random variable y, we see

$$\begin{aligned}E\left[f(x)\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right] &= \det(\Lambda)^{-1/2}(2\pi)^{-n/2}\exp\left(-\frac{\mu^TQWQ^T\mu}{2\alpha}\right)\\ &\int_{\mathbb{R}^n} f(x)\exp\left(-\frac{1}{2}\left(x-QWQ^T\mu\right)^TQ(W\Lambda)^{-1}Q^T\left(x-QWQ^T\mu\right)\right)\\ &= \det(W)^{1/2}\exp\left(-\frac{\mu^TQWQ^T\mu}{2\alpha}\right)E\left[f(y)\right].\end{aligned}$$

□

The proof of Lemma 7 is now straightforward.

Proof of Lemma 7. For simplicity of notation, we drop the subscript i from ρ_i, μ_i, Σ_i with the understanding that all statements of expectation apply to a single component. Using the notation of Claim 8, we have

$$\rho = E\left[\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right] = \det(W)^{1/2}\exp\left(-\frac{\mu^TQWQ^T\mu}{2\alpha}\right).$$

A diagonal entry of the matrix W can expanded as

$$\frac{\alpha}{\alpha+\lambda_i} = 1 - \frac{\lambda_i}{\alpha+\lambda_i} = 1 - \frac{\lambda_i}{\alpha} + \frac{\lambda_i^2}{\alpha(\alpha+\lambda_i)},$$

so that

$$W = I - \frac{1}{\alpha}\Lambda + \frac{1}{\alpha^2}W\Lambda^2.$$

Thus,

$$\begin{aligned} E\left[x\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right] &= \rho(QWQ^T\mu) \\ &= \rho(QIQ^T\mu - \frac{1}{\alpha}Q\Lambda Q^T\mu + \frac{1}{\alpha^2}QW\Lambda^2Q^T\mu) \\ &= \rho(\mu - \frac{1}{\alpha}\Sigma\mu + f), \end{aligned}$$

where $\|f\| = O(\alpha^{-2})$.

We analyze the perturbed covariance in a similar fashion.

$$\begin{aligned} E\left[xx^T\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right] &= \rho\left(Q(W\Lambda)Q^T + QWQ^T\mu\mu^TQWQ^T\right) \\ &= \rho\Bigg(Q\Lambda Q^T - \frac{1}{\alpha}Q\Lambda^2Q^T + \frac{1}{\alpha^2}QW\Lambda^3Q^T \\ &\quad +(\mu - \frac{1}{\alpha}\Sigma\mu + f)(\mu - \frac{1}{\alpha}\Sigma\mu + f)^T\Bigg) \\ &= \rho\left(\Sigma + \mu\mu^T - \frac{1}{\alpha}(\Sigma\Sigma + \mu\mu^T\Sigma + \Sigma\mu\mu^T) + F\right), \end{aligned}$$

where $\|F\| = O(\alpha^{-2})$. □

3.3.2 Mixture moments

The second step is to approximate the first and second moments of the entire mixture distribution. Let ρ be the vector where $\rho_i = E_i\left[\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right]$ and let $\bar{\rho}$ be the average of the ρ_i. We also define

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

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

with $\|f\| = O(\alpha^{-2})$ and $\|F\| = O(\alpha^{-2})$. We denote the estimates of these quantities computed from samples by $\hat{u}$ and $\hat{M}$ respectively.

Lemma 9. *Let* $v = \sum_{i=1}^k \rho_i\mathrm{w}_i\mu_i$. *Then*

$$\|u - v\|^2 \leq \frac{4k^2}{\alpha^2\mathrm{w}}\phi.$$

Proof of Lemma 9. We argue from Eqn. 2 and Eqn. 3 that

$$\begin{aligned}\|u-v\| &= \frac{1}{\alpha}\left\|\sum_{i=1}^{k} \mathrm{w}_i\rho_i\Sigma_i\mu_i\right\| + O(\alpha^{-2}) \\ &\leq \frac{1}{\alpha\sqrt{\mathrm{w}}}\sum_{i=1}^{k}\rho_i\|(\mathrm{w}_i\Sigma_i)(\sqrt{\mathrm{w}_i}\mu_i)\| + O(\alpha^{-2}) \\ &\leq \frac{1}{\alpha\sqrt{\mathrm{w}}}\sum_{i=1}^{k}\rho_i\|[A_i, B_i^T]^T\|\|(\sqrt{\mathrm{w}_i}\mu_i)\| + O(\alpha^{-2}).\end{aligned}$$

From isotropy, it follows that $\|\sqrt{\mathrm{w}_i}\mu_i\| \leq 1$. To bound the other factor, we argue

$$\|[A_i, B_i^T]^T\| \leq \sqrt{2}\max\{\|A_i\|, \|B_i\|\} \leq \sqrt{2\phi}.$$

Therefore,

$$\|u-v\|^2 \leq \frac{2k^2}{\alpha^2\mathrm{w}}\phi + O(\alpha^{-3}) \leq \frac{4k^2}{\alpha^2\mathrm{w}}\phi,$$

for sufficiently large n, as $\alpha \geq n/\mathrm{w}$. □

Lemma 10. *Let*

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

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

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

Before giving the proof, we summarize some of the necessary calculation in the following claim.

