Self-Aggregation in Scaled Principal Component Space

Chris H.Q. Ding[a], Xiaofeng He[a,b], Hongyuan Zha[b], Horst D. 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

October 5, 2001

Abstract

Automatic grouping of voluminous data into meaningful structures is a challenging task frequently encountered in broad areas of science, engineering and information processing. These *data clustering* tasks are frequently performed in Euclidean space or a subspace chosen from principal component analysis (PCA). Here we describe a space obtained by a nonlinear scaling of PCA in which data objects *self-aggregate* automatically into clusters. Projection into this space gives sharp distinctions among clusters. Gene expression profiles of cancer tissue subtypes, Web hyperlink structure and Internet newsgroups are analyzed to illustrate interesting properties of the space.

Introduction

Scientific discoveries are often made in the appropriate spaces. For example, many physical processes are analyzed in Fourier space rather than in Euclidean space. In recent decades, principal component analysis (PCA) [1] 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. This is particularly useful for *data clustering*, organizing observed data into groups or clusters, to discover meaningful differences and commonalities among data objects [2, 3]. For example, in micro-array gene expression profiling, tissue samples can be automatically grouped together according to their phenotypes [4, 5] (see Figure 1). Such groupings automatically generate meaningful structures, that are particularly useful for many knowledge discovery process [6], in climate patterns[7], image segmentation[8, 9], etc.

PCA is a coordinate rotation such that the principal components span the dimensions of largest variance. The *linear* transformation preserves local properties and global topologies, and can be efficiently computed [10]. However, it is not effective in revealing nonlinear structures and several nonlinear mappings [12, 11] have been recently developed. Here, we report that a nonlinear scaling of PCA leads to a space in which data objects self-aggregate, and therefore is most suitable for data clustering.

Scaled principal components

Associations between data objects are usually based on a similarity metric, such as the correlation, covariance or quantities inversely proportional to the distance metric. The scaled principal component approach starts with a nonlinear (non-uniform) scaling of the similarity matrix S. The scaling factor D is a diagonal matrix and each diagonal element is the sum of the corresponding row ($d_i = \sum_j S_{ij}$). Noting that $S = D^{1/2}(D^{-1/2}SD^{-1/2})D^{1/2}$, we apply PCA or spectral decomposition on the scaled matrix $D^{-1/2}SD^{-1/2}$ instead on S directly, leading to

$$S = D^{1/2}(\sum_k \mathbf{u}_k \lambda_k \mathbf{u}_k^T)D^{1/2} = D\sum_k \mathbf{f}_k \lambda_k \mathbf{f}_k^T D \tag{1}$$

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

$$D^{-1/2}SD^{-1/2}\mathbf{u} = \lambda\mathbf{u}. \tag{2}$$

Self-aggregation

The K-dimensional space spanned by the first K scaled principal components (SPCA space) has an interesting self-aggregation property enforced by intra-cluster association (connectivity). First, when no overlap exists between different clusters, the K scaled

principal components get the same maximum eigenvalue: $\lambda_1 = \cdots = \lambda_K = 1$. This is due to the nonlinear scaling and is independent of the size of each cluster. As a consequence, objects within a cluster will coincide with each other in SPCA space, resulting in K distinguishable points for K clusters [13].

Second, when overlaps between different clusters exist, objects within the same cluster become much more closer in SPCA space than they are in Euclidean space. This is because the coordinates (elements of any scaled principal component $\mathbf{f}$) can be equivalently obtained by [14, 18] minimizing the objective function

$$\min_{\{f_i\}} \frac{\sum_{ij}(f_i - f_j)^2 S_{ij}}{\sum_i d_i f_i^2}. \tag{3}$$

Thus adjacent objects have close values, such that $(f_i - f_j)^2$ is close to zero for non-zero S_{ij}. The self-aggregation is evident in Figure 1 showing gene expression profiles of lymphoma cancer (from Alizadeh et al. [4]).

