



University of Pennsylvania
ScholarlyCommons

Statistics Papers

Wharton Faculty Research

2009

Multi-View Clustering via Canonical Correlation Analysis

Kamalika Chaudhuri

Sham M. Kakade
University of Pennsylvania

Karen Livescu

Karthik Sridharan

Follow this and additional works at: http://repository.upenn.edu/statistics_papers

 Part of the [Statistics and Probability Commons](#), and the [Theory and Algorithms Commons](#)

Recommended Citation

Chaudhuri, K., Kakade, S. M., Livescu, K., & Sridharan, K. (2009). Multi-View Clustering via Canonical Correlation Analysis. *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09* 129-136. <http://dx.doi.org/10.1145/1553374.1553391>

At the time of publication, author Sham M. Kakade was affiliated with Toyota Technological Institute at Chicago. Currently, he is a faculty member at the Statistics Department at the University of Pennsylvania.

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/statistics_papers/193
For more information, please contact repository@pobox.upenn.edu.

Multi-View Clustering via Canonical Correlation Analysis

Abstract

Clustering data in high-dimensions is believed to be a hard problem in general. A number of efficient clustering algorithms developed in recent years address this problem by projecting the data into a lower-dimensional subspace, e.g. via Principal Components Analysis (PCA) or random projections, before clustering. Such techniques typically require stringent requirements on the separation between the cluster means (in order for the algorithm to be successful).

Here, we show how using multiple views of the data can relax these stringent requirements. We use Canonical Correlation Analysis (CCA) to project the data in each view to a lower-dimensional subspace. Under the assumption that conditioned on the cluster label the views are uncorrelated, we show that the separation conditions required for the algorithm to be successful are rather mild (significantly weaker than those of prior results in the literature). We provide results for mixture of Gaussians, mixtures of log concave distributions, and mixtures of product distributions.

Disciplines

Statistics and Probability | Theory and Algorithms

Comments

At the time of publication, author Sham M. Kakade was affiliated with Toyota Technological Institute at Chicago. Currently, he is a faculty member at the Statistics Department at the University of Pennsylvania.



Technical Report
TTI-TR-2008-5

Multi-View Clustering via Canonical Correlation Analysis

Kamalika Chaudhuri

UC San Diego

Sham M. Kakade

Toyota Technological Institute at Chicago

ABSTRACT

Clustering data in high-dimensions is believed to be a hard problem in general. A number of efficient clustering algorithms developed in recent years address this problem by projecting the data into a lower-dimensional subspace, e.g. via Principal Components Analysis (PCA) or random projections, before clustering. Such techniques typically require stringent requirements on the separation between the cluster means (in order for the algorithm to be successful).

Here, we show how using multiple views of the data can relax these stringent requirements. We use Canonical Correlation Analysis (CCA) to project the data in each view to a lower-dimensional subspace. Under the assumption that conditioned on the cluster label the views are uncorrelated, we show that the separation conditions required for the algorithm to be successful are rather mild (significantly weaker than those of prior results in the literature). We provide results for mixture of Gaussians, mixtures of log concave distributions, and mixtures of product distributions.

1 Introduction

The multi-view approach to learning is one in which we have ‘views’ of the data (sometimes in a rather abstract sense) and, if we understand the underlying relationship between these views, the hope is that this relationship can be used to alleviate the difficulty of a learning problem of interest [BM98, KF07, AZ07]. In this work, we explore how having ‘two views’ of the data makes the clustering problem significantly more tractable.

Much recent work has gone into understanding under what conditions we can learn a mixture model. The basic problem is as follows: we obtain iid samples from a mixture of k distributions and our task is to either: 1) infer properties of the underlying mixture model (e.g. the mixing weights, means, etc) or 2) classify a random sample according to which distribution it was generated from.

Under no restrictions on the underlying distribution, this problem is considered to be hard. However, in many applications, we are only interested in clustering the data when the component distribution are “well separated”. In fact, the focus of recent clustering algorithms [Das99, VW02, AM05, BV08] is on efficiently learning with as little separation as possible. Typically, these separation conditions are such that when given a random sample from the mixture model, the Bayes optimal classifier is able to reliably (with high probability) recover which cluster generated that point.

This work assumes a rather natural multi-view assumption: the assumption is that the views are (conditionally) uncorrelated, if we condition on which mixture distribution generated the views. There are many natural applications for which this assumption is applicable. For example, we can consider multi-modal views, with one view being a video stream and the other an audio stream — here conditioned on the speaker identity and maybe the phoneme (both of which could label the generating cluster), the views may be uncorrelated.

Under this multi-view assumption, we provide a simple and efficient subspace learning method, based on Canonical Correlation Analysis (CCA). This algorithm is affine invariant and is able to learn with some of the weakest separation conditions to date. The intuitive reason for this is that under our multi-view assumption, we are able to (approximately) find the subspace spanned by the means of the component distributions. Furthermore, the number of samples we need scales as $O(d)$, where d is the ambient dimension.