Claim 11. *The matrix of second moments*

$$M = E\left[xx^T\exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right] = \begin{bmatrix}\Gamma_{11} & 0\\ 0 & \Gamma_{22}\end{bmatrix} + \begin{bmatrix}\Delta_{11} & \Delta_{21}^T\\ \Delta_{21} & \Delta_{22}\end{bmatrix} + F,$$

where

$$\begin{aligned}\Gamma_{11} &= \sum_{i=1}^{k}\rho_i(\mathrm{w}_i\tilde{\mu}_i\tilde{\mu}_i{}^T + A_i)\\ \Gamma_{22} &= \sum_{i=1}^{k}\rho_i D_i - \frac{\rho_i}{\mathrm{w}_i\alpha}D_i^2\\ \Delta_{11} &= -\sum_{i=1}^{k}\frac{\rho_i}{\mathrm{w}_i\alpha}B_i^TB_i + \frac{\rho_i}{\mathrm{w}_i\alpha}\left(\mathrm{w}_i\tilde{\mu}_i\tilde{\mu}_i{}^TA_i + \mathrm{w}_iA_i\tilde{\mu}_i\tilde{\mu}_i{}^T + A_i^2\right)\\ \Delta_{21} &= \sum_{i=1}^{k}\rho_iB_i - \frac{\rho_i}{\mathrm{w}_i\alpha}\left(B_i(\mathrm{w}_i\tilde{\mu}_i\tilde{\mu}_i{}^T) + B_iA_i + D_iB_i\right)\\ \Delta_{22} &= -\sum_{i=1}^{k}\frac{\rho_i}{\mathrm{w}_i\alpha}B_iB_i^T,\end{aligned}$$

and $\|F\| = O(\alpha^{-2})$.

Proof. The calculation is straightforward. □

Proof of Lemma 10. We begin by bounding the 2-norm of each of the blocks. Since $\|\mathrm{w}_i\tilde{\mu}_i\tilde{\mu}_i{}^T\| < 1$ and $\|A_i\| \le \phi$ and $\|B_i\| \le \sqrt{\phi}$, we can bound

$$\begin{aligned}
\|\Delta_{11}\| &= \max_{\|y\|=1} \sum_{i=1}^{k} \frac{\rho_i}{\mathrm{w}_i\alpha} y^T B_i^T B_i y^T - \frac{\rho_i}{\mathrm{w}_i\alpha} y^T \left(\mathrm{w}_i\tilde{\mu}_i\tilde{\mu}_i{}^T A_i + \mathrm{w}_i A_i\tilde{\mu}_i\tilde{\mu}_i{}^T + A_i^2\right) y + O(\alpha^{-2}) \\
&\le \sum_{i=1}^{k} \frac{\rho_i}{\mathrm{w}_i\alpha}\|B_i\|^2 + \frac{\rho_i}{\mathrm{w}_i\alpha}(2\|A\| + \|A\|^2) + O(\alpha^{-2}) \\
&\le \frac{4k}{\mathrm{w}\alpha}\phi + O(\alpha^{-2}).
\end{aligned}$$

By a similar argument, $\|\Delta_{22}\| \le k\phi/(\mathrm{w}\alpha) + O(\alpha^{-2})$. For Δ_{21}, we observe that $\sum_{i=1}^{k} B_i = 0$. Therefore,

$$\begin{aligned}
\|\Delta_{21}\| &\le \left\|\sum_{i=1}^{k}(\rho_i - \bar{\rho})B_i\right\| + \left\|\sum_{i=1}^{k} \frac{\rho_i}{\mathrm{w}_i\alpha}\left(B_i(\mathrm{w}_i\tilde{\mu}_i\tilde{\mu}_i^T) + B_iA_i + D_iB_i\right)\right\| + O(\alpha^{-2}) \\
&\le \sum_{i=1}^{k}|\rho_i - \bar{\rho}|\|B_i\| + \sum_{i=1}^{k}\frac{\rho_i}{\mathrm{w}_i\alpha}\left(\|B_i(\mathrm{w}_i\tilde{\mu}_i\tilde{\mu}_i^T)\| + \|B_iA_i\| + \|D_iB_i\|\right) + O(\alpha^{-2}) \\
&\le k\|\rho - 1\bar{\rho}\|_\infty\sqrt{\phi} + \sum_{i=1}^{k}\frac{\rho_i}{\mathrm{w}_i\alpha}(\sqrt{\phi} + \phi\sqrt{\phi} + \sqrt{\phi}) + O(\alpha^{-2}) \\
&\le k\|\rho - 1\bar{\rho}\|_\infty\sqrt{\phi} + \frac{3k\bar{\rho}}{\mathrm{w}\alpha}\sqrt{\phi} \\
&\le \frac{7k}{2\mathrm{w}\alpha}\sqrt{\phi} + O(\alpha^{-2}).
\end{aligned}$$