Besides the self-aggregation, the nonlinearity in SPCA can alter the topology in a useful way to reveal structures (as shown in Figure 2) which are otherwise difficult to detect using standard PCA. Thus the SPCA space is a more useful space to explore the structures.

Iterative-aggregation

The self-aggregation process can be repeated to obtain sharper clusters (see Figure 1(C)). This is done by truncating the expansion in Eq.(1) at K terms, and setting $\lambda_1 = \cdots = \lambda_K = 1$. We have $S \simeq D \sum_{k=1}^{K} \mathbf{f}_k \lambda_k \mathbf{f}_k^T D = DFF^T D$, where $F = (\mathbf{f}_1, \cdots, \mathbf{f}_K)$. As in PCA, $DFF^T D$ 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: $S^{(2)} = (1-\alpha)DFF^T D + \alpha S$, where $\alpha = 0.5$. Applying SPCA on $S^{(2)}$ leads to further aggregation (see Figure 1).

The structure of $DFF^T D$ is determined by FF^T. When data contains K well separated clusters, FF^T has a diagonal block structure[13]. When clusters are not well separated but can be meaningfully distinguished, FF^T has approximate block-diagonal form[19]. Thus one can interpret FF^T as the probability that two objects i, j belong to the same cluster: $p_{ij} = (FF^T)_{ij}/(FF^T)_{ii}(FF^T)_{jj}$. To reduce noise in the above iterative aggregation, we set $(FF^T)_{ij} = 0$ if $p_{ij} < \beta$ where $0 < \beta < 1$ and we chose $\beta = 0.8$. In general, the final results are insensitive to α, β [22] and the method is stable. The above iterative aggregation process repeatedly projects data into SPCA space and the self-aggregation makes clusters well separated and their principal eigenvalues approaching 1 (see insert in Fig.1C).

After self-aggregation in SPCA, the cluster structure is usually very evident. One can use the traditional approaches such as K-means and EM to get precise cluster

structures. Alternatively, one can examine the projection matrix FF^T and use the probability interpretation to obtain cluster structures.

The hyperlinked network of the World Wide Web (specified in a hyperlink adjacency matrix) can be usefully analyzed in SPCA space. First, fragmented networks (isolated segments) [23] appear as distinct points in SPCA space and are easily identified (Fig.3). Second, a large segment is split into sparsely connected sub-segments. Several segments are shown in Fig.4. All discovered segments are meaningful structures [24].

Mutual Dependence

In many scientific research and information processing tasks, we look for inter-dependence between 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[4]. Thus it is meaningful to consider gene-gene correlations as characterized by their expressions across all tissue samples. This is different from tissue-tissue relationship considered above and shown in Figure 1.

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 matrix. Here we restrict our consideration to the cases where all entries of data matrix P are non-negative, and therefore can be viewed as the probability of association between column objects (as news articles or tissue samples) and row objects (words or genes). This kind of data is sometimes called a contingency table.

SPCA applies to these inter-dependence problems 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), and write $P = D_r^{1/2}(D_r^{-1/2}PD_c^{-1/2})D_c^{1/2}$. Applying PCA on $\widehat{P} = D_r^{-1/2}PD_c^{-1/2}$, we obtain

$$P = D_r^{1/2}(\sum_k \mathbf{u}_k\lambda_k\mathbf{v}_k^T)D_c^{1/2} = D_r\sum_k \mathbf{f}_k\lambda_k\mathbf{g}_k^T D_c. \tag{4}$$

Scaled principal components $\mathbf{f}_k = D_r^{-1/2}\mathbf{u}_k, \mathbf{g}_k = D_c^{-1/2}\mathbf{v}_k$ have the same self-aggregation and related properties because they can be viewed [25] as simultaneous solutions to Eq.(2). Again, as a low-dimensional projection, we obtain projection matrices FF^T, GG^T, FG^T [26]. The block structure of FF^T gives the clusters on row objects (words) while the block structure of GG^T simultaneously gives the clusters on column objects (news articles). Furthermore, FG^T gives the correspondence between words and documents. Results of an analysis of newsgroups on the internet is shown in Fig.5.

