Multi-View Clustering via Canonical Correlation Analysis

Kamalika Chaudhuri KAMALIKA@SOE.UCSD.EDU
ITA, UC San Diego, 9500 Gilman Drive, La Jolla, CA

Sham M. Kakade SHAM@TTI-C.ORG
Karen Livescu KLIVESCU@TTI-C.ORG
Karthik Sridharan KARTHIK@TTI-C.ORG
Toyota Technological Institute at Chicago, 6045 S. Kenwood Ave., Chicago, IL

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. Here, we consider constructing such projections using multiple views of the data, via Canonical Correlation Analysis (CCA).

Under the assumption that the views are uncorrelated given the cluster label, we show that the separation conditions required for the algorithm to be successful are significantly weaker than prior results in the literature. We provide results for mixtures of Gaussians and mixtures of log concave distributions. We also provide empirical support from audio-visual speaker clustering (where we desire the clusters to correspond to speaker ID) and from hierarchical Wikipedia document clustering (where one view is the words in the document and the other is the link structure).

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 the goal is to use the relationship between these views to alleviate the difficulty of a learning problem of interest (Blum & Mitchell, 1998; Kakade & Foster, 2007; Ando & Zhang, 2007). In this work, we explore how having two 'views' makes the clustering problem significantly more tractable.

Much recent work has been done on understanding under what conditions we can learn a mixture model. The basic problem is as follows: We are given independent 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 in the mixture it was generated from.

Under no restrictions on the underlying mixture, this problem is considered to be hard. However, in many applications, we are only interested in clustering the data when the component distributions are "well separated". In fact, the focus of recent clustering algorithms (Dasgupta, 1999; Vempala & Wang, 2002; Achlioptas & McSherry, 2005; Brubaker & Vempala, 2008) is on efficiently learning with as little separation as possible. Typically, the separation conditions are such that when given a random sample from the mixture model, the Bayes optimal classifier is able to reliably recover which cluster generated that point.

This work makes a natural multi-view assumption: that the views are (conditionally) uncorrelated, conditioned on which mixture component generated the views. There are many natural applications for which this assumption applies. For example, we can consider multi-modal views, with one view being a video stream and the other an audio stream, of a speaker — here, conditioned on the speaker identity and maybe the phoneme (both of which could label the generating cluster), the views may be uncorrelated. A second example is the words and link structure in a document from a corpus such as Wikipedia – here, conditioned on the category of each document, the words in it and its link structure may be uncorrelated. In this paper, we provide experiments for both settings.

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

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 low-dimensional subspace spanned by the means of the component distributions. This subspace is important, because, when projected onto this subspace, the means of the distributions are well-separated, yet the typical distance between points from the same distribution is smaller than in the original space. The number of samples we require to cluster correctly scales as $O(d)$, where d is the ambient dimension. Finally, we show through experiments that CCA-based algorithms consistently provide better performance than standard PCA-based clustering methods when applied to datasets in the two quite different domains of audio-visual speaker clustering and hierarchical Wikipedia document clustering by category.

Our work adds to the growing body of results which show how the multi-view framework can alleviate the difficulty of learning problems.

Related Work. Most provably efficient clustering algorithms first project the data down to some low-dimensional space and then cluster the data in this lower dimensional space (an algorithm such as single linkage usually suffices here). Typically, these algorithms also work under a separation requirement, which is measured by the minimum distance between the means of any two mixture components.

One of the first provably efficient algorithms for learning mixture models is due to (Dasgupta, 1999), who learns a mixture of spherical Gaussians by *randomly* projecting the mixture onto a low-dimensional subspace. (Vempala & Wang, 2002) provide an algorithm with an improved separation requirement that learns a mixture of k spherical Gaussians, by projecting the mixture down to the k-dimensional subspace of highest variance. (Kannan et al., 2005; Achlioptas & McSherry, 2005) extend this result to mixtures of general Gaussians; however, they require a separation proportional to the maximum directional standard deviation of any mixture component. (Chaudhuri & Rao, 2008) use a canonical correlations-based algorithm to learn mixtures of axis-aligned Gaussians with a separation proportional to σ^*, the maximum directional standard deviation in the subspace containing the means of the distributions. Their algorithm requires a coordinate-independence property, and an additional "spreading" condition. None of these algorithms are affine invariant.

