Unsupervised Learning: Self-aggregation in Scaled Principal Component Space*

Chris Ding[a], Xiaofeng He[a], Hongyuan Zha[b], Horst Simon[a]

[a]NERSC Division, Lawrence Berkeley National Laboratory
University of California, Berkeley, CA 94720

[b]Department of Computer Science and Engineering
Pennsylvania State University, University Park, PA 16802

{chqding,xhe,hdsimon}@lbl.gov, zha@cse.psu.edu

Abstract

We demonstrate that data clustering amounts to a dynamic process of *self-aggregation* in which data objects move towards each other to form clusters, revealing the inherent pattern of similarity. Self-aggregation is governed by connectivity and occurs in a space obtained by a nonlinear scaling of principal component analysis (PCA). The method combines dimensionality reduction with clustering into a single framework. It can apply to both square similarity matrices and rectangular association matrices.

1 Introduction

Organizing observed data into groups or clusters is the first step in discovering coherent patterns and useful structures. This unsupervised learning process (data clustering) is frequently encountered in science, engineering, commercial data mining and information processing. There exists a large number of data clustering methods for different situations. In recent decades, unsupervised learning methods related to the principal component analysis (PCA)[13] has being increasingly widely used: the low-dimensional space spanned by the principal components is effective in revealing structures of the observed high-dimensional data. PCA is a coordinate rotation such that the principal components span the dimensions of largest variance. The *linear* transformation preserves the local properties and global topologies, and can be efficiently computed. However, PCA is not effective in revealing nonlinear structures [9, 15, 21, 18, 19]. To overcome the short-comings of *linear* transformation of PCA, nonlinear PCAs have been proposed, such as principal curves [8], auto-associative networks [15], and kernel PCA [19]. But they do not posses the self-aggregation property. Recently, nonlinear mappings [21, 18] have been developed. But they are not primarily concerned with data clustering.

*LBNL Tech Report 49048, October 5, 2001. Supported by Department of Energy (Office of Science, through a LBNL LDRD) under contract DE-AC03-76SF00098.

Here we introduce a new concept of *self-aggregation* and show that a nonlinear scaling of PCA leads to a low-dimensional space in which data objects self-aggregate into distinct clusters, revealing inherent patterns of similarity, in contrast to existing approaches. Thus data clustering becomes a *dynamic* process, performing nonlinear dimensionality reduction and cluster formation simultaneously; the process is governed by the connectivity among data objects, similar to dynamic processes in recurrent networks [12, 10].

2 Scaled Principal Components

Associations among data objects are mostly quantified by a similarity metric. The scaled principal component approach starts with a nonlinear (non-uniform) scaling of the similarity matrix $W = (w_{ij})$. where $w_{ij} = w_{ji} \geq 0$ measures the similarity, association, or correlation between data objects i, i. The scaling factor $D = (d_i)$ is a diagonal matrix with each diagonal element being the sum of the corresponding row ($d_i = \sum_j w_{ij}$). Noting that $W = D^{1/2}(D^{-1/2}WD^{-1/2})D^{1/2}$, we apply PCA or spectral decomposition on the scaled matrix $\widehat{W} = D^{-1/2}WD^{-1/2}$ instead of on W directly, leading to

$$W = D^{1/2}(\sum_k \mathbf{z}_k \lambda_k \mathbf{z}_k^T)D^{1/2} = D \sum_k \mathbf{q}_k \lambda_k \mathbf{q}_k^T D \tag{1}$$

Here we call $\mathbf{q}_k = D^{-1/2}\mathbf{z}_k$ the *scaled principal components* ($\mathbf{q}_k, \mathbf{z}_k$ are n-vectors[1]); they are obtained by solving the eigenvalue system

$$D^{-1/2}WD^{-1/2}\mathbf{z} = \lambda \mathbf{z}. \tag{2}$$

or equivalently, solving

$$W\mathbf{q} = \lambda D\mathbf{q}. \tag{3}$$

Self-aggregation

The K-dimensional space spanned by the first K scaled principal components (SPCA space) has an interesting self-aggregation property enforced by within-cluster association (connectivity). This property is first noted in [2].

First, we consider the case where clusters are well separated, i.e., no overlap (no connectivity) exists among the clusters.

