Spectral Algorithms

Ravi Kannan         Santosh Vempala

August, 2007
# Contents

## I  Applications

1 **The Best-Fit Subspace** 3  
   1.1 Singular Value Decomposition 3  
   1.2 Algorithms for computing the SVD 7  
   1.3 The $k$-variance problem 7  
   1.4 Discussion 10  

2 **Mixture Models** 11  
   2.1 Probabilistic separation 12  
   2.2 Geometric separation 12  
   2.3 Spectral Projection 14  
   2.4 Weakly Isotropic Distributions 16  
   2.5 Mixtures of general distributions 17  
   2.6 Spectral projection with samples 18  
   2.7 Discussion 19  

## II  Algorithms 21
Part I

Applications
Chapter 1

The Best-Fit Subspace

Finding the best-fit line for a set of data points is a classic problem. The most commonly used measure of the quality of a line is the least squares measure where we take the sum of squared (perpendicular) distances of the points to the line. More generally, for a set of data points in $\mathbb{R}^n$, one could ask for the best-fit $k$-dimensional subspace. The Singular Value Decomposition (SVD) can be used to find a subspace that minimizes the sum of squared distances of the points in polynomial time. In contrast, for other measures such as the sum of distances or the maximum distance, no polynomial-time algorithms are known.

Two clustering problems widely studied in theoretical computer science are the $k$-center and $k$-median problems. In the first problem, the goal is to find a set of $k$ points ("facilities") to minimize the maximum distance of any data point to its nearest facility. In the second problem, one finds a set of $k$ points so that the sum of the distances of each data point to its nearest facility is minimized. A natural relaxation of the $k$-median problem is to find the $k$-dimensional subspace for which the sum of the (perpendicular) distances of the data points to the subspace is minimized (we will later see that this is a relaxation).

1.1 Singular Value Decomposition

For an $n \times n$ matrix $A$, an eigenvalue $\lambda$ and corresponding eigenvector $v$ satisfy the equation $Av = \lambda v$. In general, i.e., if the matrix has nonzero determinant, it will have $n$ nonzero eigenvalues (not necessarily distinct) and $n$ corresponding eigenvectors.

Here we deal with an $m \times n$ rectangular matrix $A$, where the $m$ rows denoted $A^{(1)}, A^{(2)}, \ldots, A^{(m)}$ are points in $\mathbb{R}^n$.

If $m \neq n$, the notion of an eigenvalue and eigenvector does not make sense, since the vectors $Av$ and $\lambda v$ have different dimensions. Instead, a singular value $\sigma$ and corresponding singular vectors $u \in \mathbb{R}^m, v \in \mathbb{R}^n$ simultaneously satisfy the following two equations
1. \( Av = \sigma u \)
2. \( u^T A = \sigma v^T \).

These conditions are quite special. In general, we would not expect an arbitrary pair of vectors to satisfy them.

We can assume, without loss of generality, that \( u \) and \( v \) are unit vectors. To see this, note that a pair of singular vectors \( u \) and \( v \) must have equal magnitude, since \( u^T Av = \sigma \|u\|^2 = \sigma \|v\|^2 \). If this magnitude is not 1, we can rescale both of them by the same factor without violating the above equations.

Now we turn our attention to the value \( \max_{\|v\|=1} \|Av\|^2 \). Since the rows of \( A \) form a set of \( m \) vectors in \( \mathbb{R}^n \), \( Av \) is a list of the projections of those vectors onto the line spanned by \( v \), and \( \|Av\|^2 \) is simply the sum of the squares of those projections.

Instead of choosing \( v \) to maximize \( \|Av\|^2 \), the Pythagorean theorem allows us to equivalently choose \( v \) to minimize the sum of the squared distances of the points to the line through \( v \). In this sense, \( v \) defines the line through the origin that best fits the points.

![Figure 1.1: The vector \( A^{(i)} \) projected onto \( v \).](image)

To argue this more formally, let \( d(A^{(i)}, v) \) denote the distance of the point \( A^{(i)} \) to the line through \( v \). Alternatively, we can write

\[
d(A^{(i)}, v) = \| A^{(i)} - (A^{(i)}v)v \|.
\]

This is illustrated in Figure 1. For a unit vector \( v \), the Pythagorean theorem tells us that

\[
\|A^{(i)}\|^2 = \| (A^{(i)}v)v \|^2 + d(A^{(i)}, v)^2.
\]

Thus we get the following proposition:

**Proposition 1.1.**

\[
\max_{\|v\|=1} \|Av\|^2 = \min_{\|v\|=1} \| A - (Av)v^T \|_F^2 = \min_{\|v\|=1} \sum_i \| A^{(i)} - (A^{(i)}v)v \|^2
\]

**Proof.** We simply use the identity:

\[
\|Av\|^2 = \sum_i \| (A^{(i)}v)v \|^2 = \sum_i \| A^{(i)} \|^2 - \sum_i \| A^{(i)} - (A^{(i)}v)v \|^2
\]
1.1. SINGULAR VALUE DECOMPOSITION

The proposition says that the \( v \) which maximizes \( \|Av\|^2 \) is the “best-fit” vector which also minimizes \( \sum_i d(A^{(i)}, v)^2 \).

