# On the Benefit of Spectral Projection for Document Clustering

Santosh Vempala*
Mathematics Department and CSAIL, MIT
vempala@math.mit.edu

Grant Wang*
CSAIL, MIT
gjw@theory.csail.mit.edu

**Abstract**

Spectral algorithms have been applied with much experimental success to the problem of document clustering. However, their success is not fully explained – why do spectral algorithms perform better than traditional clustering algorithms such as $k$-means? In this paper, we propose that the benefit of spectral projection is based on two simple properties that aid clustering and give experimental evidence of these properties. The first property is that the spectral projection is a good approximation of the original matrix. The second is that the true clustering becomes more apparent in the spectral projection. We also discuss a connection between the second property and recent theoretical work on learning mixtures of distributions.

## 1 Introduction

Document clustering is a fundamental problem in information retrieval; it has been used to organize results of a query [CKPT92] and produce summaries of documents [HKH$^+$01]. In this problem, an algorithm is given a set of documents. The algorithm's task is to partition the set of documents so that each subset comprises documents about the same topic. Spectral clustering algorithms have shown impressive experimental results in this domain [Bol98, ZDG$^+$01, DHZ$^+$01, Dhi01, ZK02, XLG03, LMO04, CKVW05][1]. A spectral clustering algorithm is one that partitions a data set using the eigenvalues and eigenvectors of a matrix representation of the data. Such a description is vague, but a more precise description cannot be given because spectral methods have been applied to clustering in many different ways. The varied uses of spectral methods and their common experimental success motivate the following question: do the proposed spectral algorithms share a common benefit for document clustering?

In this paper, we focus on spectral projection and give experimental evidence that the answer is yes; clustering algorithms that use the spectral projection of a document-term matrix do indeed share a common benefit in the form of two properties: *approximation* and *distinguishability*. The second property explains why traditional clustering algorithms such as $k$-means perform better on the spectral projection of a document-term matrix than on the document-term matrix itself. The purpose of this work is not to show that spectral methods *do* perform better than or can improve traditional clustering methods — this is well known. The goal of this paper is to explain the benefit of spectral projection by means of two specific properties.

The first property, approximation, means that the spectral projection of a document-term matrix remains close to the original document-term matrix. Thus, spectral projection reduces the dimension of the data, which speeds up clustering algorithms running on the data, while not incurring too much error. The experiments and results are described in Section 3.

The second property, distinguishability, is that clusters are more clearly demarcated after spectral projection. In particular, for natural (and commonly used) definitions of distance and similarity for documents, we give experimental evidence that inter-cluster distance/similarity is substantially more distinguishable from intra-cluster distance/similarity after spectral projection. Before spectral projection, these two quantities are indistinguishable. This explains why clustering algorithms that work solely on the basis of pairwise distances/similarities perform better on the projection compared to the original document-term matrix. This property of spectral projection coincides with recent theoretical work on spectral algorithms for learning mixtures of distributions [VW04, KSV04]. We explain this connection and the experimental results in Section 4.

### 1.1 Previous work

An early use of spectral techniques in information retrieval was the work of [DDL$^+$90], which showed that a low rank approximation to a document-term matrix improved precision and recall for queries. Much work on spectral techniques

*Both authors are supported in part by NSF Award CCR-0312339.

[1] For instance, see http://eigencluster.csail.mit.edu

for clustering followed in the text domain, including [Bol98, SS97, ZDG+01, DHZ+01, Dhi01, ZK02, XLG03, LMO04, CKVW05]. Empirical success of spectral methods has also appeared in other domains such as image segmentation [SM00, NJW01]. Theoretical work explaining the success of spectral methods in clustering has focused on proposing generative models for data and showing that particular spectral algorithms perform well [PRTV00, AFK+01, KVV04, McS01, VW04, KSV04]; An exception is the work of [KVV04] that gives a worst-case approximation guarantee for recursive spectral partitioning. The work done in [DM01] and [SS97] is most similar to the work in this paper. In [DM01], Dhillon and Modha perform similar experiments investigating the approximation and distinguishability properties of *concept decomposition*, a technique distinct from SVD that represents a document-term matrix in a lower-dimensional space. In [SS97], Schutze and Silverstein empirically study the effect of projections on clustering speed and efficacy. They conclude that spectral projection significantly speeds up clustering but has no effect on efficacy. However, their measure of efficacy is respect to a fixed clustering algorithm. Although a fixed clustering algorithm may not see improved performance after spectral projection, other clustering algorithms may benefit from spectral projection. Our work suggests that spectral projection has a positive effect on efficacy; we measure this effect independent of any particular clustering algorithm.