Finally, (Brubaker & Vempala, 2008) provide 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. However, their separation involves a large polynomial dependence on $\frac{1}{w_{\min}}$.

The two results most closely related to ours are the work of (Vempala & Wang, 2002) and (Chaudhuri & Rao, 2008). (Vempala & Wang, 2002) show that it is sufficient to find the subspace spanned by the means of the distributions in the mixture for effective clustering. Like our algorithm, (Chaudhuri & Rao, 2008) use a projection onto the top $k-1$ singular value decomposition subspace of the canonical correlations matrix. They also require a *spreading condition*, which is related to our requirement on the rank. We borrow techniques from both of these papers.

(Blaschko & Lampert, 2008) propose a similar algorithm for multi-view clustering, in which data is projected onto the top directions obtained by kernel CCA across the views. They show empirically that for clustering images using the associated text as a second view (where the target clustering is a human-defined category), CCA-based clustering methods out-perform PCA-based algorithms.

This Work. Our input is data on a fixed set of objects from two views, where View j is assumed to be generated by a mixture of k Gaussians $(D_1^j, \ldots, D_k^j)$, for $j = 1, 2$. To generate a sample, a source i is picked with probability w_i, and $x^{(1)}$ and $x^{(2)}$ in Views 1 and 2 are drawn from distributions D_i^1 and D_i^2. Following prior theoretical work, our goal is to show that our algorithm recovers the correct clustering, provided the input mixture obeys certain conditons.

We impose two requirements on these mixtures. First, we require that conditioned on the source, the two views are uncorrelated. Notice that this is a weaker restriction than the condition that given source i, the samples from D_i^1 and D_i^2 are drawn *independently*. Moreover, this condition allows the distributions in the mixture within each view to be completely general, so long as they are uncorrelated across views. Although we do not prove this, our algorithm seems robust to small deviations from this assumption.

Second, we require the rank of the CCA matrix across the views to be at least $k-1$, when each view is in isotropic position, and the $k-1$-th singular value of this matrix to be at least $\lambda_{\min}$. This condition ensures that there is sufficient correlation between the views. If the first two conditions hold, then we can recover the subspace containing the means in both views.

In addition, for mixtures 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, then our algorithm can also determine which component each sample came from. Here μ_i^1 is the mean of the i-th component in View 1 and σ^* is the maximum directional standard deviation in the subspace containing the means in View 1. Moreover, the number of samples required to learn this mixture grows (almost) linearly with d.

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, which can be considerably lower than the maximum directional variance over all directions. The only other algorithm which provides affine-invariant guarantees is due to (Brubaker & Vempala, 2008) — the implied separation in their work is rather large and grows with decreasing $w_{\min}$, the minimum mixing weight. To get our improved sample complexity bounds, we use a result due to (Rudelson & Vershynin, 2007) which may be of independent interest.

We stress that our improved results are really due to the multi-view condition. Had we simply combined the data from both views, and applied previous algorithms on the combined data, we could not have obtained our guarantees. 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 the multi-view and rank conditions are obeyed.

Finally, we study through experiments the performance of CCA-based algorithms on data sets from two different domains. First, we experiment with audio-visual speaker clustering, in which the two views are audio and face images of a speaker, and the target cluster variable is the speaker. Our experiments show that CCA-based algorithms perform better than PCA-based algorithms on audio data and just as well on image data, and are more robust to occlusions of the images. For our second experiment, we cluster documents in Wikipedia. The two views are the words and the link structure in a document, and the target cluster is the category. Our experiments show that a CCA-based hierarchical clustering algorithm out-performs PCA-based hierarchical clustering for this data.

2. The Setting

We assume that our data is generated by a mixture of k distributions. In particular, we assume that we obtain samples $x = (x^{(1)}, x^{(2)})$, where $x^{(1)}$ and $x^{(2)}$ are the two views, 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 $i = 1, \ldots, k$ and $j = 1, 2$, be the mean of distribution i in view j, and let w_i be the mixing weight for distribution i.