Next, we claim that \( v \) is in fact a singular vector.

**Proposition 1.2.** The vector \( v_1 = \arg \max_{\|v\|=1} \|Av\|^2 \) is a singular vector, and moreover \( \|Av_1\| = \sigma \) for the largest (or “top”) singular value \( \sigma \).

**Proof.** Note that for any singular vector \( v \), \((A^T A)v = \sigma Av = \sigma^2 v\). Thus, \( v \) is an eigenvector of \( A^T A \) with corresponding eigenvalue \( \sigma^2 \). Conversely, an eigenvector of \( A^T A \) is also a singular vector of \( A \). To see this, let \( v \) be an eigenvector of \( A^T A \) with corresponding eigenvalue \( \lambda \). Note that \( \lambda \) is positive, since \( \|Av\|^2 = v^T A^T A v = \lambda v^T v = \lambda \|v\|^2 \) and thus \( \lambda = \|Av\|^2/\|v\|^2 \). Now if we let \( \sigma = \sqrt{\lambda} \) and \( u = Av/\sigma \), it is easy to verify that \( u, v, \) and \( \sigma \) satisfy the singular value requirements.

The right singular vectors \( \{v_i\} \) are thus exactly equal to the eigenvectors of \( A^T A \). Since \( A^T A \) is a real, symmetric matrix, it has \( n \) orthonormal eigenvectors, which we can label \( v_1, \ldots, v_n \). Expressing a unit vector \( v \) in terms of \( \{v_i\} \) (i.e. \( v = \sum_i \alpha_i v_i \) where \( \sum_i \alpha_i^2 = 1 \)), we see that \( \|Av\|^2 = \sum_i \sigma_i \alpha_i^2 \) which is maximized exactly when \( v \) corresponds to the top eigenvector of \( A^T A \).

A line through the origin is a one-dimensional subspace, but we might also ask for a plane through the origin that best fits the data. Or more generally, a \( k \)-dimensional subspace that best fits the data. It turns out that this space is specified by the top \( k \) singular vectors, as the following proposition states.

**Theorem 1.3.** Define the \( k \)-dimensional subspace \( V_k \) as the span of the following \( k \) vectors:

\[
\begin{align*}
v_1 &= \arg \max_{\|v\|=1} \|Av\| \\
v_2 &= \arg \max_{\|v\|=1, v \cdot v_1 = 0} \|Av\| \\
& \quad \vdots \\
v_k &= \arg \max_{\|v\|=1, v \cdot v_i = 0 \quad \forall i < k} \|Av\|
\end{align*}
\]

Then \( V_k \) is optimal in the sense that

\[
V_k = \arg \min_{\text{dim}(V) = k} \sum_i d(A^{(i)}, V)^2.
\]

Further, \( v_1, v_2, \ldots, v_n \) are all singular vectors, with corresponding singular values \( \sigma_1, \sigma_2, \ldots, \sigma_n \) and

\[
\sigma_1 = \|Av_1\| \geq \sigma_2 = \|Av_2\| \geq \ldots \geq \sigma_n = \|Av_n\|.
\]

Finally, \( A = \sum_{i=1}^n \sigma_i u_i v_i^T \).

Such a decomposition where,
1. The sequence of $\sigma_i$'s is nonincreasing

2. The sets $\{u_i\}, \{v_i\}$ are orthonormal

is called the singular value decomposition of $A$.

Proof. We first prove that $V_k$ are optimal by induction on $k$. The case $k = 1$ is by definition. Assume that $V_{k-1}$ is optimal.

Suppose $V'_k$ is an optimal subspace of dimension $k$. Then we can choose an orthonormal basis for $V'_k$, say $w_1, w_2, \ldots, w_k$, such that $w_k$ is orthogonal to $V_{k-1}$.

By the definition of $V'_k$, we have that

$$\|Aw_1\|^2 + \|Aw_2\|^2 + \ldots + \|Aw_k\|^2$$

is maximized (among all sets of $k$ orthonormal vectors.) If now we replace $w_i$ by $v_i$ for $i = 1, 2, \ldots, k-1$, we have

$$\|Av_1\|^2 + \|Av_2\|^2 + \ldots + \|Av_k\|^2 \leq \|Av_1\|^2 + \ldots + \|Av_{k-1}\|^2 + \|Aw_k\|^2.$$

Therefore we can assume that $V'_k$ is the span of $V_{k-1}$ and $w_k$. It then follows that $\|Aw_k\|^2$ maximizes $\|Ax\|^2$ over all unit vectors $x$ orthogonal to $V_{k-1}$.

Proposition 1.2 can be extended to show that $v_1, v_2, \ldots, v_n$ are all singular vectors. The assertion that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$ follows from the definition of the $v_i$'s, and the argument given in Proposition 2 that all the singular values are non-negative.

We can verify that the decomposition $A = \sum_{i=1}^n \sigma_i u_i v_i^T$ is accurate. This is because the vectors $v_1, v_2, \ldots, v_n$ form an orthonormal basis for $R^n$, and the action of $A$ on any $v_i$ is equivalent to the action of $\sum_{i=1}^n \sigma_i u_i v_i^T$ on $v_i$. Note that we could actually decompose $A$ into the form $\sum_{i=1}^n \sigma_i u_i v_i^T$ by picking $\{v_i\}$ to be any orthogonal basis of $R^n$, but the proposition actually states something stronger: that we can pick $\{v_i\}$ in such a way that $\{u_i\}$ is also an orthogonal set.

