Concept Decompositions for Large Sparse Text Data using Clustering

Inderjit Dhillon and Dharmendra Modha

Outline

- vector space model
- data sets
- concept vectors
- spherical k-means algorithm
- Experimental Results
- difference from ordinary k-means algorithm
- clustering as matrix approximation
- concept vector vs singular vectors

vector space model

2.1. PARSING AND PREPROCESSING

1. Ignoring case, extract all unique words from the entire set of documents.

2. Eliminate non-content-bearing stopwords such as a, and, the, etc. For sample lists of stopwords

3. For each document, count the number of occurrences of each word.

4. Using heuristic or information-theoretic criteria, eliminate non-content-bearing highfrequency and low-frequency words

$x_{ji} = t_{ji} \times g_j \times s_i,$

$$s_i = \left(\sum_{j=1}^{d} (t_{ji} g_j)^2 \right)^{-1/2}.$$

Data Sets

(CLASSIC3) We obtained the CLASSIC3 data set containing 3893 documents by merging the popular MEDLINE, CISI, and CRANFIELD sets.

(NSF) downloaded 13297 NSF grant abstracts between March 1996 and August 1997 in aread such as astronomy, population studies, undergraduate education, materials, mathematics, biology and health, oceanography, computer science, and chemistry.

The Spherical k-means Algorithm

- COSINE SIMILARITY $\mathbf{x}^T\mathbf{y} = \|\mathbf{x}\|\,\|\mathbf{y}\|\cos(\theta(\mathbf{x},\mathbf{y})) = \cos(\theta(\mathbf{x},\mathbf{y}))$.
- center: $\mathbf{m}_j = \frac{1}{n_j}\sum_{\mathbf{x}\in\pi_j}\mathbf{x}$,
- Concept Vector: $\mathbf{c}_j = \frac{\mathbf{m}_j}{\|\mathbf{m}_j\|}$.
- objective fcn: $Q\left(\{\pi_j\}_{j=1}^k\right) = \sum_{j=1}^{k}\sum_{\mathbf{x}\in\pi_j}\mathbf{x}^T\mathbf{c}_j$,
- Spherical k-means Algorithm:
- Start with a partitioning of the vectors into clusters π_j.
- For each cluster, compute corresonding concept vector.
- Repeat:
  - reassign each document to closest concept vector (wrt cosine similarity)
  - recompute concept vectors.
- until some stopping test satisfied.

Experimental Results

- Confusion Matrix for CLASSIC3 data set:

	$\pi_1^\dagger$	$\pi_2^\dagger$	$\pi_3^\dagger$
MEDLINE	1004	18	11
CISI	5	1440	15
CRANFIELD	4	16	1380