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

$$\Sigma = \mathbf{E}[xx^\top], \qquad \Sigma_{11} = \mathbf{E}[x^{(1)}(x^{(1)})^\top]$$
$$\Sigma_{22} = \mathbf{E}[x^{(2)}(x^{(2)})^\top], \qquad \Sigma_{12} = \mathbf{E}[x^{(1)}(x^{(2)})^\top]$$
$$\text{Hence, we have:} \qquad \Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{21} \\ \hline \Sigma_{12} & \Sigma_{22} \end{array}\right] \quad (1)$$

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

Assumption 1 *(Multi-View Condition) We assume that conditioned on the source distribution l in the mixture (where $l = i$ is picked with probability w_i), the two views are uncorrelated. More precisely, we assume that for all $i \in [k]$,*

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

This assumption implies that: $\Sigma_{12} = \sum_i w_i \mu_i^1 \cdot (\mu_i^2)^T$. To see this, observe that

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

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 an 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 preprocessing 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 $\lambda_{\min}$ is the minimal non-zero correlation coefficient, $\lambda_{\min} = \min_{u,v:\rho(u,v)\neq 0} \rho(u,v)$. Note $1 \geq \lambda_{\min} > 0$.

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

3. The Clustering Algorithm

The following lemma provides 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).*

The lemma is a consequence of Equation 2 and the rank assumption. Since samples from a mixture are well-separated in the space containing the means of the distributions, the lemma suggests the following strategy: use CCA to (approximately) project the data down to the subspace spanned by the means to get an easier clustering problem, and then apply standard clustering algorithms in this space.

Our clustering algorithm, based on the above idea, is stated below. We can show that this algorithm clusters correctly with high probability, when the data in at least one of the views obeys a separation condition, in addition to our assumptions.

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 A and B of equal size.
2. Let $\widehat{\Sigma}_{12}(A)$ ($\widehat{\Sigma}_{12}(B)$ resp.) denote the empirical covariance matrix between views 1 and 2, computed from the sample set A (B resp.). Compute the top $k-1$ left singular vectors of $\widehat{\Sigma}_{12}(A)$ ($\widehat{\Sigma}_{12}(B)$ resp.), and project the samples in B (A resp.) on the subspace spanned by these vectors.
3. Apply single linkage clustering (Dunn & Everitt, 2004) (for mixtures of log-concave distributions), or the algorithm in Section 3.5 of (Arora & Kannan, 2005) (for mixtures of Gaussians) on the projected examples in View 1.

We note that in Step 3, we apply either single linkage or the algorithm of (Arora & Kannan, 2005); this allows us to show theoretically that if the distributions in the mixture are of a certain type, and given the right separation conditions, the clusters can be recovered correctly. In practice, however, these algorithms do not perform as well due to lack of robustness, and one would use an algorithm such as k-means or EM to cluster in this low-dimensional subspace. In particular, a variant of the EM algorithm has been shown (Dasgupta & Schulman, 2000) to cluster correctly mixtures of Gaussians, under certain conditions.

Moreover, in Step 1, we divide the data set into two halves to ensure independence between Steps 2 and 3 for our analysis; in practice, however, these steps can be executed on the same sample set.

Main Results. Our main theorem is 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,*

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

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

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

for some constant c.

Here we assume that a separation condition holds in View 1, but a similar theorem also applies to View 2. An analogous theorem can also be shown for mixtures 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,

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

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

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

for some constant c.

The proof follows from the proof of Theorem 1, along with standard results on log-concave probability distributions – see (Kannan et al., 2005; Achlioptas & McSherry, 2005). We do not provide a proof here due to space constraints.

4. Analyzing Our Algorithm

In this section, we prove our main theorems.

Notation. In the sequel, we assume that we are given samples from a mixture which obeys Assumptions 2 and 1. We use the notation S^1 (resp. S^2) to denote the subspace containing the centers of the distributions in the mixture in View 1 (resp. View 2), and notation S'^1 (resp. S'^2) to denote the orthogonal complement to the subspace containing the centers of the distributions in the mixture in View 1 (resp. View 2).

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

Proofs. Now, we are ready to prove our main theorem. First, we show the following two lemmas, which demonstrate properties of the expected cross-correlational matrix across the views. Their proofs are immediate from Assumptions 2 and 1.