Theorem 1. When overlaps among K clusters are zero, the K scaled principal components $(\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_K) = Q_K$ get the same maximum eigenvalue: $\lambda_1 = \cdots = \lambda_K = 1$. Each $\mathbf{q}_k$ is a multistep (piecewise-constant) function (assuming objects within a cluster are indexed consecutively). In the SPCA space spanned by Q_K, all objects within the same cluster self-aggregate into a single point. □

[1] Here bold-face lowercase letters are vectors of size n, with $q_k(i)$ as the ith element of $\mathbf{q}_k$. Matrices are denoted by uppercase letters.

Proof. Now $W = (W_{pq})$ is block diagonal: $W_{pq} = 0, p \neq q$. Assume $K = 3$. Define basis vectors:

$$\mathbf{x}^{(k)} = (0 \cdots 0, D_{kk}^{1/2}\mathbf{e}_k, 0 \cdots 0)^T, \tag{4}$$

where $s_{pq} = \sum_{i \in G_p} \sum_{j \in G_q} w_{ij}$, $D_{pq} = \mathrm{diag}(W_{pq}\mathbf{e}_q)$, and $\mathbf{e}_k = (1, \cdots, 1)^T$ with the size of cluster G_k. $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}$ are eigenvectors of Eq.(2) with $\lambda^{(0)} = 1$. For any K real numbers $\mathbf{c} = (c_1, c_2, \cdots, c_K)^T$, $\mathbf{z} = X_K\mathbf{c} = c_1\mathbf{x}^{(1)} + \cdots + c_3\mathbf{x}^{(K)}$ is also an eigenvector of Eq.(2) with $\lambda^{(0)} = 1$. The corresponding scaled principal component

$$\mathbf{q} = D^{-1/2}\mathbf{z} = (c_1\mathbf{e}_1/s_{11}^{1/2}, \cdots, c_K\mathbf{e}_K/s_{KK}^{1/2})^T, \tag{5}$$

is a K-step piece-wise constant function. Clearly, all data objects within the same cluster have *identical* elements in $\mathbf{q}$. The coordinate of object i in the K-dim SPCA space is $\mathbf{r}_i = (q_1(i), \cdots, q_K(i))^T$. Thus objects within a cluster are located at (self-aggregate into) the same point. □

Scaled principal components are not *unique* when no overlap between clusters exist. For a set of K scaled principal components $(\mathbf{q}_1, \cdots, \mathbf{q}_K) = Q_K$, and another arbitrary $K \times K$ orthonormal matrix R, $Q_K R$ are also a vaild set of scaled principal components. However, the expansion of Eq.(1) is unique, because $\mathbf{q}_k\mathbf{q}_k^T$ is unique. Thus, self-aggregation of cluster member is equivalent to the fact that $Q_K Q_K^T$ has a block diagonal structure,

$$Q_K Q_K^T = \mathrm{diag}(\mathbf{e}_1\mathbf{e}_1^T/s_{11}, \cdots, \mathbf{e}_K\mathbf{e}_K^T/s_{KK}), \tag{6}$$

where elements within the same diagonal block all have the same value. In graph theory, the scaled PCA represents each cluster as a complete graph (clique). For this reason, the truncated SPCA expansion

$$W_K = D\sum_{k=1}^{K}\mathbf{q}_k\mathbf{q}_k^T D = DQ_K Q_K^T D \tag{7}$$

is particularly useful in discovering cluster structure. Here we retain only first K terms and set $\lambda_k = 1$ which is crucial for enforcing the cluster structure later.

Second, we consider the case when overlaps among different clusters exist. We apply perturbation analysis by writing $\widehat{W} = \widehat{W}^{(0)} + \widehat{W}^{(1)}$, where $\widehat{W}^{(0)}$ is the similarity matrix for the zero-overlap case considered above, and $\widehat{W}^{(1)}$ accounts for the overlap among clusters and is treated as a perturbation.

Theorem 2. At the first order, the K scaled principal components and their eigenvalues have the form

$$\mathbf{q} = D^{-1/2}X_K\mathbf{y}, \ \lambda = 1 - \zeta,$$

where $\mathbf{y}$ and ζ satisfy the eigensystem $\Gamma\mathbf{y} = \zeta\mathbf{y}$. The matrix Γ has the form $\Gamma = \Omega^{-1/2}\ \bar{\Gamma}\ \Omega^{-1/2}$, where

$$\bar{\Gamma} = \begin{pmatrix} h_{11} & -s_{12} & \cdots & -s_{1K} \\ -s_{21} & h_{22} & \cdots & -s_{2K} \\ \vdots & \vdots & \cdots & \vdots \\ -s_{K1} & -s_{K2} & \cdots & h_{KK} \end{pmatrix} \tag{8}$$