Thus, we have $\max\{\|\Delta_{11}\|, \|\Delta_{22}\|, \|\Delta_{21}\|\} \le 4k\sqrt{\phi}/(\mathrm{w}\alpha) + O(\alpha^{-2})$, so that

$$\|M - \Gamma\| \le \|\Delta\| + O(\alpha^{-2}) \le 2\max\{\|\Delta_{11}\|, \|\Delta_{22}\|, \|\Delta_{21}\|\} \le \frac{8k}{\mathrm{w}\alpha}\sqrt{\phi} + O(\alpha^{-2}) \le \frac{16k}{\mathrm{w}\alpha}\sqrt{\phi}.$$

for sufficiently large n, as $\alpha \ge n/\mathrm{w}$. □

3.4 Sample Convergence

We now give some bounds on the convergence of the transformation to isotropy ($\hat{\mu} \to 0$ and $\hat{\Sigma} \to I$) and on the convergence of the reweighted sample mean $\hat{u}$ and sample matrix of second moments $\hat{M}$ to their expectations u and M. For the convergence of second moment matrices, we use the following lemma due to Rudelson [12], which was presented in this form in [13].

Lemma 12. *Let y be a random vector from a distribution D in R^n, with $\sup_D \|y\| = M$ and $\|\mathsf{E}(yy^T)\| \le 1$. Let $y_1, \ldots, y_m$ be independent samples from D. Let*

$$\eta = CM\sqrt{\frac{\log m}{m}}$$

where C is an absolute constant. Then,

(i) If $\eta < 1$, then

$$\mathsf{E}\left(\|\frac{1}{m}\sum_{i=1}^{m} y_i y_i^T - \mathsf{E}(yy^T)\|\right) \leq \eta.$$

(ii) For every $t \in (0,1)$,

$$\mathsf{P}\left(\|\frac{1}{m}\sum_{i=1}^{m} y_i y_i^T - \mathsf{E}(yy^T)\| > t\right) \leq 2e^{-ct^2/\eta^2}.$$

This lemma is used to show that a distribution can be made nearly isotropic using only $O^*(kn)$ samples [12, 10]. The isotropic transformation is computed simply by estimating the mean and covariance matrix of a sample, and computing the affine transformation that puts the sample in isotropic position.

Theorem 4. *There is an absolute constant C such that for an isotropic mixture of k logconcave distributions, with probability at least $1-\delta$, a sample of size*

$$m > C\frac{kn\log^2(n/\delta)}{\epsilon^2}$$

gives a sample mean $\hat{\mu}$ and sample covariance $\hat{\Sigma}$ so that

$$\|\hat{\mu}\| \leq \epsilon \quad \text{and} \quad \|\hat{\Sigma} - I\| \leq \epsilon.$$

We now consider the reweighted moments.

Lemma 13. *Let $\epsilon, \delta > 0$ and let $\hat{\mu}$ be the reweighted sample mean of a set of m points drawn from an isotropic mixture of k Gaussians in n dimensions, where*

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

Then

$$\mathsf{P}\left[\|\hat{u} - u\| > \epsilon\right] \leq \delta$$

Proof. We first consider only a single coordinate of the vector $\hat{u}$. Let $y = x_1 \exp\left(-\|x\|^2/(2\alpha)\right) - u_1$. We observe that

$$\left|x_1 \exp\left(-\frac{\|x\|^2}{2\alpha}\right)\right| \leq |x_1| \exp\left(-\frac{x_1^2}{2\alpha}\right) \leq \sqrt{\frac{\alpha}{e}} < \sqrt{\alpha}.$$