The relationship between row- and column-type objects is sometimes analyzed using a statistical technique called correspondence analysis [27, 28], which emphasizes a geometric interpretation. SPCA re-derives correspondence analysis [29] from the general principles of PCA and also bring along the above useful tools for analysis.

The key to understanding 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 [13, 19]. This happens independent of cluster sizes. For example, as illustrated in Fig.3, one cluster has 2181 webpages and another cluster has 2, but both of their corresponding principal eigenvalues are one. This "equalization" of importance has two desirable consequences.

1. In automatic data analysis, outliers often skew the picture and should be detected and eliminated prior to analysis. In SPCA space, outliers appear as independent clusters, since they are far away from other objects. Thus they can be easily detected.

2. SPCA is effective for unbalanced clusters (number of objects in each cluster vary substantially), which are usually difficult for many other clustering methods. Without the nonlinear scaling, direct PCA on S will be dominated by the large clusters and no self-aggregation will occur.

In self-aggregation, data objects move towards each other guided by connectivity. This differs from hill climbing, where data objects move towards to direction of higher density [30], which is difficult to estimate in high dimensions and is sensitive to parameter choices. Both of these differ further from the self-organizing map [31], where the feature vectors move around to form a feature map while data objects remain stationary.

The scaled PCA has a connection[14] to spectral graph partitioning, which uses the second eigenvector of the Laplace matrix to partition a graph into components (clusters) [15, 16, 17, 32, 9, 33]. Using several eigenvectors for partitioning have also been studied [17, 34, 35, 36, 9]. SPCA reformulates the problem as the dimensionality reduction of PCA and points out the subtle difference between principal components $\mathbf{f}$ and the eigenvectors $\mathbf{u}$. Self-aggregation then becomes clear following the perturbation analysis outlined in [13, 19, 21].

Given the wide use of PCA and the diverse applications illustrated here, we anticipate that more challenging problems will be solved and interesting new structures will be discovered in the scaled PCA space.

Acknowledgements. We thank M. Gu for discussions. This work is supported by Department of Energy under contract DE-AC03-76SF00098 and in part by National Science Foundation Grant CCR-9901986.

References

[1] K.V. Mardia, J.T. Kent, and J.B. Bibby. *Multivariate Analysis.* Academic Press, 1979.

[2] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31:264–323, 1999.

[3] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification, 2nd ed.* Wiley, 2000.

[4] A.A. Alizadeh, M.B. Eisen, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[5] T.R. Golub, D.K. Slonim, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[6] U. Fayyad, D. Haussler, and P. Stolorz. Mining scientific data. *Communications of ACM*, 39:1501–1531, 1996.

[7] P. Smyth, M. Ghil, and K. Ide. Multiple regimes in northern hemisphere height fields via mixture model clustering. *Journal of the Atmospheric Sciences*, 56:3704–3723, 1999.

[8] T. Taxt and A. Lundervold. Multi-spectral analysis of the brain using magnetic resonance imaging. *IEEE Trans. Medical Imaging*, 13:470–481, 1994.

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

[10] G. Golub and C. V. Loan. *Matrix Computations, 3rd edition.* Johns Hopkins, Baltimore, 1996.

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

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

[13] In this case, different cluster decouples and the scaled principal components are simply linear combinations of step functions. Mathematically, S is block diagonal and a vector $\mathbf{e}_i = (0 \cdots 0, 1 \cdots 1, 0 \cdots 0)^T$ is a solution to Eq.2 for the appropriate cluster with $\lambda = 1$. Any eigenvector with $\lambda = 1$ can be written as $\mathbf{f} = c_1\mathbf{e}_1/\sqrt{w_1} + \cdots + c_K\mathbf{e}_K/\sqrt{w_K}$. where w_k is the weight of kth cluster C_k, $w_k = \sum_{e_{ij} \in C_k} S_{ij}$. Hence all objects within a cluster have the same $c_1, \cdots, c_K$ and thus the same coordinates in SPCA space. Also, $FF^T = \mathrm{diag}(\mathbf{e}_1\mathbf{e}_1^T/w_1 + \cdots + \mathbf{e}_K\mathbf{e}_K^T/w_K)$ has a diagonal block structure.