This shows how the multi-view framework can provide substantial improvements to the clustering problem, adding to the growing body of results which show how the multi-view framework can alleviate the difficulty of learning problems.

1.1 Related Work

Most of the provably efficient clustering algorithms first project the data down to some low dimensional space and then cluster the data in this lower dimensional space (typically, an algorithm such as single linkage suffices here). This projection is typically achieved either randomly or by a spectral method.

One of the first provably efficient algorithms for learning mixture models is due to Dasgupta [Das99], who learns a mixture of spherical Gaussians by *randomly* projecting the mixture onto a low-dimensional subspace. The separation requirement for the algorithm in [Das99] is $\sigma\sqrt{d}$, where d is the dimension and σ is the *maximal* directional standard deviation (the maximum variance in any direction of one of the component distributions). Vempala and Wang [VW02] removes the dependence of the separation requirement on the dimension of the data: given a mixture of k spherical Gaussians, they project the mixture down to the k -dimensional subspace of highest variance, and as a result, they can learn such mixtures with a separation of $\sigma k^{1/4}$. [KSV05] and [AM05] extend this result to a mixture of general Gaussians; however, for their algorithm to work correctly, they require a separation of $\frac{\sigma}{\sqrt{w_{\min}}}$, where again σ is the maximum directional standard deviation in any direction, and w_{\min} is the minimum mixing weight. [CR08] use a canonical-correlations based

algorithm to learn mixtures of axis-aligned Gaussians with a separation of about $\sigma^* \sqrt{k}$, where σ^* is the maximum directional standard deviation in the subspace containing the centers of the distributions. However, their algorithm requires the coordinate-independence property, and requires an additional “spreading” condition, which states that the separation between any two centers should be spread along many coordinates. All these algorithms are not affine invariant (and one can show that a linear transformation of the data could cause these algorithms to fail).

Finally, [BV08] provides an affine-invariant algorithm for learning mixtures of general Gaussians, so long as the mixture has a suitably low Fisher coefficient when in isotropic position. Their result is stronger than the results of [KSV05, AM05, Das99], and more general than [CR08]; however, their implied separation condition involves a rather large polynomial dependence on $\frac{1}{w_{\min}}$. Other than these, there are also some provably efficient algorithms which do not use projections — such as [DS00] and [AK05].

The two results most closely related to ours are the work of [VW02] and [CR08].

1. [VW02] shows that it is sufficient to find the subspace spanned by the means of the distributions in the mixture for effective clustering.
2. [CR08] is related to ours, because they use a projection onto the top k singular value decomposition subspace of the canonical correlations matrix. They also provide a *spreading condition*, which is related to the requirement on the rank in our work.

We borrow techniques from both of these papers.

1.2 This Work

In this paper, we study the problem of *multi-view clustering*. In our setting, we have data on a fixed set of objects from two separate sources, which we call the two *views*, and our goal is to use this data to cluster more effectively than with data from a single source.

The conditions required by our algorithm are as follows. In the sequel, we assume that the mixture is in an isotropic position in each view individually. First, we require that conditioned on the source distribution in the mixture, the two views are uncorrelated. Notice that this condition allows the distributions in the mixture within each view to be completely general, so long as they are uncorrelated across views. Second, we require the rank of the CCA matrix across the views to be at least k , and the k -th singular value of this matrix to be at least λ_{\min} . This condition ensures that there is sufficient correlation between the views, and if this condition holds, then, we can recover the subspace containing the means of the distributions in both views.

In addition, for the case of mixture of Gaussians, if in at least one view, say view 1, we have that for every pair of distributions i and j in the mixture,

$$\|\mu_i^1 - \mu_j^1\| > C\sigma^* k^{1/4} \sqrt{\log(n/\delta)}$$

for some constant C , where μ_i^1 is the mean of the i -th component distribution in view one and σ^* is the maximum directional standard deviation in the subspace containing the means of the distributions in view 1, then our algorithm can also cluster correctly, which means that it can determine which distribution each sample came from.

This separation condition is considerably weaker than previous results in that σ^* only depends on the directional variance in the subspace spanned by the means — as opposed to directional variance over all directions. Also, the only other affine invariant algorithm is that in [BV08] — while this result does not explicitly state results in terms of separation between the means (it uses a Fisher coefficient concept), the implied separation is rather large.

We stress that our improved results are really due the multi-view condition. We also emphasize that for our algorithm to cluster successfully, it is sufficient for the distributions in the mixture to obey the separation condition in *one view*; so long as our rank condition holds, and the separation

condition holds in one view, our algorithm produces a correct clustering of the input data in that view.