We state one more classical theorem. We have said the span of the top $k$ singular vectors is the best-fit $k$-dimensional subspace for the rows of $A$. Along the same lines, the partial decomposition of $A$ obtained by using only the top $k$ singular vectors is the best rank-$k$ matrix approximation to $A$.

Theorem 1.4. Among all rank $k$ matrices $D$, $A_k = \sum_{i=1}^k \sigma_i u_i v_i^T$ is the one which minimizes $\|A - D\|_F^2 = \sum_{ij} (A_{ij} - D_{ij})^2$. Further,

$$\|A - A_k\|_F^2 = \sum_{i=k+1}^n \sigma_i^2.$$

Proof. We have

$$\|A - D\|_F^2 = \sum_{i=1}^m \|A^{(i)} - D^{(i)}\|^2.$$
1.2. ALGORITHMS FOR COMPUTING THE SVD

Since \( D \) is of rank at most \( k \), we can assume that all the \( D^{(i)} \) are projections of \( A^{(i)} \) to some rank \( k \) subspace and therefore,

\[
\sum_{i=1}^{m} \|A^{(i)} - D^{(i)}\|^2 = \sum_{i=1}^{m} \|A^{(i)}\|^2 - \|D^{(i)}\|^2 \\
= \|A\|^2_F - \sum_{i=1}^{m} \|D^{(i)}\|^2.
\]

Thus the subspace is exactly the SVD subspace given by the span of the first \( k \) singular vectors of \( A \).

1.2 Algorithms for computing the SVD

Computing the SVD is a major topic of numerical analysis \([?, ?, ?]\). Here we describe a simple algorithm called the power method.

Assume that \( A \) is symmetric.

1. Let \( x \) be a random unit vector.
2. Repeat:
   \[
   x := \frac{Ax}{\|Ax\|}
   \]

Exercise 1.1. Show that the power iteration applied \( k \) times to a symmetric matrix \( A \) finds a vector \( x^k \) such that

\[
\|Ax^k\| \geq \left(\frac{\delta}{n}\right)^k \sigma_1(A).
\]

In the case of a nonsymmetric matrix, we can simply apply the iteration to \( A^T A \).

1.3 The \( k \)-variance problem

This section contains a description of a clustering problem which is often called \( k \)-means in the literature and can be solved approximately using SVD. This illustrates a typical use of SVD and has a provable bound.

We are given \( m \) points \( \mathcal{A} = \{A^{(1)}, A^{(2)}, \ldots, A^{(m)}\} \) in \( n \)-dimensional Euclidean space and a positive integer \( k \). The problem is to find \( k \) points \( \mathcal{B} = \{B^{(1)}, B^{(2)}, \ldots, B^{(k)}\} \) such that

\[
f_{\mathcal{A}}(\mathcal{B}) = \sum_{i=1}^{m} (\text{dist}(A^{(i)}, \mathcal{B}))^2
\]

is minimized. Here dist\((A^{(i)}, \mathcal{B})\) is the Euclidean distance of \( A^{(i)} \) to its nearest point in \( \mathcal{B} \). Thus, in this problem we wish to minimize the sum of squared distances to the nearest “cluster center”. We call this the \( k \)-variance problem.
CHAPTER 1. THE BEST-FIT SUBSPACE

Note that the solution is given by $k$ clusters $S_j, j = 1, 2, \ldots, k$. The cluster center $B^{(j)}$ will be the centroid of the points in $S_j$, $j = 1, 2, \ldots, k$. This is seen from the fact that for any set $S = \{X^{(1)}, X^{(2)}, \ldots, X^{(r)}\}$ and any point $B$ we have

$$
\sum_{i=1}^{r} |X^{(i)} - B|^2 = \sum_{i=1}^{r} |X^{(i)} - \bar{X}|^2 + r|B - \bar{X}|^2, 
$$

(1.1)

where $\bar{X}$ is the centroid $(X^{(1)} + X^{(2)} + \cdots + X^{(r)})/r$ of $S$. The next exercise makes this clear.

**Exercise 1.2.** Show that for a set of points $X^1, \ldots, X^k \in \mathbb{R}^n$, the point $Y$ that minimizes $\sum_{i=1}^{k} |X^i - Y|^2$ is their centroid. Give an example when the centroid is not the optimal choice if we minimize sum of distances rather than squared distances.

The $k$-variance problem is thus the problem of partitioning a set of points into clusters so that the sum of the variances of the clusters is minimized.

We define a relaxation called the Continuous Clustering Problem (CCP), as the problem of finding the subspace $V$ of $\mathbb{R}^n$ of dimension at most $k$ which minimizes

$$
g_A(V) = \sum_{i=1}^{m} \text{dist}(A^{(i)}, V)^2.
$$

The reader will recognize that this is given by the SVD. It is easy to see that the optimal value of the $k$-variance problem is an upper bound for the optimal value of the CCP. Indeed for any set $B$ of $k$ points,

$$
f_A(B) \geq g_A(V_B) 
$$

(1.2)

where $V_B$ is the subspace generated by the points in $B$.