$h_{kk} = \sum_{p \neq k} s_{kp}$($p$ sums over all indices except k) and $\Omega = \mathrm{diag}(s_{11}, \cdots, s_{KK})$. This analysis is accurate to order $||\widehat{W}^{(1)}||^2/||\widehat{W}^{(0)}||^2$ for eigenvalues and to order $||\widehat{W}^{(1)}||/||\widehat{W}^{(0)}||$ for eigenvectors. $\square$

The proof is a bit involved and is omitted here. Several features of SPCA can be obtained from Theorem 2:

Corollary 1. SPCA expansion $W_K = DQ_K Q_K^T D = D^{1/2} X_K X_K^T D^{1/2}$ has the same block diagonal form of Eq.(6) within the accuracy of Theorem 1.

Corollary 2. The first scaled principal component is $\mathbf{q}_1 = D^{-1/2} X_K \mathbf{y}_1 = (1, \cdots, 1)^T$ with $\lambda_1 = 1$. λ_1 and $\mathbf{q}_1$ are also the exact solutions to the original Eq.(3).

Corollary 3. The second principal component for $K = 2$ is

$$\mathbf{q}_2 = D^{-1/2} X_2 \mathbf{y}_2 = (\sqrt{\frac{s_{22}}{s_{11}}}\ \mathbf{e}_1, -\sqrt{\frac{s_{11}}{s_{22}}}\ \mathbf{e}_2)^T. \tag{9}$$

The eigenvalue is

$$\lambda_2 = 1 - (s_{12}/s_{11} + s_{12}/s_{22}). \tag{10}$$

The diagonal block structure of the SPCA expansion W_K (Corollary 1) implies that objects within the same cluster will self-aggregate as in Theorem 1. We can also see this more intuitively. A scaled principal component $\mathbf{q} = (q(1), \cdots, q(n))^T$, as an eigenvector of Eq.(3), can be equivalently obtained by minimizing the objective function

$$\min_{\mathbf{q}} \frac{\sum_{ij} w_{ij} [q(i) - q(j)]^2}{\sum_i d_i [q(i)]^2}. \tag{11}$$

Thus adjacent objects have close coordinates such that $[q(i) - q(j)]^2$ is small for non-zero w_{ij}: the larger w_{ij} is, the closer $q(i)$ is to $q(j)$.

To illustrate the above analysis, we provide the following example and applications.

Example 1. A dataset of 3 clusters with substantial random overlap between the clusters. All edge weights are 1. The similarity matrix and results are shown in Fig.1, where nonzero matrix elements are shown as dots. The exact λ_2 and approximate $\tilde{\lambda}_2$ from Theorem 2 are close:

$$\lambda_2 = 0.300,\ \tilde{\lambda}_2 = 0.268.$$

The SPCA expansion $W_K = DQ_K Q_K^T D$ reveals the correct block structure clearly due to self-aggregation: in W_K connections between different clusters are substantially suppressed while connections within clusters are substantially enhanced. Thus W_K is much sharper than the original weight matrix W. In SPCA space using coordinates $\mathbf{r}_i = (q_1(i), \cdots, q_3(i))^T$, objects within the same cluster become almost on top of each other (not shown) as the result of self-aggregation.

Application 1. In DNA micro-array gene expression profiling, responses of thousands of genes from tumor tissues are simultaneously measured. We SPCA to gene expression profiles of lymphoma cancer data from Alizadeh et al. [1]. Discovered

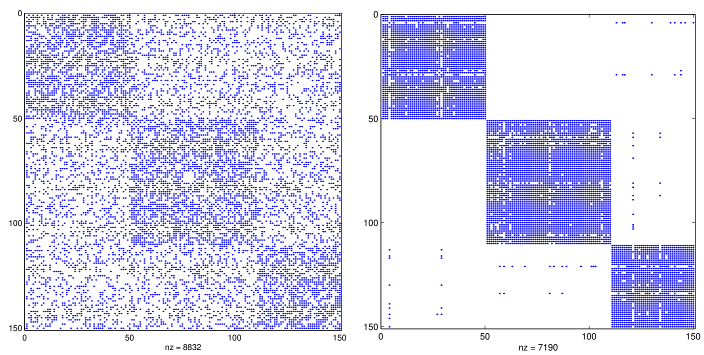

Figure 1: Left: similarity matrix W. Diagonal blocks represent weights inside clusters and off-diagonal blocks represent overlaps between clusters. Right: Computed W_K.

clusters clearly correspond to normal or cancerous subtypes identified by human expertise. 100 most informative genes (defines the Euclidean space) are selected out of the original 4025 genes based on the F-statistic. Pearson correlation c_{ij} is computed and similarity $w_{ij} = \exp(c_{ij}/\langle c \rangle)$, where $\langle c \rangle = 0.1$. Three cancer and three normal subtypes are shown with symbols explained in Figure 2B (the number of samples in each subtype is shown in parentheses). This is a difficult problem due to large variances in cluster sizes. Self-aggregation is evident in Figures 2B and 2C.