2 The Setting

We assume that our data is generated by a mixture of k distributions. In particular, we assume we obtain samples $x = (x^{(1)}, x^{(2)})$, where $x^{(1)}$ and $x^{(2)}$ are the two views of the data, which live in the vector spaces \mathcal{V}_1 of dimension d_1 and \mathcal{V}_2 of dimension d_2 , respectively. We let $d = d_1 + d_2$. Let μ_i^j , for $i = 1, \dots, k$ and $j = 1, 2$ be the center of distribution i in view j , and let w_i be the mixing weight for distribution i . Let w_i be the probability of cluster i .

For simplicity, assume that data have mean 0. We denote the covariance matrix of the data as:

$$\Sigma = E[xx^\top], \Sigma_{11} = E[x^{(1)}(x^{(1)})^\top], \Sigma_{22} = E[x^{(2)}(x^{(2)})^\top], \Sigma_{12} = E[x^{(1)}(x^{(2)})^\top]$$

Hence, we have:

$$\Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{21} \\ \hline \Sigma_{12} & \Sigma_{22} \end{array} \right]. \quad (1)$$

The multi-view assumption we work with is as follows:

Assumption 1 (*Multi-View Condition*) *We assume that conditioned on the source distribution s in the mixture (where $s = i$ with probability w_i), the two views are uncorrelated. More precisely, we assume that:*

$$E[x^{(1)}(x^{(2)})^\top | s = i] = E[x^{(1)} | s = i]E[(x^{(2)})^\top | s = i]$$

for all $i \in [k]$.

This assumption implies that:

$$\Sigma_{12} = \sum_i w_i \mu_i^1 \cdot (\mu_i^2)^T.$$

To see this, observe that

$$\begin{aligned} E[x^{(1)}(x^{(2)})^\top] &= \sum_i E_{D_i}[x^{(1)}(x^{(2)})^\top] \Pr[D_i] \\ &= \sum_i w_i E_{D_i}[x^{(1)}] \cdot E_{D_i}[(x^{(2)})^\top] \\ &= \sum_i w_i \mu_i^1 \cdot (\mu_i^2)^T \end{aligned} \quad (2)$$

As the distributions are in isotropic position, we observe that $\sum_i w_i \mu_i^1 = \sum_i w_i \mu_i^2 = 0$. Therefore, the above equation shows that the rank of Σ_{12} is at most $k - 1$. We now assume that it has rank precisely $k - 1$.

Assumption 2 (*Non-Degeneracy Condition*) *We assume that Σ_{12} has rank $k - 1$ and that the minimal non-zero singular value of Σ_{12} is $\lambda_{\min} > 0$ (where we are working in a coordinate system where Σ_{11} and Σ_{22} are identity matrices).*

For clarity of exposition, we also work in a isotropic coordinate system, in each view. Specifically, the expected covariance matrix of the data, in each view, is the identity matrix, i.e.

$$\Sigma_{11} = I_{d_1}, \Sigma_{22} = I_{d_2}$$

As our analysis shows, our algorithm is robust to errors, so we assume that data is whitened as a pre-processing step.

One way to view the Non-Degeneracy Assumption is in terms of correlation coefficients. Recall that for two directions $u \in \mathcal{V}_1$ and $v \in \mathcal{V}_2$, the correlation coefficient is defined as:

$$\rho(u, v) = \frac{\mathbf{E}[(u \cdot x^{(1)})(v \cdot x^{(2)})]}{\sqrt{\mathbf{E}[(u \cdot x^{(1)})^2]\mathbf{E}[(v \cdot x^{(2)})^2]}}.$$

An alternative definition of λ_{\min} is just the minimal non-zero, correlation coefficient i.e.

$$\lambda_{\min} = \min_{u, v: \rho(u, v) \neq 0} \rho(u, v).$$

Note $1 \geq \lambda_{\min} > 0$.

We use $\widehat{\Sigma}_{11}$ and $\widehat{\Sigma}_{22}$ to denote the sample covariance matrix in views 1 and 2 respectively. We use $\widehat{\Sigma}_{12}$ to denote the sample covariance matrix combined across views 1 and 2. We assumed these are obtained through empirical averages from i.i.d. samples from the underlying distribution.

For any matrix A , we use $\|A\|$ to denote the L_2 norm or maximum singular value of A .

2.1 A Summary of Our Results

The following lemma provide the intuition for our algorithm:

Lemma 1 *Under Assumption 2, if U, D, V is the ‘thin’ SVD of Σ_{12} (where the thin SVD removes all zero entries from the diagonal), then the subspace spanned by the means in view 1 is precisely the column span of U (and we have the analogous statement for view 2).*

This follows directly from Equation 2 and the rank assumption. Essentially, our algorithm uses a CCA to (approximately) project the data down to the subspace spanned by the means.

Our main theorem can be stated as follows.