- NSF density

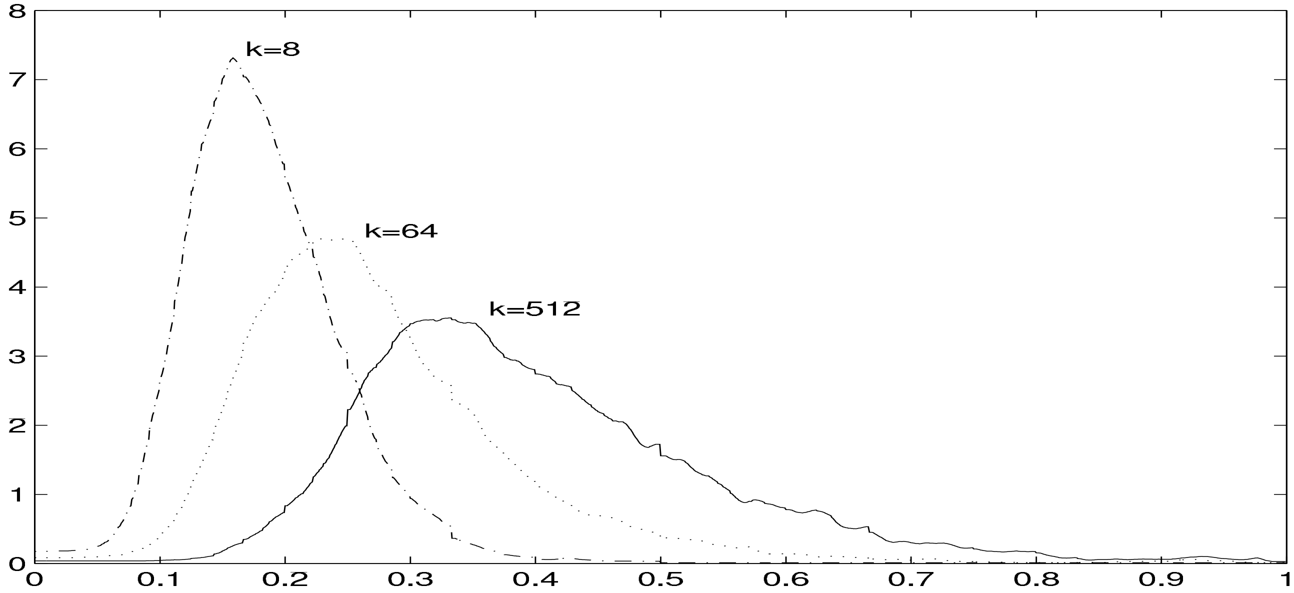


Figure 2. Average *intra*-cluster estimated probability density functions for $k = 8$, 64, and 512 clusters of the NSF data set.

- For $k = 8$, few documents are close to their concept vectors.

Compare with ordinary k-means

- Traditional objective function: $\sum_{j=1}^{k} \sum_{\mathbf{x} \in \pi_j} \|\mathbf{x} - \mathbf{m}_j\|^2$,
- Boundaries: locus of points satisfying

  $\mathbf{x}^T(\mathbf{c}_j - \mathbf{c}_\ell) = 0.$ $\rightarrow$ great circle on hypersphere.

  $\mathbf{x}^T(\mathbf{m}_j - \mathbf{m}_\ell) = \frac{1}{2}(\mathbf{m}_j^T \mathbf{m}_j - \mathbf{m}_\ell^T \mathbf{m}_\ell).$ $\rightarrow$ not a great circle
- Objective fcns related by

$$\mathcal{Q}\left(\{\pi_j\}_{j=1}^{k}\right) = \sum_{j=1}^{k} \sum_{\mathbf{x} \in \pi_j} \mathbf{x}^T \mathbf{c}_j = 2n - 2 \sum_{j=1}^{k} \sum_{\mathbf{x} \in \pi_j} \|\mathbf{x} - \mathbf{c}_j\|^2 \equiv 2n - 2\mathcal{F}\left(\{\pi_j\}_{j=1}^{k}\right).$$

Clustering as Matrix Approximation

- Concept Decomp: $\|\mathbf{X}-\hat{\mathbf{X}}_k\|_F^2 = \sum_{j=1}^{k}\sum_{\mathbf{X}\in\pi_j}\|\mathbf{x}-\mathbf{c}_j\|^2 = \mathcal{F}\left(\{\pi_j\}_{j=1}^{k}\right).$

- Truncated SVD: $\bar{\mathbf{X}}_k = \mathbf{U}_k\blacksquare_k\mathbf{V}_k^T = \mathbf{U}_k(\mathbf{U}_k^T\mathbf{U}_k)^{-1}\mathbf{U}_k^T\mathbf{X} = \mathbf{U}_k\mathbf{U}_k^T\mathbf{X}$