Besides the self-aggregation, the nonlinearity in SPCA can alter the topology in a useful way to reveal structures which are otherwise difficult to detect using standard PCA. Thus the SPCA space is a more useful space to explore the structures.

Application 2. 1000 points form two interlocking rings (but not touching each other) in 3D Euclidean space. The similarities between data points are computed same as in Application 1. In SPCA space, rings are separated. Objects self-aggregate into much thinner rings (shown in right panel of Figure 2).

Dynamic aggregation

The self-aggregation process can be repeated to obtain sharper clusters. W_K is the low-dimensional projection that contains the essential cluster structure. Combining this structure with the original similarity matrix, we obtain a new similarity matrix containing sharpened cluster information:

$$W^{(t+1)} = (1-\alpha)W_K^{(t)} + \alpha W^{(t)}, \tag{12}$$

where $W_K^{(t)}$ is the SPCA representation (Eq.7) of $W^{(t)}$, the weight matrix at t-th iteration, $\alpha = 0.5$, and $W^{(1)} = W$.

Applying SPCA on $W^{(2)}$ leads to further aggregation (see Figure 2C). The eigenvalues of the 1st and 2nd SPCA are shown in the insert in Figure 1C. As

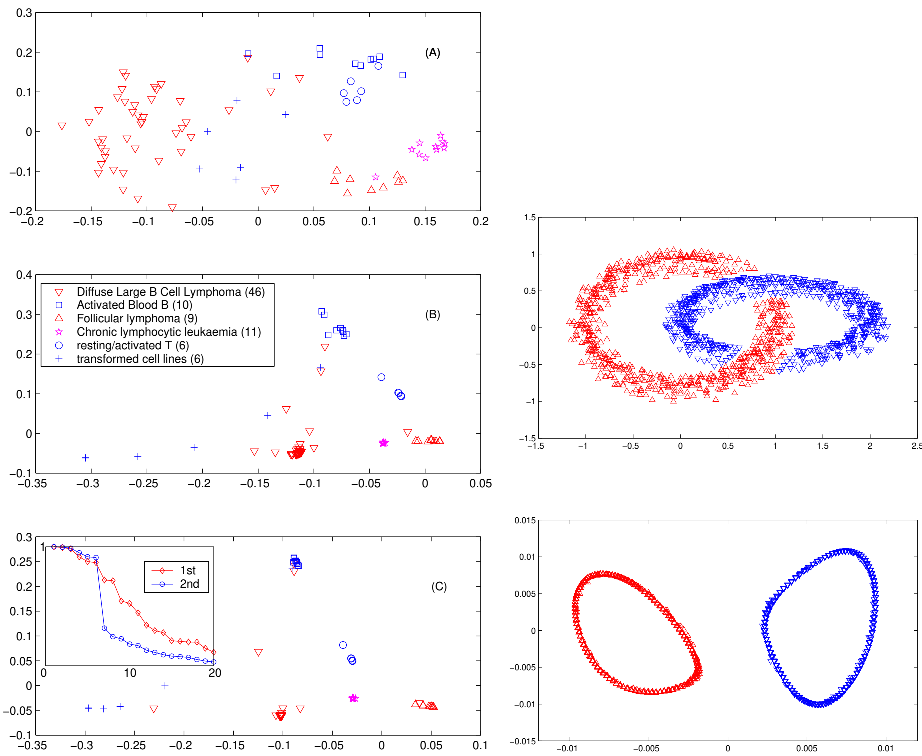

Figure 2: Left: Gene expression profiles in original Euclidean space (A), in SPCA space (B), and in SPCA space after one iteration of Eq.13 (C). In all 3 panels, objects are plotted in 2D-view spanned by the first two PCA components. Cluster structures become clearer due to self-aggregation. The insert in (C) shows the eigenvalues of the 1st and 2nd SPCA.
Right: Data objects in 3D Euclidean space (top) and in SPCA space (bottom).

iteration proceeds, a clear gap is developed, indicating that clusters becoming more separated.