Theorem 1 (Gaussians) *Suppose the source distribution is a mixture of Gaussians, and suppose Assumptions 1 and 2 hold. Let σ^* be the maximum directional standard deviation of any distribution in the subspace spanned by $\{\mu_i^1\}_{i=1}^k$. If, for each pair i and j and for a fixed constant C ,*

$$\|\mu_i^1 - \mu_j^1\| \geq C\sigma^*k^{1/4}\sqrt{\log\left(\frac{kn}{\delta}\right)}$$

then, with probability $1 - \delta$, Algorithm 1 correctly classifies the examples if the number of examples used is

$$c \cdot \frac{d}{(\sigma^*)^2 \lambda_{\min}^2 w_{\min}^2} \log^2\left(\frac{d}{\sigma^* \lambda_{\min} w_{\min}}\right) \log^2(1/\delta)$$

for some constant c .

Here we assume that a separation condition holds in View 1, but a similar theorem also applies to View 2.

Our next theorem is for mixtures of log-concave distributions, with a different separation condition, which is larger in terms of k (due to the different concentration properties of log-concave distributions).

Theorem 2 (Log-concave Distributions) *Suppose the source distribution is a mixture of log-concave distributions, and suppose Assumptions 1 and 2 hold. Let σ^* be the maximum directional standard deviation of any distribution in the subspace spanned by $\{\mu_i^1\}_{i=1}^k$. If, for each pair i and j and for a fixed constant C ,*

$$\|\mu_i^1 - \mu_j^1\| \geq C\sigma^*\sqrt{k}\log\left(\frac{kn}{\delta}\right)$$

then, with probability $1 - \delta$, Algorithm 1 correctly classifies the examples if the number of examples used is

$$c \cdot \frac{d}{(\sigma^*)^2 \lambda_{\min}^2 w_{\min}^2} \log^3\left(\frac{d}{\sigma^* \lambda_{\min} w_{\min}}\right) \log^2(1/\delta)$$

for some constant c .

An analogous theorem also holds for mixtures of product distributions, with the same k dependence as the case for mixtures of log-concave distributions.

Theorem 3 (Product Distributions) *Suppose Assumptions 1 and 2 hold, and suppose that the distributions D_i in the mixture are product distributions, in which each coordinate has range at most R . If, for each pair i and j and for a fixed constant C ,*

$$\|\mu_i^1 - \mu_j^1\| \geq CR \sqrt{k \log\left(\frac{kn}{\delta}\right)}$$

then, with probability $1 - \delta$, Algorithm 1 correctly classifies the examples if the number of examples used is

$$c \cdot \frac{d}{\lambda_{\min}^2 w_{\min}} \log^2\left(\frac{d}{\sigma^* \lambda_{\min} w_{\min}}\right) \log^2(1/\delta)$$

where c is a constant.

3 Clustering Algorithms

In this section, we present our clustering algorithm, which clusters correctly with high probability, when the data in at least one of the views obeys a separation condition, in addition to our assumptions.

Our clustering algorithm is as follows. The input to the algorithm is a set of samples S , and a number k , and the output is a clustering of these samples into k clusters. For this algorithm, we assume that the data obeys the separation condition in View 1; an analogous algorithm can be applied when the data obeys the separation condition in View 2 as well.

Algorithm 1.

1. Randomly partition S into two subsets of equal size A and B .
2. Let $\widehat{\Sigma}_{12}(A)$ ($\widehat{\Sigma}_{12}(B)$ respectively) denote the empirical covariance matrix between views 1 and 2, computed from the sample set A (B respectively). Compute the top $k - 1$ left singular vectors of $\widehat{\Sigma}_{12}(A)$ ($\widehat{\Sigma}_{12}(B)$ respectively), and project the samples in B (A respectively) on the subspace spanned by these vectors.
3. Apply single linkage clustering (for mixtures of product distributions and log-concave distributions), or the algorithm in Section 3.5 of [AK05] (for mixtures of Gaussians) on the projected examples in View 1.

4 Proofs

In this section, we present the proofs of our main Theorem. First, the following two lemmas are useful, whose proofs follow directly from our assumptions.

We use S^1 (resp. S^2) to denote the subspace of \mathcal{V}_1 (resp. \mathcal{V}_2) spanned by $\{\mu_i^1\}_{i=1}^k$ (resp. $\{\mu_i^2\}_{i=1}^k$). We use S'^1 (resp. S'^2) to denote the orthogonal complement of S^1 (resp. S^2) in \mathcal{V}_1 (resp. \mathcal{V}_2).

Lemma 2 Let v^1 and v^2 be any vectors in S^1 and S^2 respectively. Then,

$$|(v^1)^T \Sigma_{12} v^2| > \lambda_{\min}$$

Lemma 3 Let v^1 (resp. v^2) be any vector in S^{r_1} (resp. S^{r_2}). Then, for any $u^1 \in \mathcal{V}_1$ and $u_2 \in \mathcal{V}_2$,

$$(v^1)^T \Sigma_{12} u^2 = (u^1)^T \Sigma_{12} v^2 = 0$$

Lemma 4 (Sample Complexity Lemma) If the number of samples

$$n > c \cdot \frac{d}{\epsilon^2 w_{\min}} \log^2\left(\frac{d}{\epsilon w_{\min}}\right) \log^2\left(\frac{1}{\delta}\right)$$

for some constant c , then, with probability at least $1 - \delta$,

$$\|\widehat{\Sigma}_{12} - \Sigma_{12}\| \leq \epsilon$$

where $\|\cdot\|$ denotes the L_2 -norm of a matrix (equivalent to the maximum singular value).