Thus, each term in the sum $m\hat{u}_1 = \sum_{j=1}^{m} y_j$ falls the range $[-\sqrt{\alpha} - u_1, \sqrt{\alpha} - u_1]$. We may therefore apply Hoeffding's inequality to show that

$$\mathsf{P}\left[|\hat{u}_1 - u_1| \geq \epsilon/\sqrt{n}\right] \leq 2\exp\left(-\frac{2m^2(\epsilon/\sqrt{n})^2}{m\cdot(2\sqrt{\alpha})^2}\right) \leq 2\exp\left(-\frac{m\epsilon^2}{2\alpha n}\right) \leq \frac{\delta}{n}.$$

Taking the union bound over the n coordinates, we have that with probability $1-\delta$ the error in each coordinate is at most $\epsilon/\sqrt{n}$, which implies that $\|\hat{u} - u\| \leq \epsilon$. □

Lemma 14. *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

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

Proof. We will apply Lemma 12. Define $y = x \exp\left(-\|x\|^2/(2\alpha)\right)$. Then,

$$y_i^2 \leq x_i^2 \exp\left(-\frac{\|x\|^2}{\alpha}\right) \leq x_i^2 \exp\left(-\frac{x_i^2}{\alpha}\right) \leq \frac{\alpha}{e} < \alpha.$$

Therefore $\|y\| \leq \sqrt{\alpha n}$.

Next, since M is in isotropic position (we can assume this w.l.o.g.), we have for any unit vector v,

$$\mathsf{E}((v^T y)^2)) \leq \mathsf{E}((v^T x)^2) \leq 1$$

and so $\|\mathsf{E}(yy^T)\| \leq 1$.

Now we apply the second part of Lemma 12 with $\eta = \epsilon\sqrt{c/\ln(2/\delta)}$ and $t = \eta\sqrt{\ln(2/\delta)/c}$. This requires that

$$\eta = \frac{c\epsilon}{\ln(2/\delta)} \leq C\sqrt{\alpha n}\sqrt{\frac{\log m}{m}}$$

which is satisfied for our choice of m. □

Lemma 15. *Let X be a collection of m points drawn from a Gaussian with mean μ and variance σ^2. With probability $1 - \delta$,*

$$|x - \mu| \leq \sigma\sqrt{2\log m/\delta}.$$

for every $x \in X$.

3.5 Perturbation Lemma

We will use the following key lemma due to Stewart [16] to show that when we apply the spectral step, the top $k-1$ dimensional invariant subspace will be close to the Fisher subspace.

Lemma 16 (Stewart's Theorem). *Suppose A and $A + E$ are n-by-n symmetric matrices and that*

$$A = \begin{array}{c} \\ r \\ n-r \end{array} \begin{array}{c} \begin{array}{cc} r & n-r \end{array} \\ \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \end{array} \qquad E = \begin{array}{c} \\ r \\ n-r \end{array} \begin{array}{c} \begin{array}{cc} r & n-r \end{array} \\ \begin{bmatrix} E_{11} & E_{21}^T \\ E_{21} & E_{22} \end{bmatrix} \end{array}.$$

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

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

4 Finding a Vector near the Fisher Subspace

In this section, we combine the approximations of Section 3.3 and the perturbation lemma of Section 3.5 to 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.

We first assume zero overlap and that the sample reweighted moments behave exactly according to expectation. In this case, the mean shift $\hat{u}$ 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, then $\hat{M}$ becomes

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

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 10 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 17 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 18 guarantees that this vector is close to the Fisher subspace.

Lemma 17 (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 18 (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

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

4.1 Mean Shift

Proof of Lemma 17. We will make use of the following claim.

Claim 19. *For any vectors $a, b \neq 0$,*

$$\frac{|a^T b|}{\|a\|\|b\|} \geq \left(1 - \frac{\|a-b\|^2}{\max\{\|a\|^2, \|b\|^2\}}\right)^{1/2}.$$

By the triangle inequality, $\|\hat{u} - v\| \leq \|\hat{u} - u\| + \|u - v\|$. By Lemma 9,

$$\|u - v\| \leq \sqrt{\frac{4k^2}{\alpha^2 \mathrm{w}}\phi} = \sqrt{\frac{4k^2}{\alpha^2 \mathrm{w}} \cdot \frac{\mathrm{w}^2\epsilon}{2^1 4k^2}} \leq \sqrt{\frac{\mathrm{w}\epsilon}{2^{12}\alpha^2}}.$$

By Lemma 13, for large m_1 we obtain the same bound on $\|\hat{u} - u\|$ with probability $1 - \delta$. Thus,

$$\|\hat{u} - v\| \leq \sqrt{\frac{\mathrm{w}\epsilon}{2^{10}\alpha^2}}.$$

Applying the claim gives

$$\begin{aligned} \frac{\|\hat{u}^T v\|}{\|\hat{u}\|\|v\|} &\geq 1 - \frac{\|\hat{u} - v\|^2}{\|\hat{u}\|^2} \\ &\geq 1 - \frac{\mathrm{w}\epsilon}{2^{10}\alpha^2} \cdot \frac{32^2\alpha^2}{\mathrm{w}} \\ &= 1 - \epsilon. \end{aligned}$$