Noise reduction

SPCA representation W_K has noises. For example, W_K has sometimes negative weights $(W_K)_{ij}$ whereas we expect them to be nonnegative for learning. However, a nice property of SPCA provides a solution. The structure of W_K is determined by QQ^T. When data contains K well separated clusters, QQ^T has a diagonal block structure and every elements in the block are identical (Eq.6). When clusters are not well separated but can be meaningfully distinguished, QQ^T has approximately the same block-diagonal form (Corollary 1). This property allows us to interpret QQ^T as the probability that two objects i, j belong to the same cluster:

$$p_{ij} = (W_K)_{ij}/(W_K)_{ii}^{1/2}(W_K)_{jj}^{1/2}.$$

which is the same as $p_{ij} = (QQ^T)_{ij}/(QQ^T)_{ii}^{1/2}(QQ^T)_{jj}^{1/2}$. To reduce noise in the above dynamic aggregation, we set

$$(W_K)_{ij} = 0 \quad \text{if} \quad p_{ij} < \beta, \tag{13}$$

where $0 < \beta < 1$ and we chose $\beta = 0.8$. Noise reduction is an integral part of SPCA. In general, the method is stable: the final results are insensitive to α, β. The above dynamic aggregation process repeatedly projects data into SPCA space and the self-aggregation forces data objects towards the attractors of the dynamics. The attractors are the desired clusters which are well separated and their principal eigenvalues approach 1 (see insert in Fig.1C). Usually, after one or two iterations of self-aggregation in SPCA, the cluster structure becomes evident.

3 Mutual Dependence

In many learning and information processing tasks, we look for inter-dependence among different aspects (attributes) of the same data objects. In gene expression profiles, certain genes express strongly when they are from tissues of a certain phenotype, but express mildly when they are from other phenotypes[1]. Thus it is meaningful to consider gene-gene correlations as characterized by their expressions across all tissue samples, in addition to sample-sample correlations we usually study.

In text processing, such as news articles, the content of an article is determined by the word occurrences, while the meaning of words can be inferred through their occurrences across different news articles. This kind of association between a data object (tissues, news articles) and its attributes (expressions of different genes, word occurrences) is represented by the asymmetric data association matrix. Here we restrict our consideration to the cases where all entries of assoication matrix B are non-negative, and therefore can be viewed as the probability of association (conditional probability) between column objects (news articles or tissue samples) and row objects (words or genes). This kind of data is sometimes called a contingency table. In graph theory, B is the weight matrix for a bipartite graph. Clustering row and column objects simultaneously amounts to clustering the bipartite graph as shown in Figure 3.

SPCA applies to these inter-dependence problems (bipartite graphs) as well. We introduce nonlinear scaling factors, diagonal matrices D_r (each element is the sum of a row) and D_c (each element is the sum of a column). Let $B = D_r^{1/2}(D_r^{-1/2}BD_c^{-1/2})D_c^{1/2}$. Applying PCA on $\widehat{B} = D_r^{-1/2}BD_c^{-1/2}$, we obtain

$$B = D_r^{1/2}(\sum_k \mathbf{u}_k \lambda_k \mathbf{v}_k^T)D_c^{1/2} = D_r \sum_k \mathbf{f}_k \lambda_k \mathbf{g}_k^T D_c. \tag{14}$$

Scaled principal components are $\mathbf{f}_k = D_r^{-1/2}\mathbf{u}_k$ for row objects and $\mathbf{g}_k = D_c^{-1/2}\mathbf{v}_k$ for column objects.

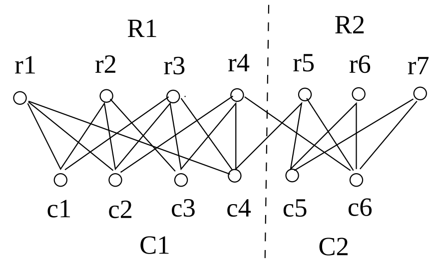

Figure 3: A bipartite graph with row-nodes and column-nodes. The dashed line indicates a possible clustering.