## 2 Preliminaries and Notation

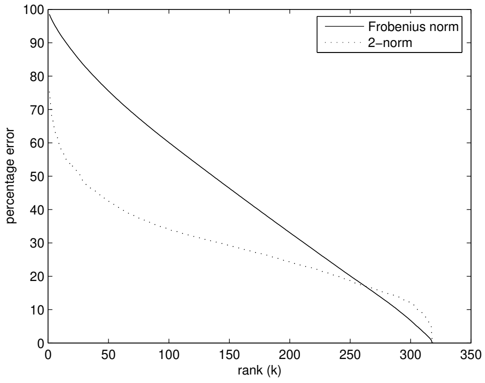
Let $A \in \mathbb{R}^{m \times n}$ be a document-term matrix; the rows of $A$ are the documents and the columns of $A$ are the terms. Let $A = U\Sigma V^T$ be the singular value decomposition of $A$, where $\Sigma$ is a diagonal matrix and $U$ and $V$ are orthonormal matrices. The diagonal entries of $\Sigma$ are the singular values of $A$, and the columns of $U$ and $V$ are the left and right singular vectors, respectively. Let $\Sigma_k$ be a diagonal matrix with the first $k$ singular values down the diagonal, and let $A_k = U\Sigma_k V^T$. We also refer to $A_k$ as the spectral projection of $A$ of rank $k$, since it is the projection of $A$ onto the top $k$ right singular vectors. $A_k$ is the best rank $k$ approximation to $A$ in the following sense:

THEOREM 2.1. (ECKHART AND YOUNG [EY39])

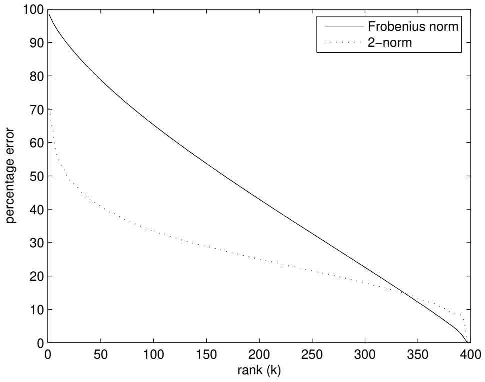
$$A_k = \mathit{argmin}_{B:rank(B) \le k} ||A - B||_F$$

$$A_k = \mathit{argmin}_{B:rank(B) \le k} ||A - B||_2$$

Recall that the 2-norm of a matrix is $\max_{v:||v||=1} ||Av||_2$ and the Frobenius norm is the sum of the squares of the matrix entries. More about the SVD and its properties can be found e.g., in [GL96].




(a) `alt.atheism`




(b) `rec.sport.hockey`

Figure 1: Percentage error vs. rank

## 3 Approximation

Reducing the dimensionality of data can speed up clustering algorithms computing over the data. However, a low-dimensional representation of high-dimensional data typically incurs some error. Thus, for dimension reduction to be effective, it is desirable that the error is small. We conducted an experiment to measure the error incurred by spectral projection of document-term matrices. From Theorem 2.1, we know that the spectral projection $A_k$ is the *best* approximation to $A$ for the Frobenius norm and 2-norm. The results show that this best approximation is indeed a good one — spectral projection to low dimensions introduces manageable error, even when the initial dimension is very high.

The data set we used was the 20 Newsgroups data set [Lan], which consists of 20,000 articles from 20