□

Proof of Claim 19. Without loss of generality, assume $\|u\| \geq \|v\|$ and fix the distance $\|u - v\|$. In order to maximize the angle between u and v, the vector v should be chosen so that it is tangent to the sphere centered at u with radius $\|u - v\|$. Hence, the vectors u,v,$(u - v)$ form a right triangle where $\|u\|^2 = \|v\|^2 + \|u - v\|^2$. For this choice of v, let θ be the angle between u and v so that

$$\frac{u^T v}{\|u\|\|v\|} = \cos\theta = (1 - \sin^2\theta)^{1/2} = \left(1 - \frac{\|u - v\|^2}{\|u\|^2}\right)^{1/2}.$$

□

4.2 Spectral Method

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 20 (Small Mean Shift Implies Balanced Second Moments). *If $\|\hat{u}| \leq \sqrt{\mathrm{w}}/(32\alpha)$ and*

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

then

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

Proof. Let $q_1, \ldots, q_k$ be the right singular vectors of the matrix $U = [\mathrm{w}_1\mu_1, \ldots, \mathrm{w}_k\mu_k]$ and let $\sigma_i(U)$ be the ith largest singular value. Because $\sum_{i=1}^k \mathrm{w}_i\mu_i = 0$, we have that $\sigma_k(U) = 0$ and $q_k = 1/\sqrt{k}$. Recall that ρ is the k vector of scalars $\rho_1, \ldots, \rho_k$ and that $v = U\rho$. Then

$$\begin{aligned}
\|v\|^2 &= \|U\rho\|^2 \\
&= \sum_{i=1}^{k-1} \sigma_i(U)^2 (q_i^T\rho)^2 \\
&\geq \sigma_{k-1}(U)^2 \|\rho - q_k(q_k^T\rho)\|_2^2 \\
&= \sigma_{k-1}(U)^2 \|\rho - 1\bar{\rho}\|_2^2.
\end{aligned}$$

Because $q_{k-1} \in \mathrm{span}\{\mu_1, \ldots, \mu_k\}$, we have that $\sum_{i=1}^k \mathrm{w}_i q_{k-1}^T \mu_i \mu_i^T q_{k-1} \geq 1 - \phi$. Therefore,

$$\begin{aligned}
\sigma_{k-1}(U)^2 &= \|Uq_{k-1}\|^2 \\
&= q_{k-1}^T \left(\sum_{i=1}^k \mathrm{w}_i^2 \mu_i \mu_i^T \right) q_{k-1} \\
&\geq \mathrm{w} q_{k-1}^T \left(\sum_{i=1}^k \mathrm{w}_i \mu_i \mu_i^T \right) q_{k-1} \\
&\geq \mathrm{w}(1-\phi).
\end{aligned}$$

Thus, we have the bound

$$\|\rho - 1\bar{\rho}\|_\infty \leq \frac{1}{\sqrt{(1-\phi)\mathrm{w}}} \|v\| \leq \frac{2}{\sqrt{\mathrm{w}}} \|v\|.$$

By the triangle inequality $\|v\| \leq \|\hat{u}\| + \|\hat{u} - v\|$. As argued in Lemma 9,

$$\|\hat{u} - v\| \leq \sqrt{\frac{4k^2}{\alpha^2 \mathrm{w}} \phi} = \sqrt{\frac{4k^2}{\alpha^2 \mathrm{w}} \cdot \frac{\mathrm{w}^2}{64^2 k^2}} =\leq \frac{\sqrt{\mathrm{w}}}{32\alpha}.$$

Thus,

$$\begin{aligned}
\|\rho - 1\bar{\rho}\|_\infty &\leq \frac{2\bar{\rho}}{\sqrt{\mathrm{w}}} \|v\| \\
&\leq \frac{2\bar{\rho}}{\sqrt{\mathrm{w}}} \left(\frac{\sqrt{\mathrm{w}}}{32\alpha} + \frac{\sqrt{\mathrm{w}}}{32\alpha} \right) \\
&\leq \frac{1}{8\alpha}.
\end{aligned}$$

□

We next show that the top $k-1$ principal components of Γ span the intermean subspace and put a lower bound on the spectral gap between the intermean and non-intermean components.

Lemma 21 (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.