We now present a factor 2 approximation algorithm for the $k$-variance problem using the relaxation to the best-fit subspace. The algorithm has two parts. First we project to the $k$-dimensional SVD subspace. Then we solve the problem in the smaller dimensional space using a brute-force algorithm with the following guarantee.

**Theorem 1.5.** The $k$-variance problem can be solved in $O(mk^2d/2)$ time when the input $A \subseteq \mathbb{R}^d$.

We now give the idea and algorithm for the low-dimensional setting. Each set $B$ of “cluster centers” defines a Voronoi diagram where cell $C_i = \{X \in \mathbb{R}^d : |X - B^{(i)}| \leq |X - B^{(j)}| \text{ for } j \neq i\}$ consists of those points whose closest point in $B$ is $B^{(i)}$. Each cell is a polyhedron and the total number of faces in $C_1, C_2, \ldots, C_k$ is no more than $\binom{k}{2}$ (since each face is the set of points equidistant from two points of $B$).

We have seen in (1.1) that it is the partition of $A$ that determines the best $B$ (via computation of centroids) and so we can move the boundary hyperplanes of the optimal Voronoi diagram, without any face passing through a point of $A$, so that each face contains at least $d$ points of $A$. 
1.3. **THE K-VARIANCE PROBLEM**

Assume that the points of $A$ are in general position and $0 \notin A$ (a simple perturbation argument deals with the general case). This means that each face now contains $d$ affinely independent points of $A$. We have lost the information about which side of each face to place these points and so we must try all possibilities for each face. This leads to the following enumerative procedure for solving the $k$-variance problem:

**Algorithm for the $k$-variance problem in $d$ dimensions**

- Enumerate all $\sum_{t=k}^{\binom{n}{2}} \binom{n}{t} = O(m^{dk^2/2})$ sets of $k \leq t \leq k(k-1)/2$ hyperplanes, each of which contains $d$ affinely independent points of $A$.
- Check that the arrangement defined by these hyperplanes has exactly $k$ cells.
- Make one of $2^{kd}$ choices as to which cell to assign each point of $A$ which is lying on a hyperplane.
- This defines a unique partition of $A$. Find the centroid of each set in the partition and compute $f_A$.

As remarked previously, CCP can be solved by Linear Algebra. Indeed, let $V$ be a $k$-dimensional subspace of $\mathbb{R}^n$ and $\bar{A}^{(1)}, \bar{A}^{(2)}, \ldots, \bar{A}^{(m)}$ be the orthogonal projections of $A^{(1)}, A^{(2)}, \ldots, A^{(m)}$ onto $V$. Let $\bar{A}$ be the $m \times n$ matrix with rows $\bar{A}^{(1)}, \bar{A}^{(2)}, \ldots, \bar{A}^{(m)}$. Thus $\bar{A}$ has rank at most $k$ and

$$
\|A - \bar{A}\|^2_F = \sum_{i=1}^{m} |A^{(i)} - \bar{A}^{(i)}|^2 = \sum_{i=1}^{m} \left(\text{dist}(A^{(i)}, V)\right)^2.
$$

Thus to solve CCP, all we have to do is find the first $k$ vectors of the SVD of $A$ (since by Theorem (1.4), these minimize $\|A - \bar{A}\|^2_F$ over all rank $k$ matrices $\bar{A}$) and take the space $V_{SV D}$ spanned by the first $k$ singular vectors in the row space of $A$.

We now show that combining SVD with the above algorithm gives a 2-approximation to the $k$-variance problem in arbitrary dimension. Let $\bar{A} = \{\bar{A}^{(1)}, \bar{A}^{(2)}, \ldots, \bar{A}^{(m)}\}$ be the projection of $A$ onto the subspace $V_k$. Let $\bar{B} = \{\bar{B}^{(1)}, \bar{B}^{(2)}, \ldots, \bar{B}^{(k)}\}$ be the optimal solution to $k$-variance problem with input $\bar{A}$.

**Algorithm for the $k$-variance problem**

- Compute $V_k$.
- Solve the $k$-variance problem with input $\bar{A}$ to obtain $\bar{B}$.
- Output $\bar{B}$.

It follows from (1.2) that the optimal value $Z_A$ of the $k$-variance problem satisfies

$$
Z_A \geq \sum_{i=1}^{m} |A^{(i)} - \bar{A}^{(i)}|^2. \quad (1.3)
$$
Note also that if \( \hat{\mathcal{B}} = \{ \hat{B}(1), \hat{B}(2), \ldots, \hat{B}(k) \} \) is an optimal solution to the \( k \)-variance problem and \( \mathcal{B} \) consists of the projection of the points in \( \hat{\mathcal{B}} \) onto \( V \), then

\[
Z_A = \sum_{i=1}^{m} \text{dist}(A^{(i)}, \hat{\mathcal{B}})^2 \geq \sum_{i=1}^{m} \text{dist}(A^{(i)}, \hat{\mathcal{B}})^2 \geq \sum_{i=1}^{m} \text{dist}(\bar{A}^{(i)}, \mathcal{B})^2.
\]

Combining this with (1.3) we get

\[
2Z_A \geq \sum_{i=1}^{m} (|A^{(i)} - \bar{A}^{(i)}|^2 + \text{dist}(\bar{A}^{(i)}, \mathcal{B})^2)
= \sum_{i=1}^{m} \text{dist}(A^{(i)}, \mathcal{B})^2
= f_A(\hat{\mathcal{B}})
\]

proving that we do indeed get a 2-approximation.