PROOF: To prove this lemma, we apply Lemma 5. Observe the block representation of Σ in Equation 1. Moreover, with Σ_{11} and Σ_{22} in isotropic position, we have that the L_2 norm of Σ_{12} is at most 1. Using the triangle inequality, we can write:

$$\|\widehat{\Sigma}_{12} - \Sigma_{12}\| \leq \frac{1}{2}(\|\widehat{\Sigma} - \Sigma\| + \|\widehat{\Sigma}_{11} - \Sigma_{11}\| + \|\widehat{\Sigma}_{22} - \Sigma_{22}\|)$$

(where we have applied the triangle inequality to the 2x2 block matrix with off-diagonal entries $\widehat{\Sigma}_{12} - \Sigma_{12}$ and with 0 diagonal entries). We now apply Lemma 5 three times, on $\widehat{\Sigma}_{11} - \Sigma_{11}$, on $\widehat{\Sigma}_{22} - \Sigma_{22}$ and a scaled version of $\widehat{\Sigma} - \Sigma$. The first two applications follow directly.

For the third application, we observe that Lemma 5 is rotation invariant, and that scaling each covariance value by some factor s scales the norm of the matrix by at most s . We claim that we can apply Lemma 5 on $\widehat{\Sigma} - \Sigma$ with $s = 4$. Since the covariance of any two random variables is at most the product of their standard deviations, and since Σ_{11} and Σ_{22} are I_{d_1} and I_{d_2} respectively, the maximum singular value of Σ_{12} is at most 1; the maximum singular value of Σ is therefore at most 4. Our claim follows.

The lemma now follows by plugging in n as a function of ϵ , d and w_{\min} \square

Lemma 5 Let X be a set of n points generated by a mixture of k Gaussians over \mathbb{R}^d , scaled such that $\mathbf{E}[x \cdot x^T] = I_d$. If M is the sample covariance matrix of X , then, for n large enough, with probability at least $1 - \delta$,

$$\|M - \mathbf{E}[M]\| \leq C \cdot \frac{\sqrt{d \log n \log\left(\frac{2n}{\delta}\right) \log(1/\delta)}}{\sqrt{w_{\min} n}}$$

where C is any constant, and w_{\min} is the minimum mixing weight of any Gaussian in the mixture.

PROOF: To prove this lemma, we use a concentration result on the L_2 -norms of matrices due to [RV07]. We observe that each vector x_i in the scaled space is generated by a Gaussian with some mean μ and maximum directional variance σ^2 . As the total variance of the mixture along any direction is at most 1,

$$w_{\min}(\mu^2 + \sigma^2) \leq 1 \tag{3}$$

Therefore, for all samples x_i , with probability at least $1 - \delta/2$,

$$\|x_i\| \leq \|\mu\| + \sigma \sqrt{d \log\left(\frac{2n}{\delta}\right)} \quad (4)$$

We condition on the fact that the event $\|x_i\| \leq \|\mu\| + \sigma \sqrt{d \log\left(\frac{2n}{\delta}\right)}$ happens for all $i = 1, \dots, n$. The probability of this event is at least $1 - \delta/2$.

Conditioned on this event, the distributions of the vectors x_i are independent. Therefore, we can apply Theorem 3.1 in [RV07] on these conditional distributions, to conclude that:

$$\Pr[||M - \mathbf{E}[M]|| > t] \leq 2e^{-cnt^2/\Lambda^2 \log n}$$

where c is a constant, and Λ is an upper bound on the norm of any vector $\|x_i\|$. Plugging in $t = \sqrt{\frac{\Lambda^2 \log(4/\delta) \log n}{cn}}$, we get that conditioned on the event $\|x_i\| \leq \Lambda$, with probability $1 - \delta/2$, $||M - \mathbf{E}[M]|| \leq \sqrt{\frac{\Lambda^2 \log(4/\delta) \log n}{cn}}$. From Equations 3 and 4, we get that with probability $1 - \delta/2$, for all the samples, $\Lambda \leq \frac{2\sqrt{d \log(2n/\delta)}}{\sqrt{w_{\min}}}$. Therefore, with probability $1 - \delta$,

$$||M - \mathbf{E}[M]|| \leq C \frac{\sqrt{d \log n \log(2n/\delta) \log(4/\delta)}}{\sqrt{nw_{\min}}}$$

which completes the proof. \square

Lemma 6 (Projection Subspace Lemma) *Let v^1 (resp. v^2) be any vector in S^1 (resp. S^2). If the number of samples $n > c \frac{d}{\tau^2 \lambda_{\min}^2 w_{\min}} \log^2\left(\frac{d}{\tau \lambda_{\min} w_{\min}}\right) \log^2\left(\frac{1}{\delta}\right)$ for some constant c , then, with probability $1 - \delta$, the length of the projection of v^1 (resp. v^2) in the subspace spanned by the top $k - 1$ left (resp. right) singular vectors of $\widehat{\Sigma}_{12}$ is at least $\sqrt{1 - \tau^2} \|v^1\|$ (resp. $\sqrt{1 - \tau^2} \|v^2\|$).*