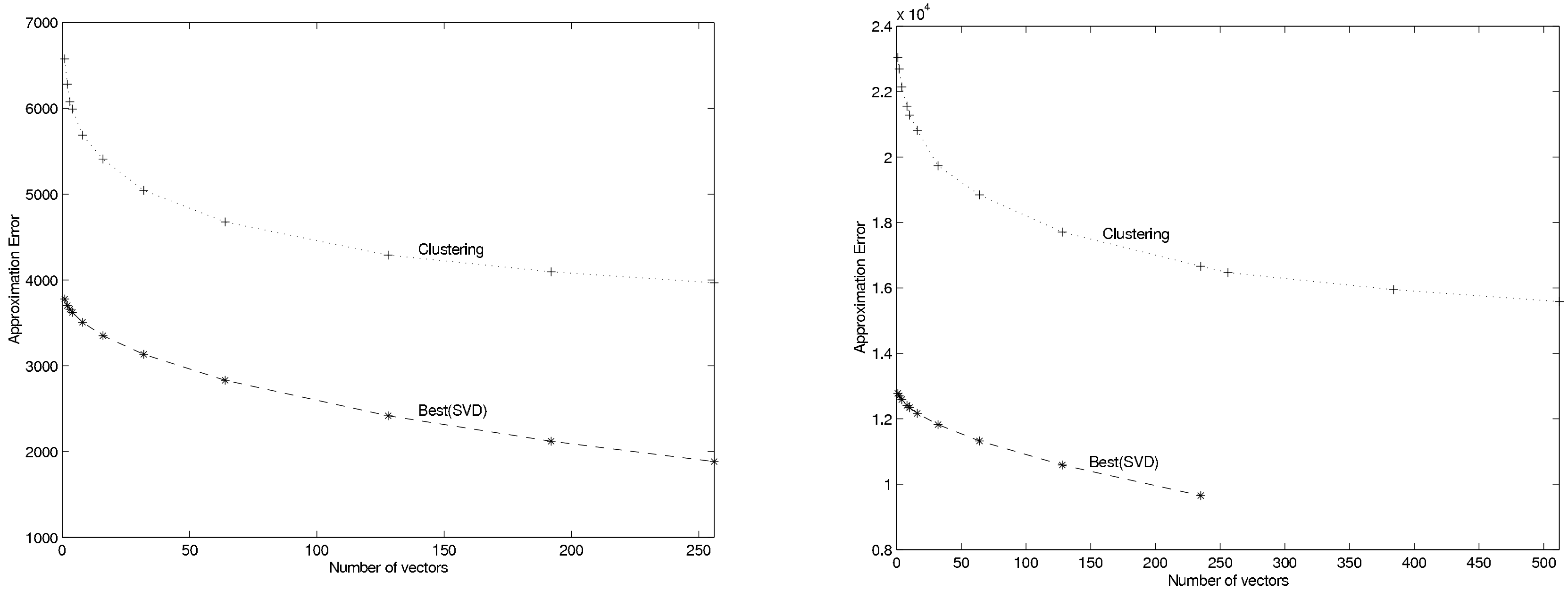


Figure 5. Comparing the approximation errors $\|\mathbf{X}-\bar{\mathbf{X}}_k\|_F^2$ and $\|\mathbf{X}-\hat{\mathbf{X}}_k^{\dagger}\|_F^2$ for the CLASSIC3 (left panel) and the NSF (right panel) data sets for various values of k.

Concept Decomposition

- $\mathbf{X} \approx \tilde{\mathbf{X}}_k = \mathbf{C}_k \mathbf{Z}^* = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_k] \cdot (\mathbf{C}_k^T \mathbf{C}_k)^{-1} \mathbf{C}_k^T \mathbf{X}.$
- $\mathbf{Z}^* = \arg\min_{\mathbf{Z}} \|\mathbf{X} - \mathbf{C}_k \mathbf{Z}\|_F.$