Scaled principal components here have the same self-aggregation and related properties as in §2. First, the singular vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ and the singular values λ_k are determined though

$$(\widehat{B}\widehat{B}^T)\mathbf{u} = \lambda^2\mathbf{u}, \quad (\widehat{B}^T\widehat{B})\mathbf{v} = \lambda^2\mathbf{v}. \tag{15}$$

They can be viewed as simultaneous solutions to Eq.(3), with

$$W = \begin{pmatrix} & B \\ B^T & \end{pmatrix}, \; D = \begin{pmatrix} D_r & \\ & D_c \end{pmatrix}, \; \mathbf{z} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix},$$

as can be easily verified. Therefore, all conclusions of Theorems 1 and 2 for undirect graphs can readily extended over to the bipartite graph case here.

When K clusters are well separated (no overlap among clusters), we have **Theorem 3**. For well separated clusters, row objects within the same cluster will self-aggregate in the SPCA space spanned by $(\mathbf{f}_1, \cdots, \mathbf{f}_K) = F_K$, while column objects within the same cluster will self-aggregate in the SPCA space spanned by $(\mathbf{g}_1, \cdots, \mathbf{g}_K) = G_K$. $\square$

When clusters overlap, a theorem almost identical to Theorem 2 can be established for bipartite graphs. The corollaries following Theorem 2 can be nearly identically extended to the bipartite graphs. We briefly summarize the results here. Let

$\mathbf{q}_k = \begin{pmatrix} \mathbf{f}_k \\ \mathbf{g}_k \end{pmatrix} = D^{-\frac{1}{2}} \begin{pmatrix} \mathbf{u}_k \\ \mathbf{v}_k \end{pmatrix}$, $Q_K = \begin{pmatrix} F_K \\ G_K \end{pmatrix}$, and $Q_K Q_K^T = \begin{pmatrix} F_K F_K^T & F_K G_K^T \\ G_K F_K^T & G_K G_K^T \end{pmatrix}$.

The low-dimensional SPCA expansion $B_K = D_r \sum_{k=1}^K \mathbf{f}_k \mathbf{g}_k^T D_c = D_r F_K G_K^T D_c$ gives the sharpened association between words and documents, the diagonal block structure of $F_K F_K^T$ gives the clusters on row objects (words) while the diagonal block structure of $G_K G_K^T$ simultaneously gives the clusters on column objects (news articles). We note that Eqs.(14,15) rediscover the correspondence analysis [6] in multivariate statistics from the SPCA point of view.

Example 2. We apply the above analysis to a bipartite graph example with association matrix shown in Fig.4. The bipartite graph has two dense clusters with large overlap between them. The SPCA representations are computed and shown in Fig.4. $F_K G_K^T$ gives a sharpened association matrix where the overlap between clusters (off-diagonal blocks) is greatly reduced. $F_K F_K^T$ reveals the cluster structure for row objects and $G_K G_K^T$ reveals the cluster structure for column objects.

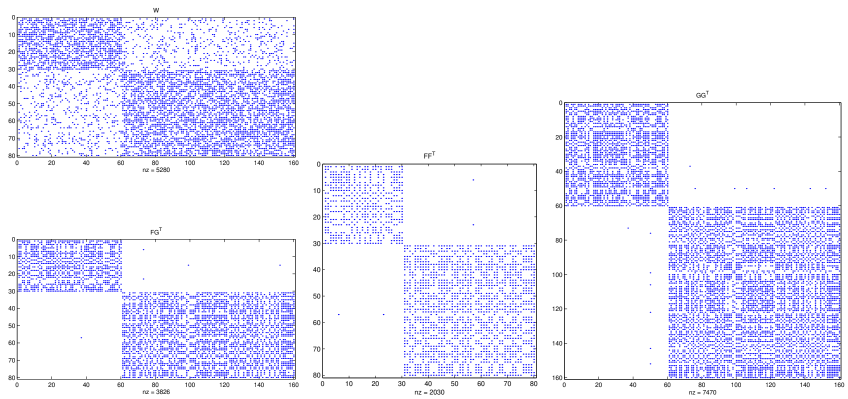

Figure 4: Left-top: association (weight) matrix of a bipartite graph of 2 dense clusters (diagonal blocks) with random overlaps (off-diagonal blocks). Left-bottom: $F_K G_K^T$ for sharpened associations. Middle: $G_K G_K^T$ for clustering column objects. Right: $F_K F_K^T$ for clustering row objects.