PROOF: For the sake of contradiction, suppose there exists a vector $v^1 \in S^1$ such that the projection of v^1 on the top $k - 1$ left singular vectors of $\widehat{\Sigma}_{12}$ is equal to $\sqrt{1 - \tau^2} \|v^1\|$, where $\tilde{\tau} > \tau$. Then, there exists some unit vector u^1 in \mathcal{V}_1 in the orthogonal complement of the space spanned by the top $k - 1$ left singular vectors of $\widehat{\Sigma}_{12}$ such that the projection of v^1 on u^1 is equal to $\tilde{\tau} \|v^1\|$. Since the projection of v^1 on u^1 is at least $\tilde{\tau}$, and u^1 and v^1 are both unit vectors, u^1 can be written as: $u^1 = \tilde{\tau} v^1 + (1 - \tilde{\tau}^2)^{1/2} y^1$, where y^1 is in the orthogonal complement of S^1 . From Lemma 2, there exists some vector u^2 in S^2 , such that $(v^1)^\top \Sigma_{12} u^2 \geq \lambda_{\min}$; from Lemma 3, as y^1 is in the orthogonal complement of S^1 , for this vector u^2 , $(u^1)^\top \Sigma_{12} u^2 \geq \tilde{\tau} \lambda_{\min}$. If $n > c \frac{d}{\tilde{\tau}^2 \lambda_{\min}^2 w_{\min}} \log^2\left(\frac{d}{\tilde{\tau} \lambda_{\min} w_{\min}}\right) \log^2\left(\frac{1}{\delta}\right)$, for some constant c , then, from Lemma 7, $(u^1)^\top \widehat{\Sigma}_{12} u^2 \geq \frac{\tilde{\tau}}{2} \lambda_{\min}$.

Now, since u_1 is in the orthogonal complement of the subspace spanned by the top $k - 1$ left singular vectors of $\widehat{\Sigma}_{12}$, for any vector y^2 in the subspace spanned by the top $k - 1$ right singular vectors of $\widehat{\Sigma}_{12}$, $(u_1)^\top \widehat{\Sigma}_{12} y^2 = 0$. This follows from the properties of the singular space of any matrix. This, in turn, means that there exists a vector z^2 in \mathcal{V}_2 in the orthogonal complement of the subspace spanned by the top $k - 1$ right singular vectors of $\widehat{\Sigma}_{12}$ such that $(u^1)^\top \widehat{\Sigma}_{12} z^2 \geq \frac{\tilde{\tau}}{2} \lambda_{\min}$. This implies that the k -th singular value of $\widehat{\Sigma}_{12}$ is at least $\frac{\tilde{\tau}}{2} \lambda_{\min}$.

However, if $n > c \frac{d}{\tau^2 \lambda_{\min}^2 w_{\min}} \log^2\left(\frac{d}{\tau \lambda_{\min} w_{\min}}\right) \log^2\left(\frac{1}{\delta}\right)$, for some constant c , then, from Lemma 7, all except the top $k - 1$ singular values of $\widehat{\Sigma}_{12}$ are at most $\frac{\tau}{3} \lambda_{\min}$, which is a contradiction. \square

Lemma 7 *Let $n > C \frac{d}{\epsilon^2 w_{\min}} \log^2\left(\frac{d}{\epsilon w_{\min}}\right) \log^2\left(\frac{1}{\delta}\right)$, for some constant C . Then, with probability $1 - \delta$, the top $k - 1$ singular values of $\widehat{\Sigma}_{12}$ have value at least $\lambda_{\min} - \epsilon$. The remaining $\min(d_1, d_2) - k + 1$ singular values of $\widehat{\Sigma}_{12}$ have value at most ϵ .*

PROOF: From Lemmas 2 and 3, Σ_{12} has rank exactly $k - 1$, and the $k - 1$ -th singular value of Σ_{12} is at least λ_{\min} . Let e_1, \dots, e_{k-1} and g_1, \dots, g_{k-1} be the top $k - 1$ left and right singular vectors of Σ_{12} . Then, using Lemma 2, for any vectors e and g in the subspaces spanned by e_1, \dots, e_{k-1} and g_1, \dots, g_{k-1} respectively, $e^T \widehat{\Sigma}_{12} g \geq e^T \Sigma_{12} g - |e^T (\widehat{\Sigma}_{12} - \Sigma_{12}) g| \geq \lambda_{\min} - \epsilon$. Therefore, there is a subspace of rank $k - 1$ with singular values at least $\lambda_{\min} - \epsilon$ from which the first part of the lemma follows.