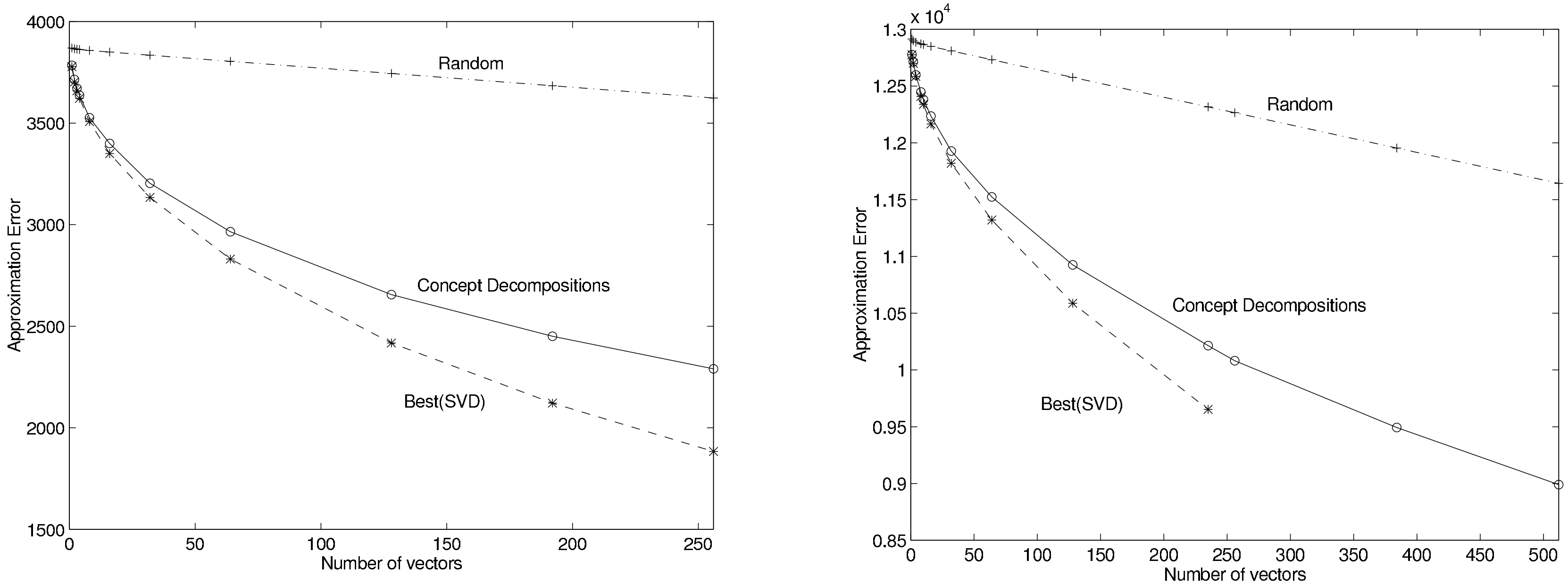
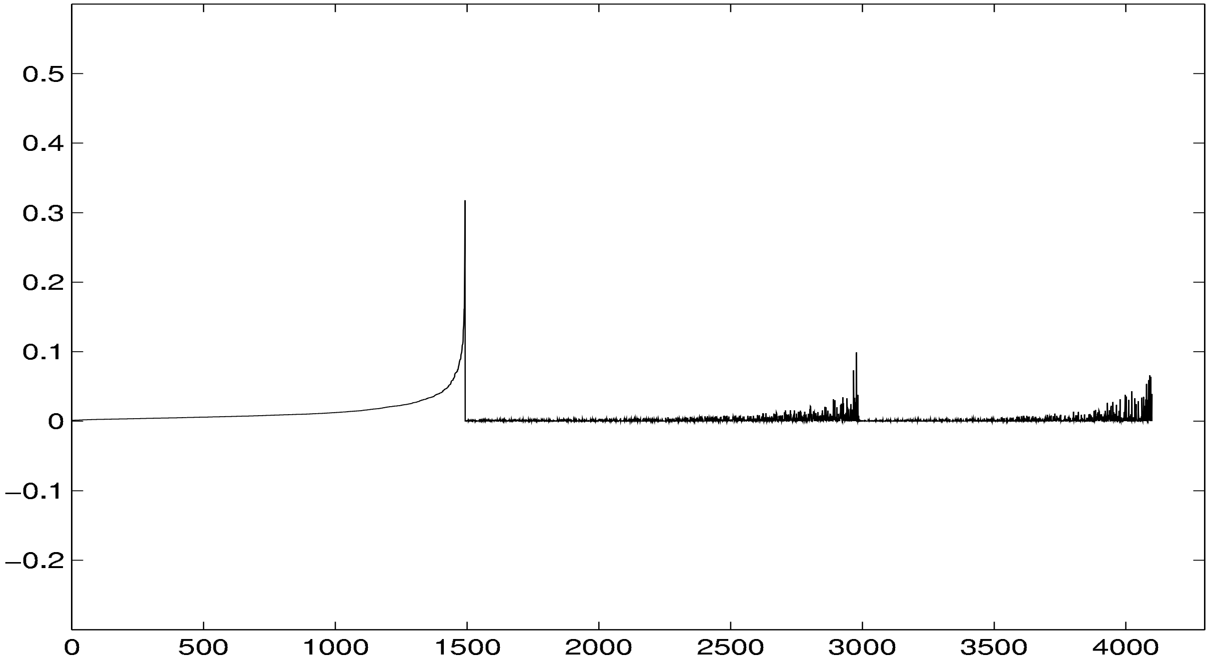