Proof of Lemma 21. We first bound $\lambda_{k-1}(\Gamma_{11})$. Recall that

$$\Gamma_{11} = \sum_{i=1}^{k} \rho_i(\mathrm{w}_i\tilde{\mu}_i\tilde{\mu}_i{}^T + A_i).$$

Thus,

$$\begin{aligned}\lambda_{k-1}(\Gamma_{11}) &= \min_{\|y\|=1} \sum_{i=1}^{k} \rho_i y^T(\mathrm{w}_i\tilde{\mu}_i\tilde{\mu}_i{}^T + A_i)y \\ &\geq \bar{\rho} - \max_{\|y\|=1} \sum_{i=1}^{k} (\bar{\rho} - \rho_i) y^T(\mathrm{w}_i\tilde{\mu}_i\tilde{\mu}_i{}^T + A_i)y.\end{aligned}$$

We observe that $\sum_{i=1}^{k} y^T(\mathrm{w}_i\tilde{\mu}_i\tilde{\mu}_i{}^T + A_i)y = 1$ and each term is non-negative. Hence the sum is bounded by

$$\sum_{i=1}^{k} (\bar{\rho} - \rho_i) y^T(\mathrm{w}_i\tilde{\mu}_i\tilde{\mu}_i{}^T + A_i)y \leq \|\rho - 1\bar{\rho}\|_\infty,$$

so,

$$\lambda_{k-1}(\Gamma_{11}) \geq \bar{\rho} - \|\rho - 1\bar{\rho}\|_\infty.$$

Next, we bound $\lambda_1(\Gamma_{22})$. Recall that

$$\Gamma_{22} = \sum_{i=1}^{k} \rho_i D_i - \frac{\rho_i}{\mathrm{w}_i\alpha} D_i^2$$

and that for any $n-k$ vector y such that $\|y\| = 1$, we have $\sum_{i=1}^{k} y^T D_i y = 1$. Using the same arguments as above,

$$\begin{aligned}\lambda_1(\Gamma_{22}) &= \max_{\|y\|=1} \bar{\rho} + \sum_{i=1}^{k} (\rho_i - \bar{\rho}) y^T D_i y - \frac{\rho_i}{\mathrm{w}_i\alpha} y^T D_i^2 y \\ &\leq \bar{\rho} + \|\rho - 1\bar{\rho}\|_\infty - \min_{\|y\|=1} \sum_{i=1}^{k} \frac{\rho_i}{\mathrm{w}_i\alpha} y^T D_i^2 y.\end{aligned}$$

To bound the last sum, we observe that $\rho_i - \bar{\rho} = O(\alpha^{-1})$. Therefore

$$\sum_{i=1}^{k} \frac{\rho_i}{\mathrm{w}_i\alpha} y^T D_i^2 y \geq \frac{\bar{\rho}}{\alpha} \sum_{i=1}^{k} \frac{1}{\mathrm{w}_i} y^T D_i^2 y + O(\alpha^{-2}).$$

Without loss of generality, we may assume that $y = e_1$ by an appropriate rotation of the D_i. Let $D_i(\ell, j)$ be element in the ℓth row and jth column of the matrix D_i. Then the sum becomes

$$\begin{aligned}\sum_{i=1}^{k} \frac{1}{\mathrm{w}_i} y^T D_i^2 y &= \sum_{i=1}^{k} \frac{1}{\mathrm{w}_i} \sum_{j=1}^{n} D_j(1,j)^2 \\ &\geq \sum_{i=1}^{k} \frac{1}{\mathrm{w}_i} D_j(1,1)^2.\end{aligned}$$

Because $\sum_{i=1}^k D_i = I$, we have $\sum_{i=1}^k D_i(1,1) = 1$. From the Cauchy-Schwartz inequality, it follows

$$\left(\sum_{i=1}^k \mathrm{w}_i\right)^{1/2} \left(\sum_{i=1}^k \frac{1}{\mathrm{w}_i} D_i(1,1)^2\right)^{1/2} \geq \sum_{i=1}^k \sqrt{\mathrm{w}_i} \frac{D_i(1,1)}{\sqrt{\mathrm{w}_i}} = 1.$$

Since $\sum_{i=1}^k \mathrm{w}_i = 1$, we conclude that $\sum_{i=1}^k \frac{1}{\mathrm{w}_i} D_i(1,1)^2 \geq 1$. Thus, using the fact that $\bar{\rho} \geq 1/2$, we have

$$\sum_{i=1}^k \frac{\rho_i}{\mathrm{w}_i \alpha} y^T D_i^2 y \geq \frac{1}{2\alpha}$$

Putting the bounds together

$$\lambda_{k-1}(\Gamma_{11}) - \lambda_1(\Gamma_{22}) \geq \frac{1}{2\alpha} - 2\|\rho - 1\bar{\rho}\|_\infty \geq \frac{1}{4\alpha}.$$

□

Proof of Lemma 18. To bound the effect of overlap and sample errors on the eigenvectors, we apply Stewart's Lemma (Lemma 16). 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 21, this implies that

$$d = \lambda_{k-1}(\Gamma) - \lambda_k(\Gamma) \geq \frac{1}{4\alpha}. \tag{5}$$

To bound $\|E\|$, we use the triangle inequality $\|E\| \leq \|\Gamma - M\| + \|M - \hat{M}\|$. Lemma 10 bounds the first term by

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

By Lemma 14, 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. 5 and 4.2, we have

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

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 21, 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,

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

□

5 Recursion

In this section, we show that 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 22. *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 $1 - \delta'$.