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'^1 (resp. S'^2). Then, for any $u^1 \in \mathcal{V}_1$ and $u_2 \in \mathcal{V}_2$, $(v^1)^T\Sigma_{12}u^2 = (u^1)^T\Sigma_{12}v^2 = 0$.*

Next, we show that given sufficiently many samples, the subspace spanned by the top $k-1$ singular vectors of $\widehat{\Sigma}_{12}$ still approximates the subspace containing the means of the distributions comprising the mixture. Finally, we use this fact, along with some results in (Arora & Kannan, 2005) to prove Theorem 1. Our main lemma of this section is the following.

Lemma 4 (Projection Subspace Lemma) *Let v^1 (resp. v^2) be any vector in S^1 (resp. S^2). If the number of samples $n > c\frac{d}{\tau^2\lambda_{\min}^2 w_{\min}}\log^2(\frac{d}{\tau\lambda_{\min}w_{\min}})\log^2(\frac{1}{\delta})$ 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. $\sqrt{1-\tau^2}||v^2||$).*

The main tool in the proof of Lemma 4 is the following lemma, which uses a result due to (Rudelson & Vershynin, 2007).

Lemma 5 (Sample Complexity Lemma) *If the number of samples*

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

for some constant c, then, with probability at least $1-\delta$, $||\widehat{\Sigma}_{12} - \Sigma_{12}|| \leq \epsilon$.

A consequence of Lemmas 5, 2 and 3 is the following.

Lemma 6 *Let $n > C\frac{d}{\epsilon^2 w_{\min}}\log^2(\frac{d}{\epsilon w_{\min}})\log^2(\frac{1}{\delta})$, 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.*

The proof follows by a combination of Lemmas 2,3, 5.

PROOF:(Of Lemma 5) To prove this lemma, we apply Lemma 7. 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 applied the triangle inequality to the 2×2 block matrix with off-diagonal entries $\widehat{\Sigma}_{12} - \Sigma_{12}$ and with 0 diagonal entries). We now apply Lemma 7 three times, on $\widehat{\Sigma}_{11} - \Sigma_{11}$, $\widehat{\Sigma}_{22} - \Sigma_{22}$, and a scaled version of $\widehat{\Sigma} - \Sigma$. The first two applications follow directly.

For the third application, we observe that Lemma 7 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 7 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; so the maximum singular value of Σ is at most 4. Our claim follows. The lemma follows by plugging in n as a function of ϵ, d and $w_{\min}$ $\square$

Lemma 7 *Let X be a set of n points generated by a mixture of k Gaussians over R^d, scaled such that $\boldsymbol{E}[x \cdot x^T] = I_d$. If M is the sample covariance matrix of X, then, for n large enough, with probability at least $1-\delta$,*

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

where C is a fixed 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 (Rudelson & Vershynin, 2007). 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, $w_{\min}(\mu^2 + \sigma^2) \leq 1$. Therefore, for all samples x_i, with probability at least $1 - \delta/2$, $||x_i|| \leq ||\mu|| + \sigma\sqrt{d \log(\frac{2n}{\delta})}$.

We condition on the fact that the event $||x_i|| \leq ||\mu|| + \sigma\sqrt{d \log(\frac{2n}{\delta})}$ happens for all $i = 1, \ldots, n$. 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 (Rudelson & Vershynin, 2007) 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 $||x_i||$. The lemma follows by plugging in $t = \sqrt{\frac{\Lambda^2 \log(4/\delta) \log n}{cn}}$, and $\Lambda \leq \frac{2\sqrt{d \log(2n/\delta)}}{\sqrt{w_{\min}}}$. □

PROOF: (Of Lemma 4) 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 $\sqrt{1-\tilde{\tau}^2}||v^1||$, 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 $\tilde{\tau}||v^1||$. This vector 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, for this vector u^2, $(u^1)^\top \Sigma_{12} u^2 \geq \tilde{\tau}\lambda_{\min}$. If $n > c\frac{d}{\tilde{\tau}^2 \lambda_{\min}^2 w_{\min}} \log^2(\frac{d}{\tilde{\tau}\lambda_{\min} w_{\min}}) \log^2(\frac{1}{\delta})$, then, from Lemma 6, $(u^1)^T \widehat{\Sigma}_{12} u^2 \geq \frac{\tilde{\tau}}{2}\lambda_{\min}$.