Figure 6. Comparing the approximation errors $\|\mathbf{X}-\bar{\mathbf{X}}_k\|_F^2$, $\|\mathbf{X}-\tilde{\mathbf{X}}_k^\dagger\|_F^2$, and $\|\mathbf{X}-\check{\mathbf{X}}_k\|_F^2$ for the CLASSIC3 (left panel) and the NSF (right panel) data sets for various values of k.

Concept vectors vs Singular vectors

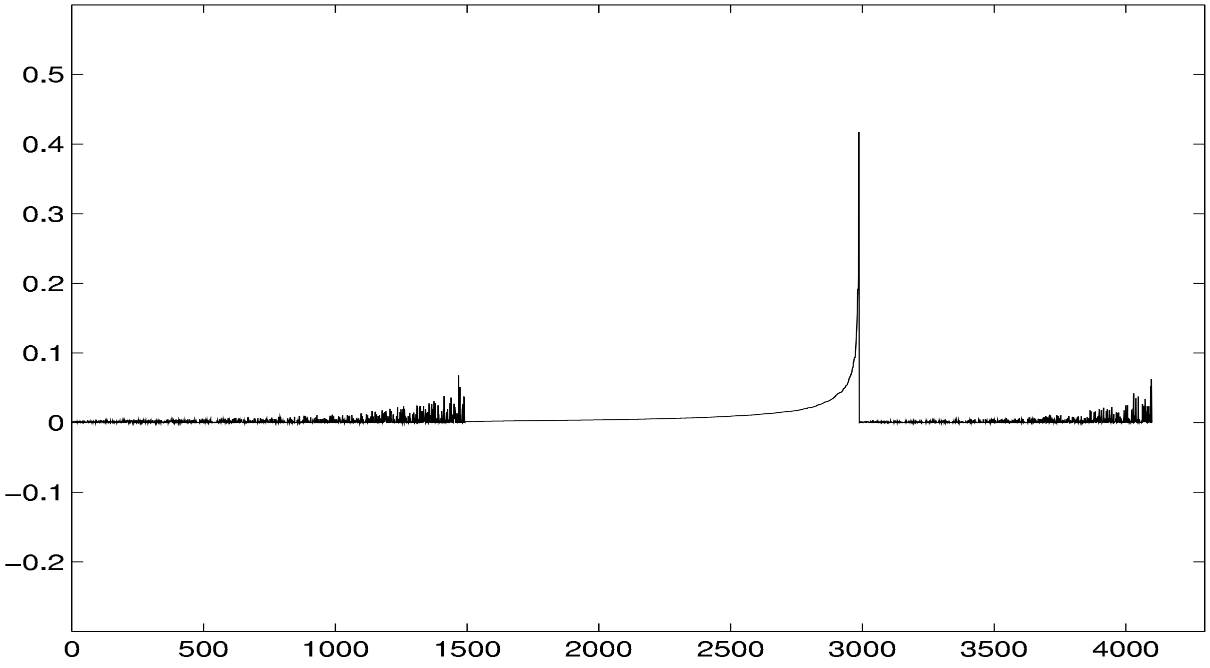
- Concept vectors are
  - sparse
  - non-negative
  - local – representing only a few words &/or documents.
  - Almost orthonormal anyway.
  - small principal angles to singular vectors.