[14] Eq.(1) can be equivalently written as $(D-S)\mathbf{f} = (1-\lambda)D\mathbf{f}$. $D-S$ is called the Laplacian matrix of a graph and the equation is the generalized eigenvalue problem of the Laplace matrix. Thus scaled PCA is related to the spectral graph partitioning [15, 16, 17]. The K highest eigenvectors of Eq.(1) correspond to the K lowest eigenvectors of this equation which can be re-written as minimizing the Rayleigh quotient of Eq.(2) [18, 15, 16].

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

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

[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] K. M. Hall. r-dimensional quadratic placement algorithm. *Management Science*, 17:219–229, 1971.

[19] We do a perturbation analysis by writing $\widehat{S} \equiv D^{-1/2}SD^{-1/2} = \widehat{S}^{(0)} + \widehat{S}^{(1)}$, where $\widehat{S}^{(0)}$ is the similarity matrix for zero-overlap case and $\widehat{S}^{(1)}$ accounts for the overlap between clusters and is treated as a small perburbation. The solution of $\widehat{S}^{(0)}\mathbf{u}^{(0)} = \lambda^{(0)}\mathbf{u}^{(0)}$ is given in [13]. $\widehat{S}^{(1)}$ is then solved in the subspace spanned by the largest K eigenvectors of $\widehat{S}^{(0)}$. This leads to a scaled Laplacian perturbation matrix $\widetilde{S}$. $\widetilde{S}$ determines the essential structure of SPCA. For example, FF^T has diagonal block structure as in [13]. This perturbation analysis is accurate to order $||\widehat{S}^{(1)}||^2/||\widehat{S}^{(0)}||^2$ for eigenvalues and to order $||\widehat{S}^{(1)}||/||\widehat{S}^{(0)}||$ for eigenvectors [20]. Similar perturbation analysis is used in [21].

[20] J. Mathews and R.L. Walker. *Mathematical Methods of Physics*. Addison-Wesley, 1971.

[21] C. Ding, X. He, and H. Zha. A spectral method to separate disconnected and nearly-disconnected web graph components. *Proc. 7th ACM Int'l Conf Knowledge Discovery and Data Mining (KDD 2001)*, pages 275–280, August 2001.

[22] Setting $\alpha = 0.3 - 0.7$ and $\beta = 0.3 - 0.9$ cause very small changes in the final results.

[23] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[24] For example, one segment (www.amanakaa.org) among the clusters shown in Figure 4. is on the indigenous people in the Amazon region, their culture, etc. Another segment (www.internext.com.br/ariau) is an upscale hotel in the middle of the Amazon, pointed to by the rest of the webpages in the segment, which are mostly upscale hotels specializing in safaris, ecology, and related travel. The large segment is dominated by amazon.com, which is clustered into sparsely connected sub-segments: one for Amazon.com and their European branches; another for auctions; another one relates to books on women warriors and women's issues. Other segments are not

shown. All the discovered (sub)segments correspond to meaningful communities. This could be an effective information discovery and display technology.

[25] In Eq.(2), we replace S by $\begin{pmatrix} & P \\ P^T & \end{pmatrix}$, D by $\begin{pmatrix} D_r & \\ & D_c \end{pmatrix}$ and $\mathbf{u}$ by $\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}$. This leads to equations $(\hat{P}\hat{P}^T)\mathbf{u} = \lambda^2\mathbf{u}$ and $(\hat{P}^T\hat{P})\mathbf{v} = \lambda^2\mathbf{v}$ that determine the singular vectors $\mathbf{u}_k, \mathbf{v}_k$.