Application 4. Clustering internet newsgroups (see Figure 5). (The newsgroup dataset is from www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html.) Five newsgroups are used in the dataset (listed in upper right corner with corresponding color). 100 news articles are randomly chosen from each newsgroup. From them 1000 words selected. Standard `tf.idf` weighting are used. Each document (column) is normalized to one. The resulting word-to-document association matrix is the input matrix B. As shown in Figure 5, words aggregate in SPCA word space (spanned by F_K) while news articles are simultaneously clustered in SPCA document space (spanned by G_K) shown by the projection matrix GG^T (the insert). One can see that GG^T indicates some overlap between *computer graphics* and *space science*, which is consistent with the relative closeness of the two corresponding word clusters in word space. The accuracy of clustering is 86%. (We also computed the cosine similarity $W = B^T B$ and use the method in §2 to obtain clusters with 89% accuracy.) This dataset has been extensively studied in [22]; the standard Kmeans clustering gives about 66% accuracy, while two improved methods get 76-80% accuracy.

4 Discussions

In this paper, we assume that objects belonging to the same cluster are consecutively indexed. However, the SPCA framework is independent of the indexing. The diagonal block structure of SPCA representation as the result of cluster member self-aggregation merely indicates the fact that connectivities between different

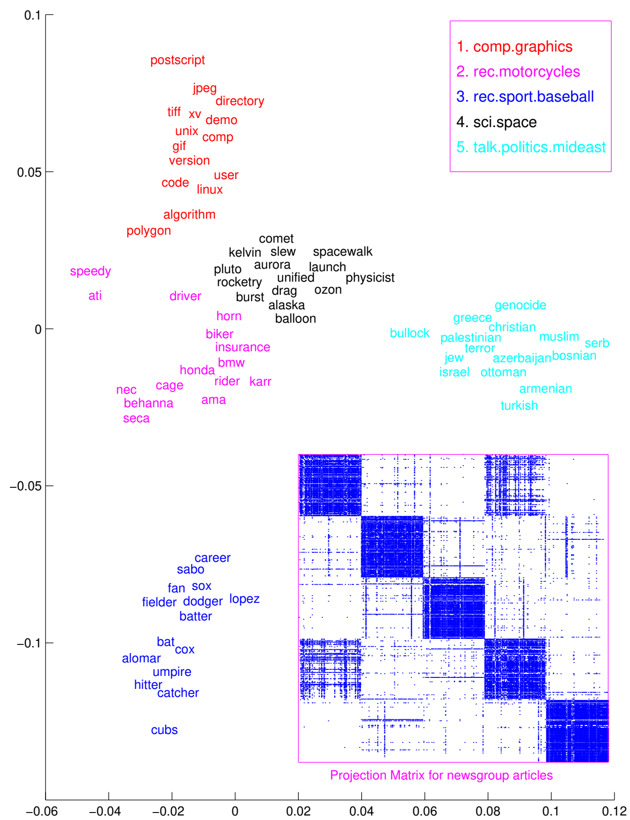

Figure 5: Words self-aggregate in SPCA word space while internet newsgroups articles are simultaneously clustered. Shown are the top 15 most frequently occurring words from each discovered cluster. (Several words in *motorcycles* are brand names, and several words in *baseball* are players' names.) The insert shows the projection matrix GG^T on clustering news articles.

clusters are substantially suppressed while connectivities within a cluster are substantially enhanced. Our main results, in essence, is that if cluster structures in the original dataset can be meaningfully distinguished, such as Figures 1,2, SPCA makes them much more well-separated so that clusters can be easily detected either by direct visual inspection or by a standard clustering method such as the K-means algorithm.

The key to understand SPCA is the nonlinear scaling factor D. Columns and rows of the similarity matrix are scaled inversely proportional to their weights such that all K principal components get the same maximum eigenvalue of one. This happens independent of cluster sizes. leading to desirable consequences. (i) Outliers are usually far away from other objects and often skew the statistics (means, covariance, etc) in original Euclidean space. However, in SPCA we focus on similarity matrix (instead of distance matrix). Outliers contribute very little to the quantities in Eqs.(4,8) thus do not adversely affects SPCA. But, their small similarities with other objects force them to appear as independent clusters and thus can be easily detected. (ii) Unbalanced clusters (in which the number of objects in each cluster varies substantially) are usually difficult to discover using many other clustering methods, but can be effectively dealt with in SPCA due to the nonlinear scaling. Directly applying PCA on W will be dominated by the large clusters and no self-aggregation will occur.