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 means that there exists a vector $z^2 \in \mathcal{V}_2$, the orthogonal complement of the subspace spanned by the top $k-1$ right singular vectors of $\widehat{\Sigma}_{12}$ such that $(u^1)^T \widehat{\Sigma}_{12} z^2 \geq \frac{\tilde{\tau}}{2}\lambda_{\min}$. This implies that the k-th singular value of $\widehat{\Sigma}_{12}$ is at least $\frac{\tilde{\tau}}{2}\lambda_{\min}$. However, from Lemma 6, all but the top $k-1$ singular values of $\widehat{\Sigma}_{12}$ are at most $\frac{\tau}{3}\lambda_{\min}$, which is a contradiction. □

PROOF:(Of Theorem 1) From Lemma 4, if $n > \frac{Cd}{\tau^2 \lambda_{\min}^2 w_{\min}} \log^2(\frac{d}{\tau\lambda_{\min} w_{\min}}) \log^2(\frac{1}{\delta})$, then, with probability at least $1-\delta$, the projection of any vector v in S^1 or S^2 onto the subspace returned by Step 2 of Algorithm 1 has length at least $\sqrt{1-\tau^2}||v||$. Therefore, the maximum directional variance of any D_i 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 D_i. When $\tau \leq \frac{\sigma^*}{\sigma}$, this is at most $2(\sigma^*)^2$. From the isotropic condition, $\sigma \leq \frac{1}{\sqrt{w_{\min}}}$. Therefore, when $n > \frac{Cd}{(\sigma^*)^2 \lambda_{\min}^2 w_{\min}^2} \log^2(\frac{d}{\sigma^* \lambda_{\min} w_{\min}}) \log^2(\frac{1}{\delta})$, the maximum directional variance of any D_i in the mixture in the space output by Step 2 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 $k-1$ SVD subspace of $\widehat{\Sigma}_{12}(A)$, the samples from B are distributed as a mixture of $(k-1)$-dimensional Gaussians. The theorem follows from the previous paragraph, and Theorem 1 of (Arora & Kannan, 2005). □

5. Experiments

5.1. Audio-visual speaker clustering

In the first set of experiments, we consider clustering either audio or face images of speakers. We use 41 speakers from the VidTIMIT database (Sanderson, 2008), speaking 10 sentences (about 20 seconds) each, recorded at 25 frames per second in a studio environment with no significant lighting or pose variation. The audio features are standard 12-dimensional mel cepstra (Davis & Mermelstein, 1980) and their derivatives and double derivatives computed every 10ms over a 20ms window, and finally concatenated over a window of 440ms centered on the current frame, for a total of 1584 dimensions. The video features are pixels of the face region extracted from each image (2394 dimensions). We consider the target cluster variable to be the speaker. We use either CCA or PCA to project the data to a lower dimensionality N. In the case of CCA, we initially project to an intermediate dimensionality M using PCA to reduce the effects of spurious correlations. For the results reported here, typical values (selected using a held-out set) are $N = 40$ and

	PCA	CCA
Images	1.1	1.4
Audio	35.3	12.5
Images + occlusion	6.1	1.4
Audio + occlusion	35.3	12.5
Images + translation	3.4	3.4
Audio + translation	35.3	13.4

Table 1. Conditional perplexities of the speaker given the cluster, using PCA or CCA bases. "+ occlusion" and "+ translation" indicate that the images are corrupted with occlusion/translation; the audio is unchanged, however.

$M = 100$ for images and 1000 for audio. For CCA, we randomize the vectors of one view in each sentence, to reduce correlations between the views due to other latent variables such as the current phoneme. We then cluster either view using k-means into 82 clusters (2 per speaker). To alleviate the problem of local minima found by k-means, each clustering consists of 5 runs of k-means, and the one with the lowest score is taken as the final clustering.