The second part of the lemma follows similarly from the fact that Σ_{12} has a $\min(d_1, d_2) - k + 1$ dimensional subspace with singular value 0. \square

Now we are ready to prove our main theorem.

PROOF:(Of Theorem 1) From Lemma 6, if $n > C \cdot \frac{d}{\tau^2 \lambda_{\min}^2 w_{\min}} \log^2(\frac{d}{\tau \lambda_{\min} w_{\min}}) \log^2(1/\delta)$, for some constant C , then, with probability at least $1 - \delta$, for any vector v in the subspace containing the centers, the projection of v onto the subspace returned by Step 2 of Algorithm 1 has length at least $\sqrt{1 - \tau^2} \|v\|$. Therefore, the directional maximum variance of any distribution D_i in the mixture along any direction in this subspace is at most $(1 - \tau^2)(\sigma^*)^2 + \tau^2 \sigma^2$, where σ^2 is the maximum directional variance of any distribution D_i in the mixture. When $\tau \leq \sigma^*/\sigma$, this variance is at most $2(\sigma^*)^2$. Since the directional variance of the entire mixture is 1 in any direction, $w_{\min} \sigma^2 \leq 1$ which means that $\sigma \leq \frac{1}{\sqrt{w_{\min}}}$. Therefore, when $n > C \cdot \frac{d}{(\sigma^*)^2 \lambda_{\min}^2 w_{\min}^2} \log^2(\frac{d}{\sigma^* \lambda_{\min} w_{\min}}) \log^2(1/\delta)$, for some constant C , the maximum directional variance of any distribution D_i in the mixture in the space output by Step 2 of the Algorithm is at most $2(\sigma^*)^2$.

Since A and B are random partitions of the sample set S , the subspace produced by the action of Step 2 of Algorithm 1 on the set A is independent of B , and vice versa. Therefore, when projected onto the top $k - 1$ singular value decomposition subspace of $\widehat{\Sigma}_{12}(A)$, the samples from B are distributed as a mixture of $(k - 1)$ -dimensional Gaussians. From the previous paragraph, in this subspace, the separation between the centers of any two distributions in the mixture is at least $c' \sigma^* \cdot k^{1/4} \sqrt{\log(\frac{kn}{\delta})}$, for some constant c' , and the maximum directional standard deviation of any distribution in the mixture is at most $2\sigma^*$. The theorem now follows from Theorem 1 of [AK05]. \square

Similar theorems with slightly worse bounds holds when the distributions in the mixture are not Gaussian, but possess certain distance concentration properties.

The following lemma is useful for the proof of Theorem 2.

Lemma 8 *Let D_i be a log-concave distribution over \mathbf{R}^d , and let S be a fixed r -dimensional subspace, such that the maximum directional variance of D_i along S is at most σ^* . If x is a point sampled from D_i , and if $P_S(x)$ denotes the projection of x onto S , then, with probability $1 - \delta$,*

$$\|P_S(x - \mu_i^1)\| \leq \sigma^* \sqrt{r} \log\left(\frac{r}{\delta}\right)$$

PROOF: From Lemma 2 in [KSV05], if v is a unit vector, and D_i is a log-concave distribution, then, for x sampled from D_i ,

$$\Pr[|v^\top(x - \mu_i^1)| > \sigma^* t] \leq e^{-t}$$

Let v_1, \dots, v_r be a basis of S . Plugging in $t = \log(\frac{r}{\delta})$, for all v_l in this basis, with probability $1 - \frac{\delta}{r}$, $|v_l^\top(x - \mu_i^1)| \leq \sigma^* \log(\frac{r}{\delta})$. The lemma now follows by observing that $\|x - \mu_i^1\|^2 = \sum_l |v_l^\top(x - \mu_i^1)|^2$. \square

Now the proof of Theorem 2 follows.

PROOF:(Of Theorem 2) We can use Lemma 8 along with an argument similar to the first part of the proof in Theorem 1 to show that if the number of samples

$$n > c \cdot \frac{d}{(\sigma^*)^2 \lambda_{\min}^2 w_{\min}^2} \log^3\left(\frac{d}{\sigma^* \lambda_{\min} w_{\min}}\right) \log^2(1/\delta)$$

then, one can find a subspace such that for any vector v in the subspace containing the centers, (1) the projection of v onto this subspace has norm at least $c' \cdot \|v\|$, where c' is a constant, and (2) the maximum directional standard deviation of any D_i in the mixture along v is at most $2\sigma^*$.

We now apply Lemma 8 on the subspace returned by our Algorithm in View 1. We remark that we can do this, because, due to the independence between the sample sets A and B , the top $k - 1$ singular value decomposition subspace of the covariance matrix of A across the views is independent of B and vice versa.