The scaled PCA has a connection to spectral graph partitioning and clustering [4, 5, 17, 7, 20, 3, 16]. Given a weighted graph G where the weight w_{ij} is the similarity between nodes i, j, one wish to partition it into two subgraphs (clusters) A, B according to the *min-max clustering principle*: the (overlapping) similarity between A and B is minimized while similarities within A or B are maximized[3]. The overlap between A and B is the sum of weights between A and B, $s(A, B) = \sum_{i \in A, j \in B} w_{ij}$. The similarity within cluster A is $s(A, A)$ (sum of all edge weights within A). The similarity within cluster B is $s(B, B)$. Thus the clustering principle of minimizing $s(A, B)$ while maximizing $s(A, A)$ and $s(B, B)$ leads to the min-max cut objective function[3],

$$J_{\text{MMC}} = \frac{s(A,B)}{s(A,A)} + \frac{s(A,B)}{s(B,B)}. \tag{16}$$

The clustering result can be represented by an indicator vector $\mathbf{q}$, where $q(i) = a$ or -b depending on node $i \in A$ or B. (a and b are positive constants.) If one relaxes $q(i)$ from discrete indicators to continuous values in $(-1, 1)$, the solution $\mathbf{q}$ for minimizing J_{MMC} is given by the eigenvector of $(D - W)\mathbf{q} = \zeta D\mathbf{q}$, which is exactly Eq.3 with $\lambda = 1 - \zeta$. This further justifies our SPCA approach for unsupervised learning. In addition, the desired clustering indicator vector $\mathbf{q}$ is precisely recovered in Eq.9 with $a = \sqrt{s_{22}/s_{11}}$ and $b = \sqrt{s_{11}/s_{22}}$ due to Theorem 1; minimizing the min-max cut objective of Eq.16 is equivalent to maximizing the eigenvalue of the second SPCA component given in Eq.10. All these indicate that SPCA is a principled and coherent framework for data clustering. One drawback of the method is the computation is in general $O(n^2)$.

In self-aggregation, data objects move towards each other guided by connectivity which determines the attractors. This is similar to the self-organizing map [14, 11], where feature vectors self-organize into a 2D feature map while data objects remain fixed. All these have a connection to recurrent networks [12, 10]. In Hopfield network, features are stored as associative memories. In more complicated networks, connection weights are dynamically adjusted to learn or discover the patterns, much like the dynamic aggregation of Eq.(12). Thus it may be possible to construct a recurrent network that implements the self-aggregation. In this network, high dimensional input data are converted into low-dimensional representations in SPCA space and cluster structures emerge as the attractors of dynamics.

References

[1] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[2] C. Ding, X. He, and H. Zha. A spectral method to separate disconnected and nearly-disconnected web graph components. In *Proc. ACM Int'l Conf Knowledge Disc. Data Mining (KDD)*, pages 275–280, 2001.

[3] C. Ding, X. He, H. Zha, M. Gu, and H. Simon. A min-max cut algorithm for graph partitioning and data clustering. *Proc. IEEE Int'l Conf. Data Mining (ICDM)*, pages 107–114, 2001.

[4] W.E. Donath and A. J. Hoffman. Lower bounds for partitioning of graphs. *IBM J. Res. Develop.*, 17:420–425, 1973.

[5] M. Fiedler. Algebraic connectivity of graphs. *Czech. Math. J.*, 23:298–305, 1973.

[6] M. J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic press, 1984.

[7] L. Hagen and A.B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE. Trans. on Computed Aided Desgin*, 11:1074–1085, 1992.

[8] T. Hastie and W. Stuetzle. Principal curves. *J. Amer. Stat. Assoc*, 84:502–516, 1989.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning*. Springer Verlag, 2001.

[10] S.S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1998, 2nd ed.

[11] J. Himberg. A som based cluster visualization and its application for false coloring. *Proc Int'l Joint Conf. Neural Networks*, pages 587–592, 2000.

[12] J.J. Hopfield. Neural networks and physical systems with emergent collective computation abilities. *Proc. Natl Acad Sci USA*, 79:2554–2558, 1982.

[13] I.T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

[14] T. Kohonen. *Self-organization and Associative Memory*. Springer-Verlag, 1989.

[15] M.A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37:233–243, 1991.

[16] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Proc. Neural Info. Processing Systems (NIPS 2001)*, 2001.

[17] A. Pothen, H. D. Simon, and K. P. Liou. Partitioning sparse matrices with egenvectors of graph. *SIAM Journal of Matrix Anal. Appl.*, 11:430–452, 1990.

[18] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[19] B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[20] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE. Trans. on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

[21] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[22] H. Zha, C. Ding, M. Gu, X. He, and H.D. Simon. Spectral relaxation for K-means clustering. *Advances in Neural Information Processing Systems 14 (NIPS'01)*, pages 1057–1064, 2002.