Similarly to (Blaschko & Lampert, 2008), we measure clustering performance using the conditional entropy of the speaker s given the cluster c, $H(s|c)$. We report the results in terms of conditional perplexity, $2^{H(s|c)}$, which is the mean number of speakers corresponding to each cluster. Table 1 shows results on the raw data, as well as with synthetic occlusions and translations of the image data. Considering the clean visual environment, we expect PCA to do very well on the image data. Indeed, PCA provides an almost perfect clustering of the raw images and CCA does not improve it. However, CCA far outperforms PCA when clustering the more challenging audio view. When synthetic occlusions or translations are applied to the images, the performance of PCA-based clustering is greatly degraded. CCA is unaffected in the case of occlusion; in the case of translation, CCA-based image clustering is degraded similarly to PCA, but audio clustering is almost unaffected. In other words, even when the image data are degraded, CCA is able to recover a good clustering in at least one of the views. [1] For a more detailed look at the clustering behavior, Figures 1(a-d) show the distributions of clusters for each speaker.

[1]The audio task is unusually challenging, as each feature vector corresponds to only a few phonemes. A typical speaker classification setting uses entire sentences. If we force the cluster identity to be constant over each sentence (the most frequent cluster label in the sentence), performance improves greatly; e.g., in the "audio+occlusion" case, the perplexity improves to 8.5 (PCA) and 2.1 (CCA).

5.2. Clustering Wikipedia articles

Next we consider the task of clustering Wikipedia articles, based on either their text or their incoming and outgoing links. The link structure L is represented as a concatenation of "to"and "from" link incidence vectors, where each element $L(i)$ is the number of times the current article links to/from article i. The article text is represented as a bag-of-words feature vector, i.e. the raw count of each word in the article. A lexicon of about 8 million words and a list of about 12 million articles were used to construct the two feature vectors. Since the dimensionality of the feature vectors is very high (over 20 million for the link view), we use random projection to reduce the dimensionality to a computationally manageable level.

We present clustering experiments on a subset of Wikipedia consisting of 128,327 articles. We use either PCA or CCA to reduce the feature vectors to the final dimensionality, followed by clustering. In these experiments, we use a hierarchical clustering procedure, as a flat clustering is poor with either PCA or CCA (CCA still usually outperforms PCA, however). In the hierarchical procedure, all points are initially considered to be in a single cluster. Next, we iteratively pick the largest cluster, reduce the dimensionality using PCA or CCA on the points in this cluster, and use k-means to break the cluster into smaller sub-clusters (for some fixed k), until we reach the total desired number of clusters. The intuition for this is that different clusters may have different natural subspaces.

As before, we evaluate the clustering using the conditional perplexity of the article category a (as given by Wikipedia) given the cluster c, $2^{H(a|c)}$. For each article we use the first category listed in the article. The 128,327 articles include roughly 15,000 categories, of which we use the 500 most frequent ones, which cover 73,145 articles. While the clustering is performed on all 128,327 articles, the reported entropies are for the 73,145 articles. Each sub-clustering consists of 10 runs of k-means, and the one with the lowest k-means score is taken as the final cluster assignment.

Figure 1(e) shows the conditional perplexity versus the number of clusters for PCA and CCA based hierarchical clustering. For any number of clusters, CCA produces better clusterings, i.e. ones with lower perplexity. In addition, the tree structures of the PCA/CCA-based clusterings are qualitatively different. With PCA based clustering, most points are assigned to a few large clusters, with the remaining clusters being very small. CCA-based hierarchical clustering produces more balanced clusters. To see this, in Figure 1(f) we show the perplexity of the cluster distribu-