[26] Let $\mathbf{q}_k = \begin{pmatrix} \mathbf{f}_k \\ \mathbf{g}_k \end{pmatrix} = D^{-1/2}\begin{pmatrix} \mathbf{u}_k \\ \mathbf{v}_k \end{pmatrix}$. The projection matrix becomes $QQ^T = \begin{pmatrix} FF^T & FG^T \\ GF^T & GG^T \end{pmatrix}$, where $Q = (\mathbf{q}_1, \cdots, \mathbf{q}_K) = \begin{pmatrix} F \\ G \end{pmatrix}$ and $G = (\mathbf{g}_1 \cdots \mathbf{g}_K)$.

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

[28] J. P. Benzécri. *Correspondence Analysis handbook.* Marcel Dekker, 1992.

[29] One can verify that the largest singular value of $\hat{P}$ and its corresponding singular vectors are $\lambda_1 = 1$, $\mathbf{u}_1 = D_r^{1/2}\mathbf{e}_m/p_{..}^{1/2}$, $\mathbf{v}_1 = D_c^{1/2}\mathbf{e}_n/p_{..}^{1/2}$, where $p_{..} = \sum_{ij} p_{ij}$, and $\mathbf{e}_m, \mathbf{e}_m$ are vectors of ones. Therefore, Eq.(4) can be written in the familiar form $P - \mathbf{r}\mathbf{c}^T = \sum_{k=2}^{m} D_r\mathbf{f}_k\lambda_k\mathbf{g}_k^T D_c$ for correspondence analysis [27].

[30] B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 1986.

[31] T. Kohonen. *The self organizing maps. Series in Information Sciences, Vol. 30, 2nd ed.* Springer-Verlag, 1997.

[32] 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.

[33] C. Ding, X. He, H. Zha, M. Gu, and H. Simon. A min-max cut algorithm for graph partitioning and data clustering. *Proc. 1st IEEE Int'l Conf. Data Mining*, November 2001.

[34] P.K. Chan, M.Schlag, and J.Y. Zien. Spectral k-way ratio-cut partitioning and clustering. *IEEE Trans. CAD-Integrated Circuits and Systems*, 13:1088–1096, 1994.

[35] B. Hendrickson and R. Leland. An improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM J. Sci. Computing*, 16:452–469, 1995.

[36] C.J. Alpert, A.B. Kahng, and S.Z. Yao. Spectral partitioning with multiple eigenvectors. *Discrete Applied Mathematics*, 90:3–26, 1999.