Proof of Lemma 22. 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.

We first show that at least one mean must be far from the origin in the direction h. Let the columns of P_1 be the vectors $e_1, \ldots, e_{k-1}$. The span of these vectors is also the span of the means, so we have

$$\begin{aligned}
\max_i (h^T \mu_i)^2 &= \max_i (h^T P_1 P_1^T \mu_i)^2 \\
&= \|P_1^T h\|^2 \max_i \left(\frac{(P_1^T h)^T}{\|P_1 h\|} \tilde{\mu}_i \right)^2 \\
&\geq \|P_1^T h\|^2 \sum_{i=1}^{k} \mathrm{w}_i \left(\frac{(P_1^T h)^T}{\|P_1 h\|} \tilde{\mu}_i \right)^2 \\
&\geq \|P_1^T h\|^2 (1 - \phi) \\
&> \frac{1}{2}.
\end{aligned}$$

Since the origin is the mean of the means, we conclude that the maximum distance between two means in the direction h is at least $1/2$. Without loss of generality, we assume that the interval $[0, 1/2]$ is contained between two means projected to h.

We now show that every point x drawn from component i falls in a narrow interval when projected to h. That is, x satisfies $h^T x \in b_i$, where $b_i = [h^T \mu_i - (8(k-1))^{-1}, h^T \mu_i + (8(k-1))^{-1}]$. We begin by examining the variance along h. Let $e_k, \ldots, e_n$ be the columns of the matrix n-by-$(n-k+1)$ matrix P_2. Recall from Eqn. 2 that $P_1^T \mathrm{w}_i \Sigma_i P_1 = A_i$, that $P_2^T \mathrm{w}_i \Sigma_i P_1 = B_i$, and that

$P_2^T \mathrm{w}_i \Sigma_i P_2 = D_i$. The norms of these matrices are bounded according to Lemma 4. Also, the vector $h = P_1 P_1^T h + P_2 P_2^T h$. For convenience of notation we define ϵ such that $\|P_1^T h\| = 1 - \epsilon$. Then $\|P_2^T h\|^2 = 1 - (1-\epsilon)^2 \leq 2\epsilon$. We now argue

$$\begin{aligned} h^T \mathrm{w}_i \Sigma_i h &\leq \left(h^T P_1 A_i P_1^T h + 2 h^T P_2 B_i P_1 h + h^T P_2^T D_i P_2 h\right) \\ &\leq 2\left(h^T P_1 A_i P_1^T h + h^T P_2 D_i P_2^T h\right) \\ &\leq 2(\|P_1^T h\|^2 \|A_i\| + \|P_2^T h\|^2 \|\|D_i\|) \\ &\leq 2(\phi + 2\epsilon). \end{aligned}$$

Using the assumptions about ϕ and ϵ, we conclude that the maximum variance along h is at most

$$\max_i h^T \Sigma_i h \leq \frac{2}{\mathrm{w}} \left(\frac{\mathrm{w}}{2^9 (k-1)^2} \log \frac{1}{\delta'} + 2 \frac{\mathrm{w}}{2^{10}(k-1)^2} \log \frac{1}{\delta'} \right) \leq \left(2^7 (k-1)^2 \log 1/\delta'\right)^{-1}.$$

We now translate these bounds on the variance to a bound on the difference between the minimum and maximum points along the direction h. By Lemma 15, with probability $1 - \delta/2$

$$|h^T (x - \mu_{\ell(x)})| \leq \sqrt{2 h^T \Sigma_i h \log(2m_2/\delta)} \leq \frac{1}{8(k-1)} \cdot \frac{\log(2m_2/\delta)}{\log(1/\delta')} \leq \frac{1}{8(k-1)}.$$

Thus, with probability $1 - \delta/2$, every point from X falls into the union of intervals $b_1 \cup \ldots \cup b_k$ where $b_i = [h^T \mu_i - (8(k-1))^{-1}, h^T \mu_i + (8(k-1))^{-1}]$. Because these intervals are centered about the means, at least the equivalent of one interval must fall outside the range $[0, 1/2]$, which we assumed was contained between two projected means. Thus, the measure of subset of $[0, 1/2]$ that does not fall into one of the intervals is

$$\frac{1}{2} - (k-1)\frac{1}{4(k-1)} = \frac{1}{4}.$$

This set can be cut into at most $k-1$ intervals, so the smallest possible gap between these intervals is $(4(k-1))^{-1}$, which is exactly the width of an interval.