| data set | rank (k) 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 |
|---:|---|---|---|---|---|---|---|---|
| alt.atheism | 98.5% | 97.7% | 96.2% | 93.7% | 89.5% | 82.3% | 70.9% | 52.4% |
| comp.graphics | 98.7% | 98.1% | 97.1% | 95.5% | 92.5% | 87.4% | 78.7% | 63.5% |
| comp.os.ms-windows.misc | 98.7% | 98.0% | 96.8% | 94.7% | 91.1% | 85.5% | 76.2% | 60.5% |
| comp.sys.ibm.pc.hardware | 98.5% | 97.6% | 96.4% | 94.4% | 91.0% | 85.5% | 76.2% | 60.4% |
| comp.sys.mac.hardware | 98.8% | 98.1% | 96.8% | 94.7% | 91.2% | 85.2% | 75.1% | 58.5% |
| comp.windows.x | 98.9% | 98.3% | 97.2% | 95.5% | 92.5% | 87.2% | 78.2% | 62.9% |
| misc.forsale | 98.9% | 98.2% | 97.1% | 95.3% | 92.2% | 87.0% | 78.0% | 62.3% |
| rec.autos | 98.8% | 98.0% | 96.6% | 94.5% | 91.0% | 85.3% | 75.5% | 59.6% |
| rec.motorcycles | 98.9% | 98.2% | 97.0% | 94.8% | 91.2% | 85.0% | 74.6% | 58.1% |
| rec.sport.baseball | 98.7% | 97.9% | 96.5% | 94.3% | 90.8% | 84.6% | 74.4% | 58.1% |
| rec.sport.hockey | 98.6% | 97.9% | 96.5% | 94.2% | 90.6% | 84.6% | 74.7% | 58.7% |
| sci.crypt | 98.6% | 97.9% | 96.8% | 94.6% | 91.1% | 85.2% | 75.1% | 58.8% |
| sci.electronics | 98.9% | 98.1% | 96.7% | 94.6% | 91.1% | 85.3% | 75.2% | 58.5% |
| sci.med | 99.0% | 98.2% | 97.0% | 94.8% | 91.4% | 85.7% | 76.0% | 59.9% |
| sci.space | 98.7% | 98.0% | 96.7% | 94.6% | 91.0% | 84.9% | 75.2% | 59.3% |
| soc.religion.christian | 98.5% | 97.8% | 96.6% | 94.6% | 91.3% | 85.9% | 77.2% | 62.8% |
| talk.politics.guns | 98.5% | 97.7% | 96.4% | 94.3% | 90.7% | 84.7% | 74.5% | 58.0% |
| talk.politics.mideast | 98.4% | 97.3% | 95.7% | 93.0% | 88.9% | 82.5% | 72.4% | 56.2% |
| talk.politics.misc | 98.3% | 97.1% | 95.5% | 93.0% | 89.0% | 82.2% | 70.8% | 51.9% |
| talk.religion.misc | 98.5% | 97.5% | 95.8% | 92.9% | 88.3% | 80.3% | 66.6% | 43.5% |

Table 1: Percentage error (Frobenius norm) vs. rank for all 20 newsgroups

Usenet newsgroups. Each newsgroup contains roughly 1,000 articles divided into a training subset and a test subset. For each of the newsgroups, we constructed the document-term matrix for all the articles in the test subset of the newsgroup. We used the following common pre-processing steps.

- Stemming (using Porter's algorithm [Por80]).
- Removal of terms that appear too infrequently (less than two times).
- TF*IDF normalization.
- Normalization of term vectors to have Euclidean length one.

The document-term matrix was roughly 400 by $4,000$ for each newsgroup and contained roughly $24,000$ non-zero entries. The singular value decomposition was computed for each document-term matrix, and the percentage error

$$\frac{||A - A_k||}{||A||}$$

was measured for values of the rank ($k$) from 1 to 250 for both the Frobenius norm and the 2-norm.

The plots of percentage error vs. $k$ for the newsgroups `alt.atheism` and `rec.sport.hockey` appear in Figures 1(a) and 1(b). The percentage error for the Frobenius norm for all 20 newsgroups is shown in Table 1 for $k$ (rank) in powers of two. Table 2 contains the same information but for the 2-norm.

In Figures 1(a) and 1(b), the percentage error drops to 0% when $k = 319$ and $k = 399$, respectively. This is because `alt.atheism` consists of 319 documents and `rec.sport.hockey` consists of 399 documents. Thus, spectral projection to 319 and 399 dimensions results in perfect reconstructions of the document-term matrices for `alt.atheism` and `rec.sport.hockey`, respectively. Both plots show that the error for the Frobenius norm drops off roughly linearly, whereas the error for the 2-norm drops off more quickly initially. For high dimensions, though, the percentage error for the Frobenius norm is lower than that of the 2-norm.