First Concept Vector – CLASSIC3



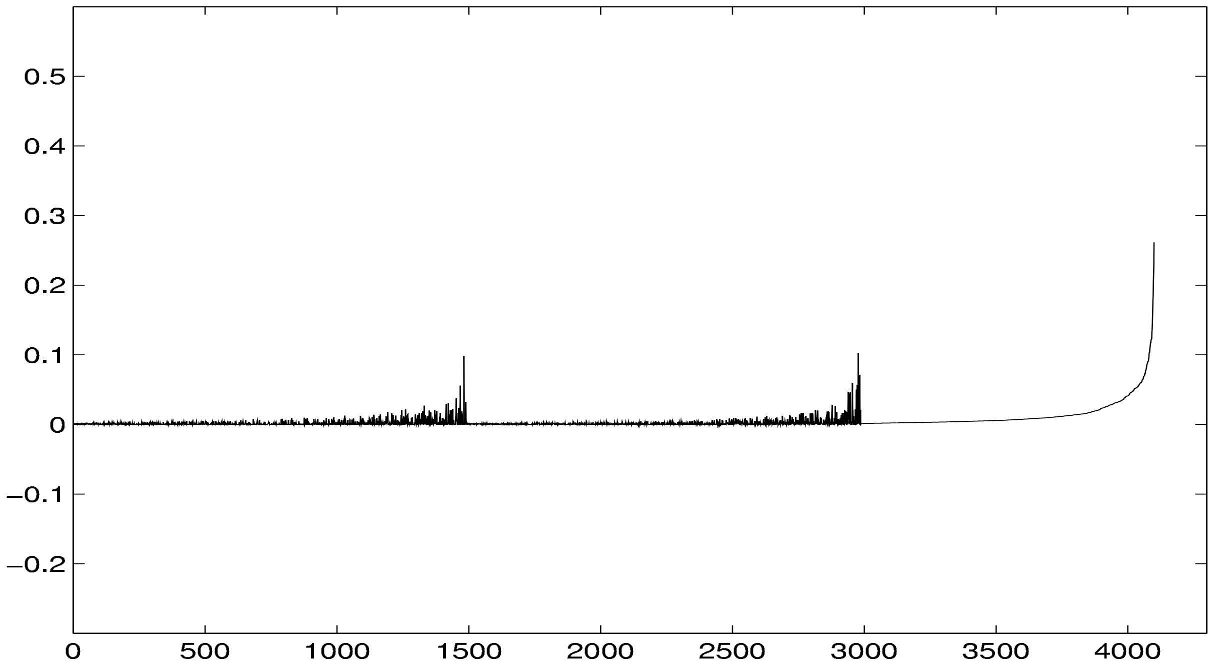
patients	0.3178
cells	0.2429
treatment	0.1617
normal	0.1614
growth	0.1489
blood	0.1376
children	0.1310
disease	0.1151
cell	0.1120
human	0.1104

Second Concept Vector – CLASSIC3



library	0.4171
research	0.2108
libraries	0.2009
retrieval	0.1921
data	0.1918
systems	0.1794
science	0.1637
scientific	0.1524
book	0.1345
study	0.1291