Theorem 1.6. The algorithm above finds a factor 2 approximation for the \( k \)-variance problem for \( m \) points in \( \mathbb{R}^n \) in \( O(mn^2 + m^{1.5}) \) time.

1.4 Discussion

We reviewed basic concepts in linear algebra from a geometric perspective. The \( k \)-variance problem is a typical example of how SVD is used: project to the SVD subspace, then solve the original problem. In many application areas, the method known as “Principal Component Analysis” (PCA) focusses on the projection of a data matrix to the span of the largest singular vectors.

The following complexity questions are open: (1) Given a matrix \( A \), is it NP-hard to find a rank-\( k \) matrix \( D \) that minimizes the error with respect to the \( L_1 \) norm, \( \sum_{i,j} |A_{ij} - D_{ij}| \) is NP-hard (more generally for \( L_p \) norm for \( p \neq 2 \))? (2) Given a set of \( m \) points in \( \mathbb{R}^n \), is it NP-hard to find a subspace of dimension at most \( k \) that minimizes the sum of distances of the points to the subspace? (It is known that finding a subspace that minimizes the maximum distance is NP-hard [?]). (3) Is the \( k \)-variance problem NP-hard?
Chapter 2

Mixture Models

Let $F$ be a probability distribution in $\mathbb{R}^n$ with the property that it is a convex combination of distributions of known types, e.g., Gaussians. We can decompose $F$ as

$$F = w_1 F_1 + w_2 F_2 + \ldots + w_k F_k$$

where each $F_i$ is a probability distribution with mixing weight $w_i \geq 0$, and $\sum_i w_i = 1$. A random point drawn from $F$ originates from distribution $F_i$ with probability $w_i$.

Given a sample of points from $F$, we consider the following problems:

1. Classify the sample according to the component distributions.

2. Learn the component distributions (find their means, covariances, etc.).

For most of this chapter, we deal with the classical setting: each $F_i$ is a Gaussian in $\mathbb{R}^n$. In fact, we begin with the special case of spherical Gaussians whose density functions (i) depend only on the distance of a point from the mean and (ii) can be written as the product of density functions on each coordinate. The density function of a spherical Gaussian in $\mathbb{R}^n$ is

$$p(x) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\|x-\mu\|^2/2\sigma^2}$$

where $\mu$ is its mean and $\sigma$ is the standard deviation along any direction.

If the component distributions are far apart, so that points from one component distribution are closer to each other than to points from other components, then classification is straightforward. In the case of spherical Gaussians, making the means sufficiently far apart achieves this setting with high probability. On the other hand, if the component distributions have large overlap, then for a large fraction of the mixture, it is impossible to determine the origin of sample points. Thus, the classification problem is by nature tied to some assumption on the separability of the component distributions.
2.1 Probabilistic separation

In order to correctly identify sample points, we require a small overlap of distributions. How can we quantify the distance between distributions? One way, if we only have two distributions, is to take the total variation distance,

\[ d_{TV}(f_1, f_2) = \frac{1}{2} \int_{\mathbb{R}^n} |f_1(x) - f_2(x)| \, dx. \]

We can require this to be large for two well-separated distributions, i.e., \( d_{TV}(f_1, f_2) \geq 1 - \epsilon \), if we tolerate \( \epsilon \) error. We can incorporate mixing weights in this condition, allowing for two components to overlap more if the mixing weight of one of them is small:

\[ d_{TV}(f_1, f_2) = \int_{\mathbb{R}^n} |w_1 f_1(x) - w_2 f_2(x)| \, dx \geq 1 - \epsilon. \]

This can be generalized to

\[ \int_{\mathbb{R}^n} \left( 2 \max_i w_i f_i(x) - \sum_{i=1}^k w_i f_i(x) \right)^+ \, dx \geq 1 - \epsilon. \quad (2.1) \]

The quantity inside the integral is simply the maximum \( w_i f_i \) at \( x \), minus the sum of the rest of the \( w_i f_i \)'s. If the supports of the components are essentially disjoint, the integral will be 1.

This is a very general measure of distance between distributions. It is an open problem to deal with mixtures when we are only given the weak guarantee (2.1). In what follows, we will make a stronger assumption.

2.2 Geometric separation

Here we assume some separation between the means of component distributions. For two distributions, we require \( ||\mu_1 - \mu_2|| \) to be large compared to \( \max\{\sigma_1, \sigma_2\} \). Note this is a stronger assumption than that of small overlap. In fact, two distribution can have the same mean, yet still have small overlap, e.g., two spherical Gaussians with different variances.

Figure 2.1: Two distributions with the same mean, but small overlap.