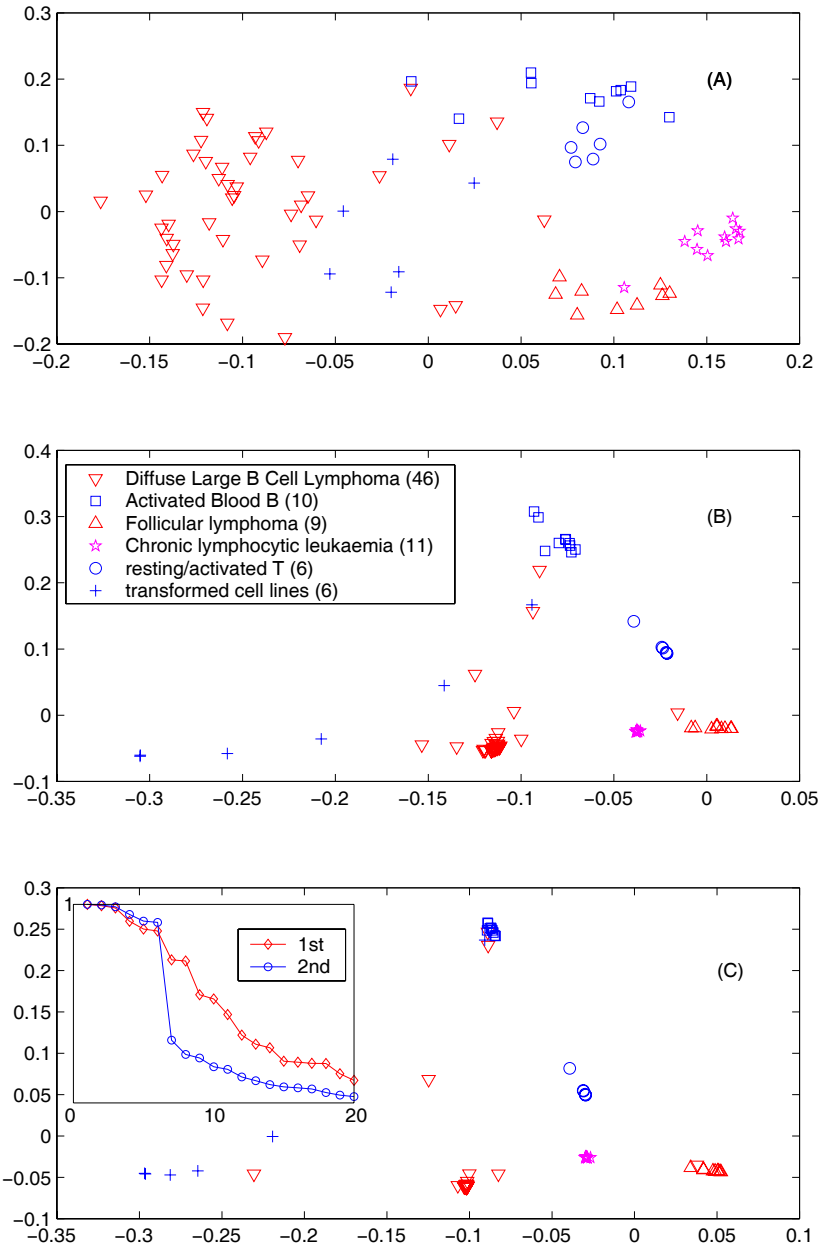

Figure 1: Gene expression profiles of cancerous and normal tissues samples from Alizadeh et al. [4] in original Euclidean space (A), in SPCA space (B), and in SPCA space after one iteration (C). Cluster structures become clearer due to self-aggregation. The insert in (C) shows the eigenvalues of the 1st and 2nd SPCA. Three cancerous and three normal subtypes are shown (with the number of samples in each subtype in parentheses). Expression levels on the 100 most informative genes define the Euclidean space; these genes are selected out of the original 4025 genes based on the F-statistic. Pearson correlation is used and similarity $S_{ij} = \exp(C_{ij}/\langle C \rangle)$, where $\langle C \rangle = 0.099$ is the average correlation. (Please view this figure in color.)