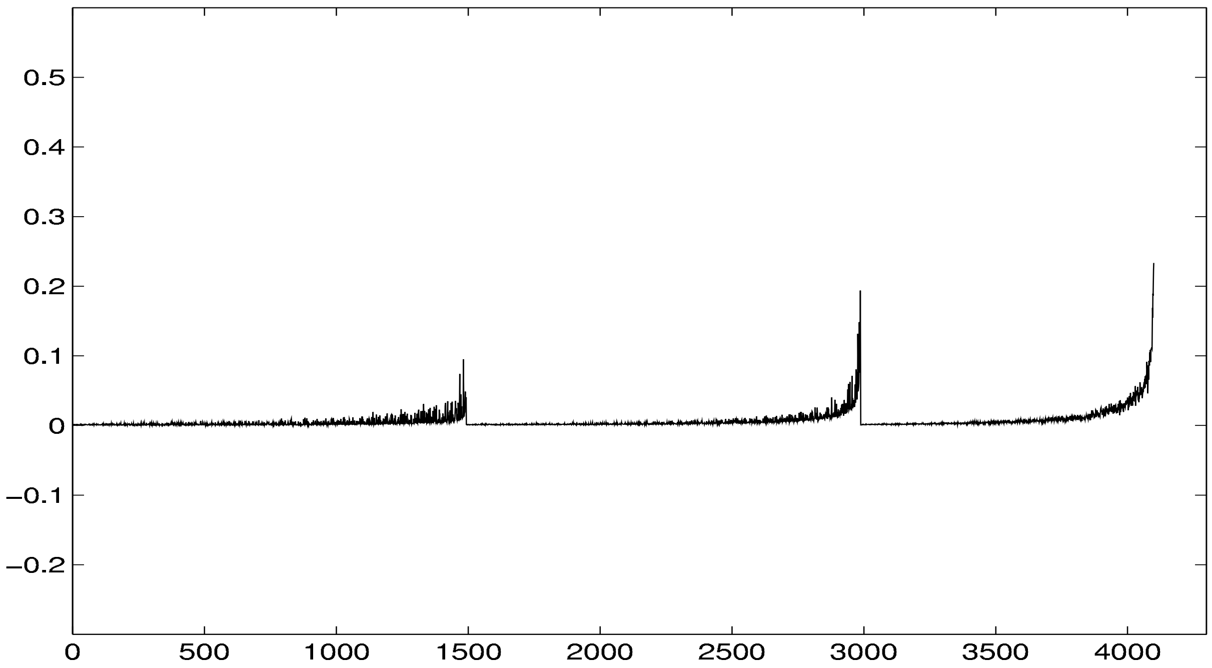
Third Concept Vector – CLASSIC3



pressure	0.2616
boundary	0.2263
theory	0.2046
layer	0.1846
mach	0.1790
method	0.1651
supersonic	0.1426
heat	0.1361
shock	0.1325
effects	0.1228

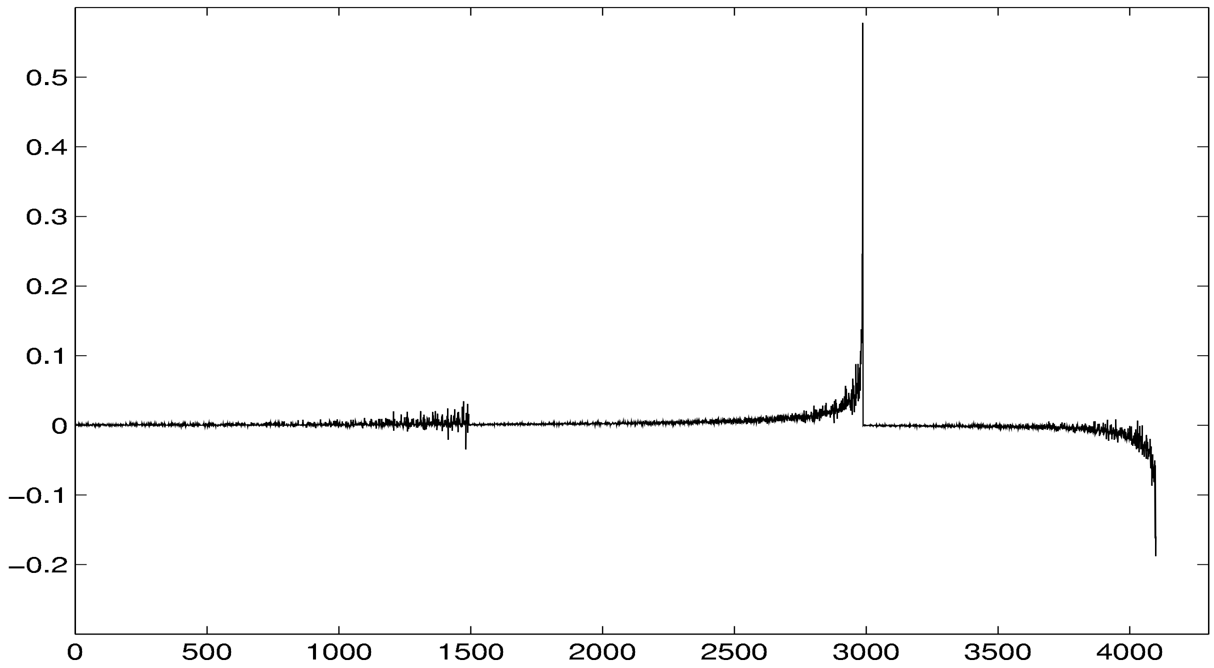
Figure 7. The three concept vectors corresponding to a clustering of the CLASSIC3 data set into 3 clusters. For each concept vector, the top ten words with the corresponding weights are shown on the right.