Given a separation between the means, we expect that sample points originating from the same component distribution will have smaller pairwise distances than points originating from different distributions. Let \( X \) and \( Y \) be two independent samples drawn from the same \( F_i \).
2.2. GEOMETRIC SEPARATION

\[ \mathbb{E} (\|X - Y\|^2) = \mathbb{E} (\|(X - \mu_i) - (Y - \mu_i)\|^2) \]
\[ = 2\mathbb{E} (\|X - \mu_i\|^2) - 2\mathbb{E} ((X - \mu_i)(Y - \mu_i)) \]
\[ = 2\mathbb{E} (\|X - \mu_i\|^2) \]
\[ = 2\mathbb{E} \left( \sum_{j=1}^{n} |x^j - \mu_i^j|^2 \right) \]
\[ = 2n\sigma_i^2 \]

Next let \( X \) be from \( F_i \) and \( Y \) from \( F_j \).

\[ \mathbb{E} (\|X - Y\|^2) = \mathbb{E} (\|(X - \mu_i) - (Y - \mu_j) + (\mu_i - \mu_j)\|^2) \]
\[ = \mathbb{E} (\|X - \mu_i\|^2) + \mathbb{E} (\|Y - \mu_j\|^2) + \|\mu_i - \mu_j\|^2 \]
\[ = n\sigma_i^2 + n\sigma_j^2 + \|\mu_i - \mu_j\|^2 \]

Note how this value compares to the previous one. If \( \|\mu_i - \mu_j\|^2 \) were large enough, points in the component with smallest variance would all be closer to each other than to any point from the other components. This suggests that we can compute pairwise distances in our sample and use them to identify the subsample from the smallest component.

We consider separation of the form
\[ \|\mu_i - \mu_j\| \geq \beta \max\{\sigma_i, \sigma_j\}, \quad (2.2) \]

between every pair of means \( \mu_i, \mu_j \). For \( \beta \) large enough, the distance between points from different components will be larger in expectation than that between points from the same component. This suggests the following classification algorithm: we compute the distances between every pair of points, and connect those points whose distance is less than some threshold. The threshold is chosen to split the graph into two (or \( k \)) cliques. A similar algorithm is to calculate a minimum spanning tree of the graph (with edge weights equal to distances between points), and drop the heaviest edge (\( k-1 \) edges) so that the graph has two \( (k) \) connected components and each corresponds to a component distribution.

Both algorithms use only the pairwise distances. In order for any algorithm of this form to work, we need to turn the above arguments about expected distance between sample points into high probability bounds. For Gaussians, we can use the following concentration bound.

**Lemma 2.1.** Let \( X \) be drawn from a spherical Gaussian in \( \mathbb{R}^{\sigma^2} \) with mean \( \mu \) and variance \( \sigma^2 \) along any direction. Then for any \( \alpha > 1 \),
\[ \Pr(|\|X - \mu\|^2 - \sigma^2| > \alpha \sigma^2 \sqrt{n}) \leq 2e^{-\alpha^2/8}. \]
Using this lemma with $\alpha = 4\sqrt{\ln(m/\delta)}$, we have that with probability at least $1 - \delta$, for $X, Y$ from the $i$'th and $j$'th Gaussians respectively,

$$n\sigma_i^2 + n\sigma_j^2 - 4\sqrt{n\ln\left(\frac{m}{\delta}\right)}(\sigma_i^2 + \sigma_j^2)\|X - Y\|^2 \leq n\sigma_i^2 + 4\sqrt{n\ln\left(\frac{m}{\delta}\right)}(\sigma_i^2 + \sigma_j^2)$$

Thus it suffices for $\beta$ in the separation bound (2.2) to grow as $\Omega(\sqrt{n})$ for either of the above algorithms (clique or MST). One can be more careful and get a bound that grows only as $\Omega(n^{1/4})$ by identifying smaller components first as mentioned earlier.

The problem with these approaches is that the separation needed grows rapidly with $n$, the dimension, which in general is much higher than $k$, the number of components. On the other hand, for classification to be achievable with high probability, the separation does not need a dependence on $n$. In particular, it suffices for the means to be separated by a small number of standard deviations.

One way to reduce the dimension and therefore the dependence on $n$ is to project to a lower-dimensional subspace. A natural idea is random projection. Consider a projection from $\mathbb{R}^n \rightarrow \mathbb{R}^\ell$ so that the image of a point $u$ is $u'$. Then it can be shown that

$$E\left(||u'||^2\right) = \frac{\ell}{n}||u||^2$$

In other words, the expected squared length of a vector shrinks by a factor of $\frac{\ell}{n}$. Further, the squared length is concentrated around its expectation.

$$\Pr\left(||u'||^2 - \frac{\ell}{n}||u||^2 > \frac{\ell}{n}||u||^2\right) \leq 2e^{-\ell^2/4}$$

The problem with random projection is that the squared distance between the means, $||\mu_i - \mu_j||^2$, is also likely to shrink by the same $\frac{\ell}{n}$ factor, and therefore random projection provides no advantage.

### 2.3 Spectral Projection

Next we consider projecting on to the best-fit subspace given by the top $k$ singular vectors of the mixture. This is a general methodology — use principal component analysis (PCA) as a preprocessing step. In this case, it will be of great value provably.
2.3. SPECTRAL PROJECTION

Algorithm.

1. Compute the singular value decomposition of the sample matrix.
2. Project the samples to the rank \( k \) subspace spanned by the top \( k \) right singular vectors.
3. Perform a distance-based classification in the \( k \)-dimensional space.