Applying this Lemma and a triangle inequality, if D_i is a log-concave distribution such that the maximum directional variance of D_i in S^1 is σ^* , then, the distance between all pairs of points from D_i when projected onto the $(k - 1)$ -dimensional subspace returned by Step 2 of Algorithm 1 is at most $O(\sigma^* \sqrt{k} \log(\frac{nk}{\delta}))$ with probability at least $1 - \delta$. This statement implies that single linkage will succeed (due to that the interclass distances are larger than the intraclass distances), which concludes the proof of the theorem. \square

Finally, we prove Theorem 3. The following lemma is useful for the proof of Theorem 3.

Lemma 9 *Let D_i be a product distribution over \mathbf{R}^d , in which each coordinate has range R , and let S be a fixed r -dimensional subspace. Let x be a point drawn from D_i , and let $P_S(x)$ be the projection of x onto S . Then, with probability at least $1 - \delta$,*

$$\|P_S(x - \mu_i^1)\| \leq R \sqrt{2r \log(\frac{2r}{\delta})}$$

PROOF: Let v_1, \dots, v_r be an orthonormal basis of S . For any vector v_l , and any x drawn from distribution D_i ,

$$\Pr[|v_l^\top(x - \mu_i^1)| > Rt] \leq 2e^{-t^2/2}$$

Plugging in $t = \sqrt{2 \log(\frac{2r}{\delta})}$, we get that with probability at least $1 - \frac{\delta}{r}$, $|v_l^\top(x - \mu_i^1)| \leq R \sqrt{2 \log(\frac{2r}{\delta})}$. The rest follows by the observation that in the subspace returned by our algorithm, $\|x - \mu_i^1\|^2 = \sum_{l=1}^r (v_l^\top(x - \mu_i^1))^2$, and a union bound over all r vectors in this basis. \square

PROOF: (Of Theorem 3) Since the distribution of each coordinate has range R , and the coordinates are distributed independently, for any distribution D_i in the mixture, the maximum directional variance is at most R . We can now use Lemma 9 and an argument similar to the first part of the proof in Theorem 1 to show that if the number of samples $n > c \cdot \frac{d}{\lambda_{\min}^2 w_{\min}} \log^2(\frac{d}{\sigma^* \lambda_{\min} w_{\min}}) \log^2(1/\delta)$, then, one can find a subspace such that for any vector v in the subspace containing the centers, the projection of v onto this subspace has norm at least $c' \cdot \|v\|$, where c' is a constant.

We now apply Lemma 9 on the subspace returned by our Algorithm in View 1. We remark that we can do this, because, due to the independence between the sample sets A and B , the top $k - 1$ singular value decomposition subspace of the covariance matrix of A across the views, is independent of B and vice versa. Applying Lemma 9 and a triangle inequality, if D_i is a product distribution where each coordinate has range R , then, the distance between all pairs of points from D_i when projected onto a specific $(k - 1)$ -dimensional subspace is at most $O(R \sqrt{k \log(\frac{nk}{\delta})})$ with probability at least $1 - \delta$. This statement implies that single linkage will succeed (due to that the interclass distances are larger than the intraclass distances), which concludes the proof of the theorem. \square

References

- [AK05] S. Arora and R. Kannan. Learning mixtures of separated nonspherical gaussians. *Annals of Applied Probability*, 15(1A):69–92, 2005.

- [AM05] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the 18th Annual Conference on Learning Theory*, pages 458–469, 2005.
- [AZ07] Rie Kubota Ando and Tong Zhang. Two-view feature generation model for semi-supervised learning. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 25–32, New York, NY, USA, 2007. ACM.
- [BM98] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann Publishers, pages 92–100, 1998.
- [BV08] S. C. Brubaker and S. Vempala. Isotropic pca and affine-invariant clustering. In *Proc. of Foundations of Computer Science*, 2008.
- [CR08] K. Chaudhuri and S. Rao. Learning mixtures of distributions using correlations and independence. In *In Proc. of Conference on Learning Theory*, 2008.
- [Das99] S. Dasgupta. Learning mixtures of gaussians. In *Proceedings of the 40th IEEE Symposium on Foundations of Computer Science*, pages 634–644, 1999.
- [DS00] S. Dasgupta and L. Schulman. A two-round variant of em for gaussian mixtures. In *Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2000.
- [KF07] Sham M. Kakade and Dean P. Foster. Multi-view regression via canonical correlation analysis. In Nader H. Bshouty and Claudio Gentile, editors, *COLT*, volume 4539 of *Lecture Notes in Computer Science*, pages 82–96. Springer, 2007.
- [KSV05] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *Proceedings of the 18th Annual Conference on Learning Theory*, 2005.
- [RV07] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, 2007.
- [VW02] V. Vempala and G. Wang. A spectral algorithm of learning mixtures of distributions. In *Proceedings of the 43rd IEEE Symposium on Foundations of Computer Science*, pages 113–123, 2002.