First Singular Vector – CLASSIC3



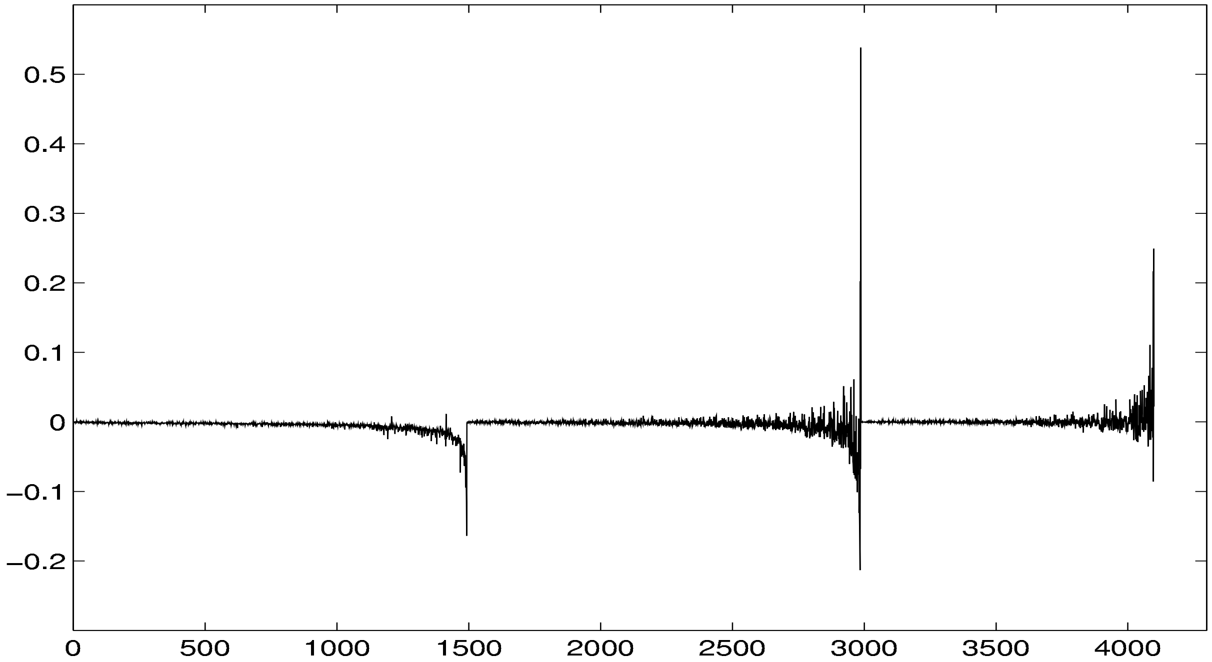
pressure	+0.2335
boundary	+0.2234
library	+0.1938
layer	+0.1894
theory	+0.1862
method	+0.1689
mach	+0.1548
data	+0.1484
analysis	+0.1318
shock	+0.1117

Second Singular Vector – CLASSIC3



library	+0.5780
libraries	+0.2464
boundary	−0.1879
research	+0.1730
layer	−0.1620
pressure	−0.1597
systems	+0.1379
data	+0.1200
retrieval	+0.1180
science	+0.1144

Third Singular Vector – CLASSIC3



library	+0.5384
boundary	+0.2491
layer	+0.2163
retrieval	−0.2132
libraries	+0.2023
patients	−0.1638
data	−0.1534
scientific	−0.1307
systems	−0.1294
cells	−0.1152

Figure 8. The three leading singular vectors for the CLASSIC3 word-by-document matrix. For each singular vector, the top ten words with the corresponding weights are shown on the right.

Conclusions

- Spherical k-means algorithm achieves high-coherence clusters
- average intra vs inter –cluster coherence structure
  preserved as one refines individual clusters.
- Concept Decomposition: approximation & orthogonality
  approaces that of the SVD
- Concept vectors: large entries $\Longleftrightarrow$ word clusters.
  Also, sparse and non-negative.