We will see that by doing this, a separation of

\[
\|\mu_i - \mu_j\| \geq c(k \log m)^{\frac{1}{4}} \max\{\sigma_i, \sigma_j\},
\]

where \( c \) is an absolute constant, is sufficient for classifying \( m \) points.

The best-fit vector for a distribution is one that minimizes the expected squared distance of a random point to the vector. Using this definition, it is intuitive that the best fit vector for a single Gaussian is simply the vector that passes through the Gaussian’s mean. We state this formally below.

Lemma 2.2. The best fit 1-dimensional subspace for a spherical Gaussian with mean \( \mu \) is given by the vector passing through \( \mu \).

Proof. For a randomly chosen \( x \), we have

\[
E \left( (x \cdot v)^2 \right) = E \left( ((x - \mu) \cdot v + \mu \cdot v)^2 \right) \\
= E \left( ((x - \mu) \cdot v)^2 \right) + E \left( (\mu \cdot v)^2 \right) + 2E( (x - \mu) \cdot v)(\mu \cdot v)) \\
= \sigma^2 + (\mu \cdot v)^2 + 0 \\
= \sigma^2 + (\mu \cdot v)^2
\]

which is maximized when \( \mu = v \).

Further, due to the symmetry of the symmetry of the sphere, the best subspace of dimension 2 or more is any subspace containing the mean.

Lemma 2.3. The \( k \)-dimensional SVD subspace for a spherical Gaussian with mean \( \mu \) is any \( k \)-dimensional subspace containing \( \mu \).

A simple consequence of this lemma is the following theorem, which states that the best \( k \)-dimensional subspace for a mixture \( F \) involving \( k \) spherical Gaussians is the space which contains the means of the Gaussians.

Theorem 2.4. The \( k \)-dim SVD subspace for a mixture of \( k \) Gaussians \( F \) contains the span of \( \{\mu_1, \mu_2, \ldots, \mu_k\} \).

Now let \( F \) be a mixture of two Gaussians. Consider what happens when we project from \( \mathbb{R}^n \) onto the best two-dimensional subspace \( \mathbb{R}^2 \). The expected distance (after projection) of two points drawn from the same distribution goes
from $n\sigma_i^2$ to $2\sigma_i^2$. And, crucially, since we are projecting onto the best two-dimensional subspace which contains the two means, the expected value of $\|\mu_1 - \mu_2\|^2$ does not change!

What property of spherical Gaussians did we use in this analysis? A spherical Gaussian projected onto the best SVD subspace is still a spherical Gaussian. In fact, this only required that the variance in every direction is equal. But many other distributions, e.g., uniform over a cube, also have this property. We address the following questions in the rest of this chapter.

1. What distributions does Theorem 2.6 extend to?
2. What about more general distributions?
3. What is the sample complexity?

### 2.4 Weakly Isotropic Distributions

Next we study how our characterization of the SVD subspace can be extended.

**Definition 2.5.** Random variable $X \in \mathbb{R}^n$ has a weakly-isotropic distribution with mean $\mu$ if

$$E (w \cdot (X - \mu))^2 = \sigma^2, \quad \forall w \in \mathbb{R}^n, \: ||w|| = 1.$$ 

**Theorem 2.6.** The $k$-dimensional SVD subspace for a mixture $F$ with component means $\mu_1, \ldots, \mu_k$ is given by span $\mu_1, \ldots, \mu_k$ if each $F_i$ is weakly-isotropic and the means are in general position.

Theorem 2.6 requires the distributions to be weakly-isotropic. A spherical Gaussian is clearly weakly-isotropic. A uniform distribution in a cube is also weakly-isotropic.

**Exercise 2.1.** Show that a cube is weakly isotropic.

**Exercise 2.2.** Show that a distribution is weakly isotropic iff its covariance matrix is a multiple of the identity.

Theorem 2.6 is not always true for other distributions, even for $k = 1$. Consider a non-spherical Gaussian random vector $X \in \mathbb{R}^2$, whose distribution is like in Figure 2.2, where the variance along the $x$-axis is much larger than that along the $y$-axis. Clearly the optimal 1-dimensional subspace for $X$ (that maximizes the squared projection in expectation) is not the one passes through its mean $\mu$; it is orthogonal to the mean.

Figure 2.2: For a non-spherical Gaussian, the subspace containing the mean is not the best subspace.
2.5 Mixtures of general distributions

For a mixture of general distributions, the subspace that maximizes the squared projections is not the best subspace for our classification purpose any more. Consider two distributions that look like two parallel pancakes. We know there is a plane that separates them, but we do not know how to find the plane. The 2-dimensional subspace that maximizes the squared projections is the one parallel to the two pancakes. But after projection to this subspace, the means collapse and we can not separate the two distributions anymore.

Figure 2.3: Two distributions that are collapsed by spectral projection.

The next theorem provides an extension of the analysis of spherical Gaussians by showing when the SVD subspace is "close" the subspace spanned by the component means.