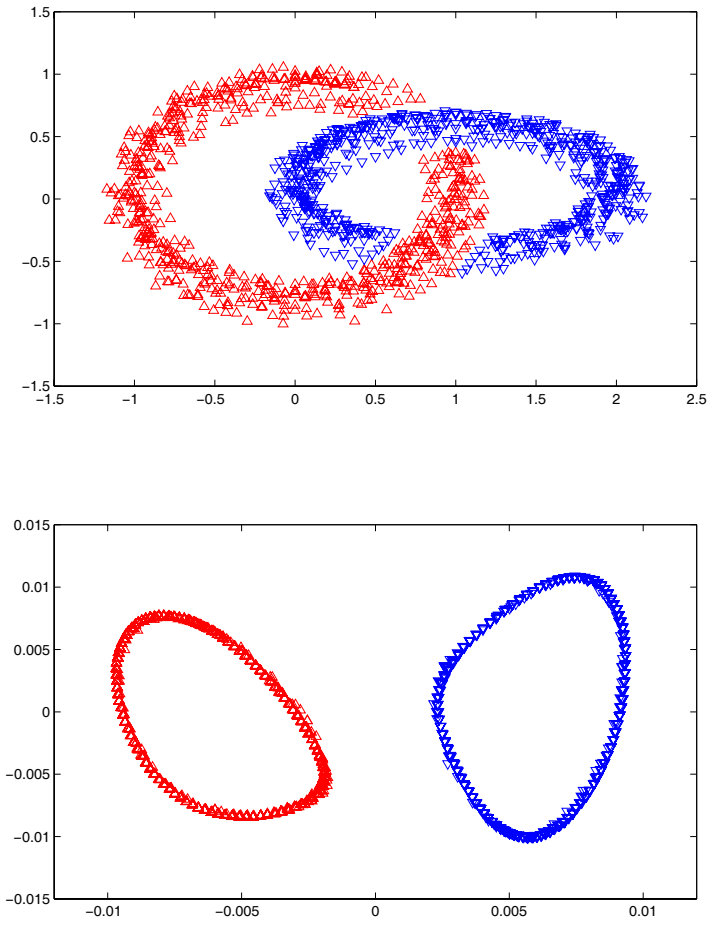

Figure 2: Data objects form two interlocking rings (but not touching each other) in 3D Euclidean space (top). In SPCA space (bottom), the rings separate. Objects self-aggregate into much thinner rings.

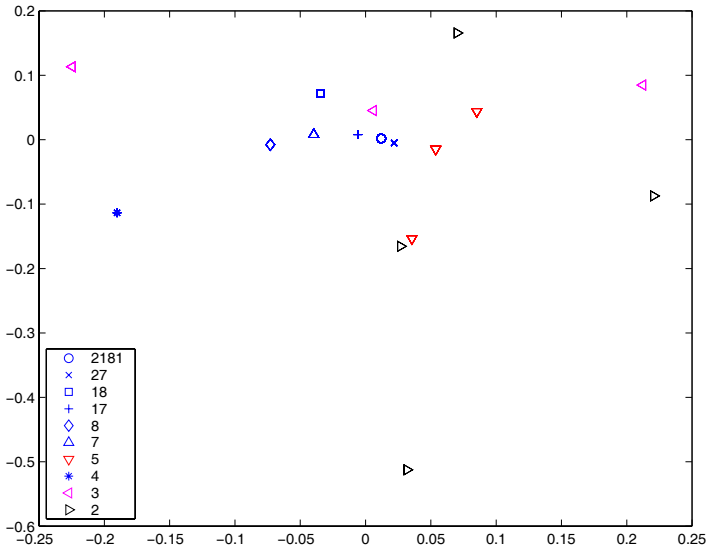

Figure 3: 2294 webpages retrieved using the query *amazon* on a search engine. They form a graph of 17 isolated segments (connected components) as shown in SPCA space (the number of webpages in each segment is indicated). Shown is the 2D-view chosen by PCA in the 17-dim SPCA space.