The best low-rank approximations with respect to the Frobenius norm and 2-norm are indeed good approximations. When the dimensionality of the data was reduced from roughly 400 to 128, the percentage error for the Frobenius norm was less than 60% for all but two newsgroups (see Table 1). For the 2-norm (Table 2), we have even better results: when the dimensionality was reduced to just 32, the percentage error for all but three newsgroups was under 50%. It is interesting to note that all newsgroups behaved

| data set | rank (k) 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|---|---|---|
| alt.atheism | 75.4% | 71.6% | 66.7% | 61.2% | 54.9% | 47.3% | 39.5% | 31.3% |
| comp.graphics | 69.6% | 63.7% | 57.6% | 54.0% | 50.9% | 45.4% | 40.2% | 33.5% |
| comp.os.ms-windows.misc | 74.1% | 71.2% | 65.0% | 59.1% | 53.0% | 45.7% | 39.8% | 33.0% |
| comp.sys.ibm.pc.hardware | 72.3% | 64.8% | 58.2% | 52.9% | 47.8% | 42.1% | 36.6% | 30.3% |
| comp.sys.mac.hardware | 74.4% | 70.6% | 67.9% | 60.9% | 55.1% | 49.0% | 42.2% | 33.8% |
| comp.windows.x | 73.2% | 70.9% | 64.8% | 59.8% | 55.0% | 49.5% | 43.5% | 36.2% |
| misc.forsale | 78.7% | 74.2% | 64.7% | 61.1% | 55.1% | 49.2% | 43.7% | 36.5% |
| rec.autos | 82.4% | 77.2% | 69.7% | 63.9% | 55.4% | 49.3% | 42.8% | 34.5% |
| rec.motorcycles | 76.8% | 72.2% | 69.4% | 64.9% | 57.8% | 52.2% | 43.8% | 34.5% |
| rec.sport.baseball | 74.7% | 71.7% | 67.7% | 58.2% | 52.3% | 47.4% | 40.0% | 31.6% |
| rec.sport.hockey | 70.3% | 69.2% | 64.7% | 57.1% | 51.4% | 45.1% | 38.4% | 30.8% |
| sci.crypt | 66.1% | 63.2% | 61.5% | 57.1% | 50.6% | 45.1% | 39.1% | 30.8% |
| sci.electronics | 87.1% | 81.2% | 71.1% | 63.7% | 58.6% | 51.7% | 45.4% | 36.1% |
| sci.med | 83.7% | 80.2% | 71.7% | 64.5% | 58.1% | 51.7% | 45.4% | 36.3% |
| sci.space | 74.0% | 73.5% | 65.5% | 60.3% | 55.4% | 47.8% | 40.8% | 33.0% |
| soc.religion.christian | 66.6% | 62.2% | 56.6% | 53.2% | 47.4% | 41.3% | 35.4% | 29.4% |
| talk.politics.guns | 71.5% | 65.4% | 59.1% | 54.1% | 49.5% | 43.9% | 37.7% | 30.3% |
| talk.politics.mideast | 78.8% | 74.9% | 65.1% | 56.2% | 49.6% | 42.5% | 35.3% | 28.5% |
| talk.politics.misc | 85.3% | 69.5% | 60.9% | 55.7% | 50.2% | 43.7% | 37.3% | 29.7% |
| talk.religion.misc | 77.7% | 76.8% | 69.4% | 62.5% | 55.2% | 49.6% | 40.9% | 31.6% |

Table 2: Percentage error (2-norm) vs. rank for all 20 newsgroups

similarly with respect to percentage error versus rank, despite the difference in content in the newsgroups. For most values of $k$ (rank), the maximum percentage error difference between any two newsgroups was roughly 10%.

## 4 Distinguishability

Many document clustering algorithms are based on a pairwise distance or similarity function. The distance function maps two documents to any non-negative number; the larger the distance, the more unlike the two documents are. On the other hand, a similarity function maps two documents to the interval $[0, 1]$; the closer to 1, the more *like* the two documents are. Many different distance and similarity functions have been used when documents are represented as term vectors. We consider two commonly used candidates for a distance function and similarity function. The distance between two term vectors $u, v$ will be their Euclidean distance:

$$d(u, v) = ||u - v||.$$

For similarity, we use the Gaussian kernel:

$$s(u, v) = e^{-||u-v||^2}.$$

We give experimental evidence that under these measures, intra-cluster distance/similarity becomes substantially more distinguishable from inter-cluster distance/similarity after spectral projection.