**Theorem 2.7.** Let $F$ be a mixture of arbitrary distributions $F_1, \ldots, F_k$. Let $w_i$ be the mixing weight of $F_i$, $\mu_i$ be its mean and $\sigma^2_{i,W}$ be the maximum variance of $F_i$ along directions in $W$, the $k$-dimensional SVD-subspace of $F$. Then

$$
\sum_{i=1}^{k} w_i d(\mu_i, W)^2 \leq \sum_{i=1}^{k} w_i \sigma^2_{i,W}
$$

where $d(\cdot, \cdot)$ is the orthogonal distance.

Theorem 2.7 says that for a mixture of general distributions, the means do not move too much after projection. The theorem is true for any mixture, thus is true for samples, with distribution means and variances replaced by sample means and variances.

**Proof.** Let $M$ be the span of $\mu_1, \mu_2, \ldots, \mu_k$. For $x \in \mathbb{R}^n$, we write $\pi_M(x)$ for the projection of $x$ to the subspace $M$ and $\pi_W(x)$ for the projection of $x$ to $W$.

We first lower bound the expected squared length of the projection to the mean subspace $M$.

$$
\mathbb{E} \left( \|\pi_M(x)\|^2 \right) = \sum_{i=1}^{k} w_i \mathbb{E}_{F_i} \left( \|\pi_M(x)\|^2 \right)
$$

$$
= \sum_{i=1}^{k} w_i \left( \mathbb{E}_{F_i} \left( \|\pi_M(x) - \mu_i\|^2 \right) + \|\mu_i\|^2 \right)
$$

$$
\geq \sum_{i=1}^{k} w_i \|\mu_i\|^2
$$

$$
= \sum_{i=1}^{k} w_i \|\pi_W(\mu_i)\|^2 + \sum_{i=1}^{k} w_i d(\mu_i, W)^2.
$$
We next upper bound the expected squared length of the projection to the SVD subspace $W$. Let $\vec{e}_1, ..., \vec{e}_k$ be an orthonormal basis for $W$.

$$
E \left( \|\pi_W(x)\|^2 \right) = \sum_{i=1}^{k} w_i \left( E_{F_i} \left( \|\pi_W(x - \mu_i)\|^2 \right) + \|\pi_W(\mu_i)\|^2 \right)
$$

$$
\leq \sum_{i=1}^{k} w_i \sum_{j=1}^{k} E_{F_i} \left( (\pi_W(x - \mu_i) \cdot \vec{e}_j)^2 \right) + \sum_{i=1}^{k} w_i \|\pi_W(\mu_i)\|^2
$$

$$
\leq k \sum_{i=1}^{k} w_i \hat{\sigma}_{i,W}^2 + \sum_{i=1}^{k} w_i \|\pi_W(\mu_i)\|^2.
$$

The SVD subspace maximizes the sum of squared projections among all subspaces of rank at most $k$ (Theorem 1.3). Therefore,

$$
E \left( \|\pi_M(x)\|^2 \right) \leq E \left( \|\pi_W(x)\|^2 \right)
$$

and the theorem follows from the previous two inequalities.

### 2.6 Spectral projection with samples

So far we have shown that the SVD subspace of a mixture can be quite useful for classification. In reality, we only have samples from the mixture. This section is devoted to establishing bounds on sample complexity to achieve similar guarantees as we would for the full mixture. The main tool will be distance concentration of samples. In general, we are interested in inequalities such as the following for a random point $X$ from a component $F_i$ of the mixture:

$$
\Pr \left( |(X - \mu_i) \cdot v_i| > t \sigma_i \right) \leq e^{-ct}.
$$

This is useful for two reasons:

1. To ensure that the subspace computed from SVD on the sample matrix has similar properties as the SVD subspace for the full mixture.

2. To be able to apply simple clustering algorithms such as forming cliques or connected components, we need distances between points of the same component to be close to their expectation.

An interesting class of distributions with such concentration properties are those whose probability density functions are **logconcave**. A function $f$ is log-concave if $\forall x, y, \forall \lambda \in [0, 1],$

$$
f(\lambda x + (1 - \lambda)y) \geq f(x)^\lambda f(y)^{1-\lambda}
$$

or equivalently,

$$
\log f(\lambda x + (1 - \lambda)y) \geq \lambda \log f(x) + (1 - \lambda) \log f(y).
$$
Gaussian is logconcave. In fact, any distribution with a density function \( f(x) = e^{g(x)} \) for some concave function \( g(x) \), e.g. \( e^{-c\|x\|} \) or \( e^{c(x \cdot v)} \) is logconcave. Also any uniform distribution in a convex body is logconcave.

2.7 Discussion

Mixture models are a classical topic in statistics. Traditional methods such as EM or other local search heuristics can get stuck in local optima or take a long time to converge. Starting with Dasgupta’s paper [], there has been much progress on efficient algorithms with rigorous guarantees. PCA was analyzed in this context by Vempala and Wang [?, ?] giving the current best guarantees for mixtures of spherical Gaussians (and weakly isotropic distributions). While this has been extended to general Gaussians, the current bounds are far from optimal in that the separation required grows with the largest variance of the components or with the dimension of the underlying space.

An even more general question is “agnostic” learning of Gaussians, where we are given samples from an arbitrary distribution and would like to find the best-fit mixture of \( k \) Gaussians. This problem naturally accounts for noise and appears to be much more realistic.
Part II

Algorithms