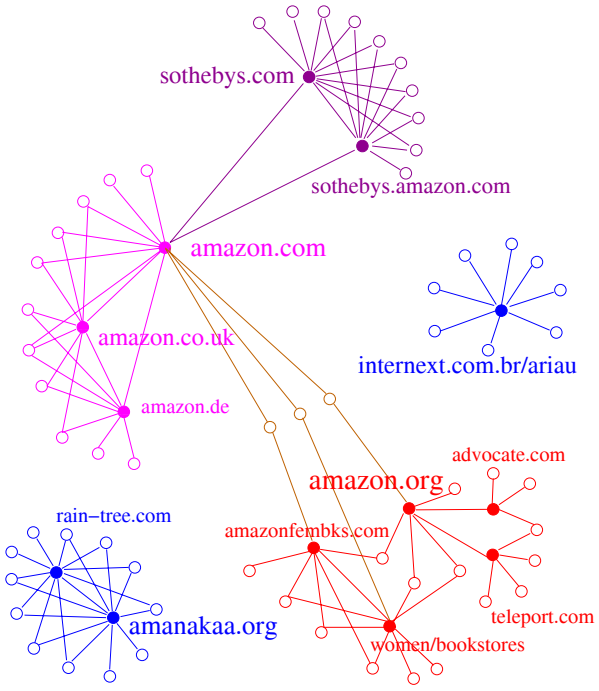

Figure 4: Connectivity of the 2181-, 27-, and 17-webpage segments. The 2181-webpage segment is further split into several sparsely connected sub-segments. Each discovered (sub)segment corresponds to a community [24].

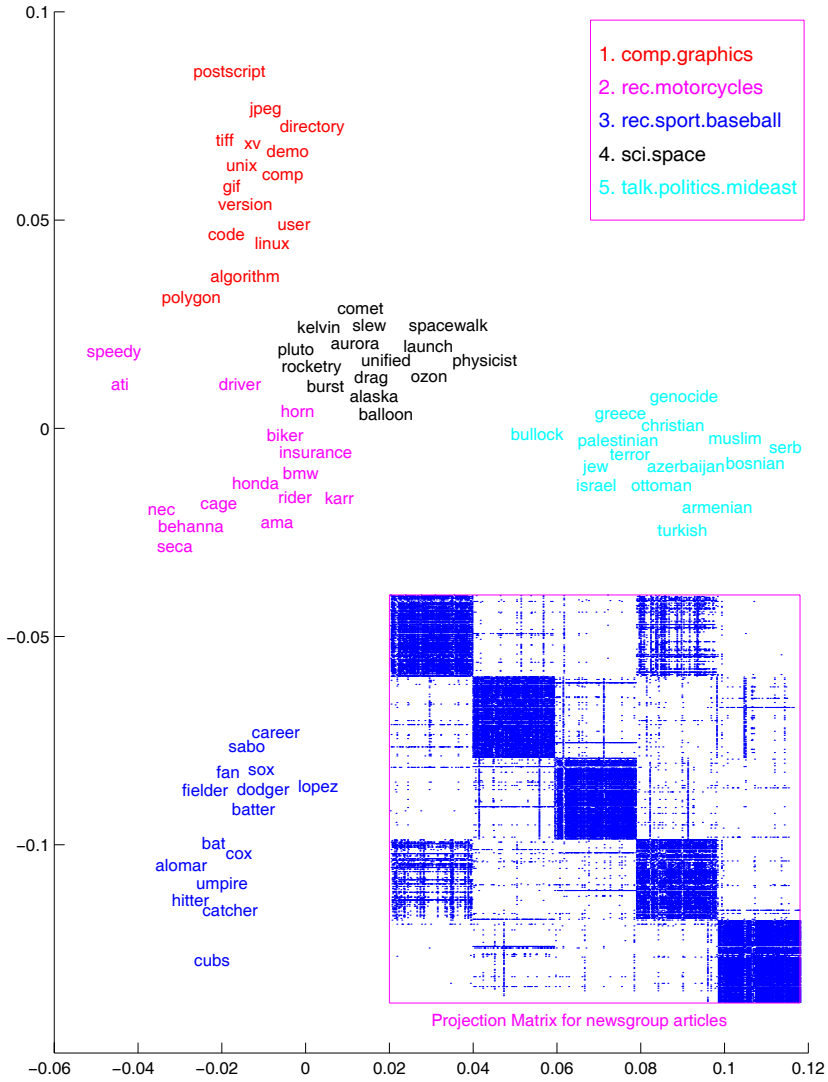

Figure 5: Word aggregation in SPCA space while articles from five internet newsgroups are simultaneously clustered. 100 news articles are randomly chosen from each newsgroup (listed in upper right corner with corresponding color). Shown are the top 15 most frequently occurring words from each discovered 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, which indicates some overlap between *computer graphics* and *space science*. The accuracy of clustering is 86%. (Please view this figure in color.)