**4.1 Experiment and results** The data set we used for this experiment was also the 20 Newsgroup data set [Lan]. The general experimental setup was as follows: we formed the document-term matrix $A$ of a random sample of the articles from $k$ random newsgroups, where $k$ ranged from 2 to 10. We measured the inter/intra newsgroup distance/similarity of these documents. We then computed $A_k$, the spectral projection of $A$ to $k$ dimensions, and measured the inter/intra newsgroup distance/similarity in the new representation. The exact experimental details follow.

We chose 10 random sets of $k$ different newsgroups. For a random set of $k$ newsgroups, we chose 10 random samples $S_1 \ldots, S_{10}$. Each random sample consisted of 50 articles from each of the $k$ newsgroups. Clearly, the articles in each random sample, $S_1 \ldots S_{10}$ can be partitioned into $k$ clusters according to the newsgroup to which they belong (we will refer to the collection of documents from the same newsgroup as a cluster). We then computed the document-term matrix $A$ for each random sample $S_i$, using the pre-processing steps described in Section 3. The singular value decomposition of $A$ was computed and we formed the spectral projection $A_k$.

Each row vector in $A_k$ was then normalized so that its Euclidean length was 1 (note that in $A$, each row vector was also normalized to 1).

For the distance function $d(u, v) = ||u - v||$, we measured the following two quantities both before and after projection:

- Average distance between cluster means.
- Average distance between cluster mean and a vector from the same cluster.

The cluster mean is simply the vector that is the average of all the term vectors from a newsgroup. Table 3 shows the results averaged together for each value of $k$. The key property is that the ratio of average distance between cluster means, $d(\mu_i, \mu_j)$, to average distance between a cluster mean and a term vector from the same cluster, $d(u, \mu)$, increases to the point that the two quantities become distinguishable. The factor by which this ratio increases after projection is the magnification factor (the last column in the table). Consider a situation in which the ratio $d(\mu_i, \mu_j)/d(u, \mu)$ is at least 2 for every pair of clusters; then we can put a ball around each cluster mean so that 1) the balls do not intersect each other, and 2) each ball contains only the term vectors from a single cluster. In our experiments, the ratio was at least 2 only for $k < 6$. Nevertheless, the magnification factor shows that it is easier to cluster the points in balls after spectral projection (most points can be clustered).

For the similarity function $s(u, v) = e^{-||u-v||^2}$, we measured the following quantities both before and after spectral projection:

- Average intra-cluster similarity, $s(C_i, C_i)$.
- Average inter-cluster similarity, $s(C_i, C_j)$.

The results appear in Table 4. The ratio between the average intra-cluster similarity and average inter-cluster similarity ($s(C_i, C_i)/s(C_i, C_j)$) is the key quantity. Before projection (4th column), the ratio is roughly 1. After projection (7th column), the ratio approaches 2 – meaning that, on average, documents from the same cluster are twice as similar as documents from different clusters. The magnification factor, the last column in the table, shows the increase in this ratio after projection. It is interesting that the magnification factor stays roughly constant as $k$ increases. This is not true for distance; the magnification factor drops as $k$ increases (see Table 3).

**4.2 Connection to learning mixtures of distributions** Measuring the effect of spectral projection on inter-cluster distance and intra-cluster distance was motivated by recent theoretical work on the problem of learning mixtures of distributions. In this problem, we are given a random sample from a mixture of $k$ distributions $F_1, \ldots, F_k$ with mixing weights $w_1, \ldots, w_k$ that sum to 1, i.e. $\sum_i w_i = 1$. A random sample is generated from the mixture by first choosing a distribution $F_i$ according to its mixing weight $w_i$, and then choosing a random sample according to $F_i$. We say that an algorithm *learns a mixture of distributions* if it can classify each random sample according to the distribution from which it was sampled.

Recent work has shown that spectral algorithms can provably learn mixtures of distributions for some important special cases. In [VW04], a spectral algorithm is given that correctly classifies a random sample from a mixture of spherical Gaussians, assuming a weak separation between the means of the Gaussians. This was generalized in [KSV04] using a stronger separation condition to mixtures of logconcave distributions, a class of distributions which includes general Gaussians. The key insight to both algorithms was relating the spectral projection subspace $V$, spanned by the top $k$ right singular vectors of the sample matrix $A$ (the rows of $A$ are simply the sample points), to the subspace $W$ spanned by the mean vectors of the distributions $F_1, \ldots, F_k$. In [VW04], it is shown that, in expectation, $V$ is the same subspace as $W$, but only when the distributions are spherical. This is not true when the distributions are arbitrary, but an approximate theorem holds.