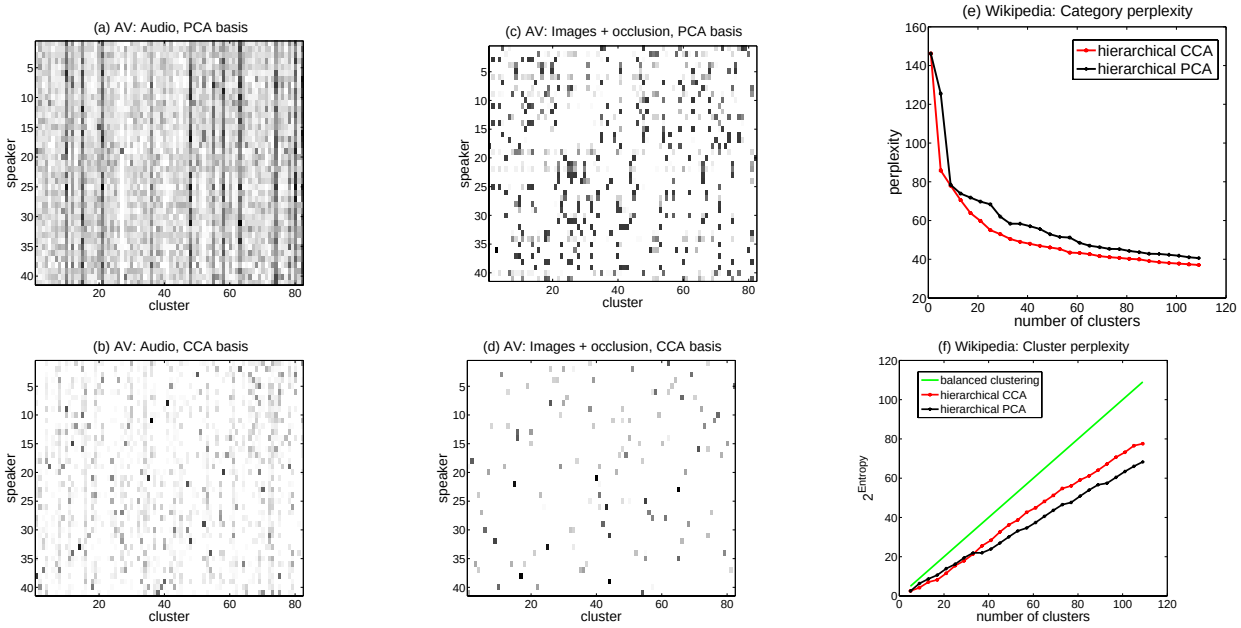

Figure 1. (a-d) Distributions of cluster assignments per speaker in audio-visual experiments. The color of each cell (s, c) corresponds to the empirical probability $p(c|s)$ (darker = higher). (e-f) Wikipedia experiments: (e) Conditional perplexity of article category given cluster ($2^{H(a|c)}$). (f) Perplexity of the cluster distribution ($2^{H(c)}$)

tion versus number of clusters. For about 25 or more clusters, the CCA-based clustering has higher perplexity, indicating a more uniform distribution of clusters.

References

Achlioptas, D., & McSherry, F. (2005). On spectral learning of mixtures of distributions. *Conf. on Learning Thy* (pp. 458–469).

Ando, R. K., & Zhang, T. (2007). Two-view feature generation model for semi-supervised learning. *Int. Conf. on Machine Learning* (pp. 25–32).

Arora, S., & Kannan, R. (2005). Learning mixtures of separated nonspherical Gaussians. *Ann. Applied Prob.*, *15*, 69–92.

Blaschko, M. B., & Lampert, C. H. (2008). Correlational spectral clustering. *Conf. on Comp. Vision and Pattern Recognition.*

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Conf. on Learning Thy.* (pp. 92–100).

Brubaker, S. C., & Vempala, S. (2008). Isotropic PCA and affine-invariant clustering. *Found. of Comp. Sci.* (pp. 551–560).

Chaudhuri, K., & Rao, S. (2008). Learning mixtures of distributions using correlations and independence. *Conf. On Learning Thy.* (pp. 9–20).

Dasgupta, S. (1999). Learning mixtures of Gaussians. *Found. of Comp. Sci.* (pp. 634–644).

Dasgupta, S., & Schulman, L. (2000). A two-round variant of EM for Gaussian mixtures. *Uncertainty in Art. Int.* (pp. 152–159).

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech, and Signal Proc.*, *28*, 357–366.

Dunn, G., & Everitt, B. (2004). *An introduction to math. taxonomy.* Dover Books.

Kakade, S. M., & Foster, D. P. (2007). Multi-view regression via canonical correlation analysis. *Conf. Learning Thy* (pp. 82–96).

Kannan, R., Salmasian, H., & Vempala, S. (2005). The spectral method for general mixture models. *Conf. on Learning Thy* (pp. 444–457).

Rudelson, M., & Vershynin, R. (2007). Sampling from large matrices: An approach through geometric functional analysis. *Jour. of ACM.*

Sanderson, C. (2008). *Biometric person recognition: Face, speech and fusion.* VDM-Verlag.

Vempala, V., & Wang, G. (2002). A spectral algorithm for learning mixtures of distributions. *Found. of Comp. Sci.* (pp. 113–123).