Because $m_2 = k/\mathrm{w} \log(2k/\delta)$ the set X contains at least one sample from every component with probability $1 - \delta/2$. Overall, with probability $1 - \delta$ every component has at least one sample and all samples from component i fall in b_i. Thus, the largest gap between the sampled points will not contain one of the intervals $b_1, \ldots, b_k$. Moreover, the midpoint t of this gap must also fall outside of $b_1 \cup \ldots \cup b_k$, ensuring that no b_i is cut by t.

By the same argument given above, any single point y from $\mathcal{F}$ is contained in $b_1 \cup \ldots \cup b_k$ with probability $1 - \delta'$ proving the Lemma. □

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

Proof of Lemma 23. Suppose that the components $j+1, \ldots k$ are removed from the mixture. Let $\omega = \sum_{i=1}^{j} \mathrm{w}_i$ be a normalizing factor for the weights. Then if $c = \sum_{i=1}^{j} \mathrm{w}_i \mu_i = -\sum_{i=j+1}^{k} \mathrm{w}_i \mu_i$, the

induced mean is $\omega^{-1}c$. Let T be the subspace that minimizes the maximum overlap for the full k component mixture. We then argue that the overlap $\tilde{\phi}^2$ of the induced mixture is bounded by

$$\begin{aligned}\tilde{\phi} &= \min_{\dim(S)=j-1} \max_{v\in S} \frac{\omega^{-1} v^T \Sigma v}{\omega^{-1}\sum_{i=1}^{j} \mathrm{w}_i v^T(\mu_i\mu_i^T - cc^T + \Sigma_i)v} \\ &\leq \max_{v\in \mathrm{span}\{e_1,\ldots,e_{k-1}\}\setminus \mathrm{span}\{\mu_{j+1},\ldots,\mu_k\}} \frac{\sum_{i=1}^{j} \mathrm{w}_i v^T\Sigma_i v}{\sum_{i=1}^{j} \mathrm{w}_i v^T(\mu_i\mu_i^T - cc^T + \Sigma_i)v}.\end{aligned}$$

Every $v \in \mathrm{span}\{e_1,\ldots,e_{k-1}\}\setminus \mathrm{span}\{\mu_{j+1},\ldots,\mu_k\}$ must be orthogonal to every μ_ℓ for $j+1 \leq \ell \leq k$. Therefore, v must be orthogonal to c as well. This also enables us to add the terms for $j+1,\ldots,k$ in both the numerator and denominator, because they are all zero.

$$\begin{aligned}\tilde{\phi} &\leq \max_{v\in \mathrm{span}\{e_1,\ldots,e_{k-1}\}\setminus \mathrm{span}\{\mu_{j+1},\ldots,\mu_k\}} \frac{v^T\Sigma v}{\sum_{i=1}^{k} \mathrm{w}_i v^T(\mu_i\mu_i^T + \Sigma_i)v} \\ &\leq \max_{v\in \mathrm{span}\{e_1,\ldots,e_{k-1}\}} \frac{v^T\Sigma v}{\sum_{i=1}^{k} \mathrm{w}_i v^T(\mu_i\mu_i^T + \Sigma_i)v} \\ &= \phi.\end{aligned}$$

□

The proofs of the main theorems are now apparent. Consider the case of $k = 2$ Gaussians first. As argued in Section 3.4, 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 22 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 23. 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] K. Chaudhuri and S. Rao. Learning mixtures of product distributions using correlations and independence. In *Proc. of COLT*, 2008.

[3] Anirban Dasgupta, John Hopcroft, Jon Kleinberg, and Mark Sandler. On learning mixtures of heavy-tailed distributions. In *Proc. of FOCS*, 2005.

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

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

[6] Jon Feldman, Rocco A. Servedio, and Ryan O'Donnell. Pac learning axis-aligned mixtures of gaussians with no separation assumption. In *COLT*, pages 20–34, 2006.

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

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

[9] 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.

[10] L. Lovász and S. Vempala. The geometry of logconcave functions and and sampling algorithms. *Random Strucures and Algorithms*, 30(3):307–358, 2007.

[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] M. Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164:60–72, 1999.

[13] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4), 2007.

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

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

[16] G.W. Stewart and Ji guang Sun. *Matrix Perturbation Theory*. Academic Press, Inc., 1990.

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