Theorem 4.1. *([KSV04]) Let $S = S_1 \cup S_2 \ldots \cup S_k$ be a sample from a mixture with $k$ distributions such that $S_i$ is from the $i$th distribution $F_i$ and let $V$ be the $k$-dimensional SVD subspace of $S$. For each $i$, let $\mu_i$ be the mean of $S_i$ and $\sigma_i^2$ be the maximum variance of $S_i$ along any direction in $V$. Then,*

$$\sum_{i=1}^{k} |S_i| d(\mu_i, V)^2 \leq k \sum_{i=1}^{k} |S_i| \sigma_i^2$$

*where $d(\mu_i, V)$ is the orthogonal distance between $\mu_i$ and $V$.*

The theorem upper bounds the average distance between means and their projection to $V$. Thus a lower bound on the distance between any two means (i.e., $d(\mu_i, \mu_j)$), gives a lower bound on the distance between the means of any two projected distributions by the triangle inequality. Roughly speaking, the theorem says that, on average, the inter-mean distances do not decrease much.

On the other hand, it is reasonable to expect that the distance between a random point and the mean

| $k$ | before projection $d(\mu_i,\mu_j)$ | $d(u,\mu)$ | $\frac{d(\mu_i,\mu_j)}{d(u,\mu)}$ | after projection $d(\mu_i,\mu_j)$ | $d(u,\mu)$ | $\frac{d(\mu_i,\mu_j)}{d(u,\mu)}$ | magnification factor |
|---|---|---|---|---|---|---|---|
| 2 | 0.2902 | 0.9690 | 0.2995 | 0.8985 | 0.2197 | 4.0889 | 13.6524 |
| 3 | 0.2858 | 0.9701 | 0.2945 | 0.8222 | 0.3269 | 2.5151 | 8.5402 |
| 4 | 0.2842 | 0.9714 | 0.2925 | 0.8259 | 0.3860 | 2.1396 | 7.3149 |
| 5 | 0.2891 | 0.9703 | 0.2980 | 0.9124 | 0.4122 | 2.2137 | 7.4285 |
| 6 | 0.2876 | 0.9707 | 0.2963 | 0.8325 | 0.4439 | 1.8755 | 6.3297 |
| 7 | 0.2876 | 0.9717 | 0.2960 | 0.7731 | 0.4627 | 1.6709 | 5.6449 |
| 8 | 0.2863 | 0.9719 | 0.2946 | 0.8376 | 0.4953 | 1.6909 | 5.7396 |
| 9 | 0.2863 | 0.9714 | 0.2947 | 0.8716 | 0.5202 | 1.6756 | 5.6858 |
| 10 | 0.2847 | 0.9725 | 0.2928 | 0.8348 | 0.5374 | 1.5534 | 5.3053 |

Table 3: Average distances before and after spectral projection to $k$ dimensions

| $k$ | before projection $s(C_i,C_i)$ | $s(C_i,C_j)$ | $\frac{s(C_i,C_i)}{s(C_i,C_j)}$ | after projection $s(C_i,C_i)$ | $s(C_i,C_j)$ | $\frac{s(C_i,C_i)}{s(C_i,C_j)}$ | magnification factor |
|---|---|---|---|---|---|---|---|
| 2 | 0.1485 | 0.1407 | 1.0556 | 0.8709 | 0.4460 | 1.9530 | 1.8501 |
| 3 | 0.1479 | 0.1404 | 1.0534 | 0.7836 | 0.4450 | 1.7609 | 1.6716 |
| 4 | 0.1471 | 0.1398 | 1.0524 | 0.7282 | 0.4025 | 1.8093 | 1.7192 |
| 5 | 0.1478 | 0.1400 | 1.0553 | 0.7007 | 0.3643 | 1.9235 | 1.8227 |
| 6 | 0.1475 | 0.1399 | 1.0543 | 0.6686 | 0.3444 | 1.9410 | 1.8410 |
| 7 | 0.1470 | 0.1393 | 1.0546 | 0.6513 | 0.3334 | 1.9533 | 1.8522 |
| 8 | 0.1468 | 0.1394 | 1.0532 | 0.6144 | 0.3099 | 1.9825 | 1.8824 |
| 9 | 0.1471 | 0.1396 | 1.0535 | 0.5889 | 0.3061 | 1.9271 | 1.8292 |
| 10 | 0.1464 | 0.1391 | 1.0525 | 0.5695 | 0.2948 | 1.9359 | 1.8393 |

Table 4: Average intra-cluster and inter-cluster similarity before and after spectral projection to $k$ dimensions

of its distribution shrinks when projected to a lower dimension. As the ratio of the inter-mean distance to the distance between a random sample and the mean of its distribution increases, clustering based on pairwise distances becomes more effective. The assumption of Gaussian (or logconcave) distributions with separated means is sufficient to be able to classify the entire sample with high probability.

Table 3 suggests that this is indeed occurring for document clusters. Note that the average distance between means before and after spectral projection is roughly the same. This corresponds to the lower bound on the distance between means from Theorem 4.1. Meanwhile, the distance between a document and the mean of its cluster drops considerably. It is interesting that text corpora, which we have no reason to believe are mixtures of logconcave distributions benefit from spectral projection.

## 5 Conclusion

We have described two properties, approximation and distinguishability, that aid clustering and have given experimental evidence that the spectral projection of a document-term matrix has these properties. Besides more extensive experiments, several other questions also arise from this work: To what extent do the preprocessing steps aid the inter-cluster/intra-cluster distinguishability? What other definitions of distance and similarity can spectral projection magnify? The connection to the problem of learning mixtures of distribution also suggests a data-driven method of designing algorithms for document clustering. In this method, properties of real-world data are first experimentally verified and then an algorithm that provably works on data with such properties is designed. With this method, one might avoid having to use generative models (which may not be accurate) to show that an algorithm works on real-world data.

## References

[AFK+01] Y. Azar, A. Fiat, A. Karlin, F. McSherry, and J. Saia. Spectral analysis of data. In *Proceedings of* *the Thirty-Third Annual ACM Symposium on Theory of Computing*, pages 619–626, 2001.

[Bol98] D. Boley. Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, 2(4):325–344, 1998.

[CKPT92] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, 1992.

[CKVW05] D. Cheng, R. Kannan, S. Vempala, and G. Wang. A divide-and-merge methodology for clustering. In *Proceedings of the 24th Annual ACM Symposium on Principle of Database Systems*, page To Appear, 2005.

[DDL+90] S.C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[Dhi01] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–274, 2001.

[DHZ+01] C. Ding, X. He, H. Zha, M. Gu, and H. Simon. Spectral min-max cut for graph partitioning and data clustering. In *Proceedings of the First IEEE International Conference on Data Mining*, pages 107–114, 2001.

[DM01] I.S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1):143–175, 2001.

[EY39] C. Eckhart and G. Young. A principal axis transformation for non-hermitian matrices. *Bulletin of the American Mathematical Society*, 45:118–121, 1939.

[GL96] G. Golub and C. Van Loan. *Matrix Computations*. The Johns Hopkins Press, third edition, 1996.

[HKH+01] V. Hatzivassiloglou, J. Klavans, M. Holcombe, R. Barzilay, M. Kan, and K. McKeown. Simfinder: A flexible clustering tool for summarization. In *NAACL Workshop on Automatic Summarization*, pages 41–49, 2001.

[KSV04] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for mixture models. In *Electronic Colloquium on Computational Complexity*, pages Report TR04–067, 2004.

[KVV04] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad, and spectral. *Journal of the ACM*, 51(3):497–515, 2004.

[Lan] K. Lang. 20 newsgroups data set. `http://www.ai.mit.edu/people/jrennie/20Newsgroups/`.

[LMO04] T. Li, S. Ma, and M. Ogihara. Document clustering via adaptive subspace iteration. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 218–225, 2004.

[McS01] F. McSherry. Spectral partitioning of random graphs. In *Proceedings of the 42nd IEEE Annual Symposium on Foundations of Computer Science*, pages 529–537, 2001.

[NJW01] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Neural Information Processing Systems*, pages 849–856, 2001.

[Por80] M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[PRTV00] C. Papdimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: A probabalistic analysis. *Journal of Computer and System Sciences*, 61:217–235, 2000.

[SM00] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[SS97] H. Schutze and C. Silverstein. Projections for efficient document clustering. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81, 1997.

[VW04] S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[XLG03] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 267–273, 2003.

[ZDG+01] H. Zha, C. Ding, M. Gu, X. He, and H. Simon. Spectral relaxation for k-means clustering. In *Neural Information Processing Systems*, pages 1057–1064, 2001.

[ZK02] Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pages 515–524, 2002.