Iterative Extraction (ITEX SEFIT (1990)): Extensions of Principal Component Analysis

Boris Mirkin

School of Computer Science
Birkbeck College
University of London

Special Issue: The Computer Journal, Profiling Expertise and Behaviour:
Deadline 15 Nov. 2006
WHAT IS CLUSTERING; WHAT IS DATA


WARD HIERARCHICAL CLUSTERING: Agglomerative; Divisive Clustering with Ward Criterion; Extensions of Ward Clustering

DATA RECOVERY MODELS: Statistics Modelling as Data Recovery; Data Recovery Model for K-Means; for Ward; Extensions to Other Data Types; One-by-One Clustering

DIFFERENT CLUSTERING APPROACHES: Extensions of K-Means; Graph-Theoretic Approaches; Conceptual Description of Clusters

GENERAL ISSUES: Feature Selection and Extraction; Similarity on Subsets and Partitions; Validity and Reliability
Talk’s outline

- Data model and Pythagorean decomposition
- Principal component analysis as a data model
- Extension of PCA to clustering and K-Means
- Principal cluster analysis for clustering
- General ITEX strategy
- Examples of ITEX: hierarchical clustering, additive clustering, box clustering, contingency data aggregation
Pythagorean framework for data analysis methods

Type of Data
- Similarity
- Temporal
- Entity-to-feature
- Co-occurrence

Type of Model
- Regression
- Principal components
- Clusters

Model:
Data = Model_Data + Residual

Pythagoras:
Data^2 = Model_Data^2 + Residual^2
Pearson’s PCA: measuring talent

**Given:** marks $x_{iv}$ (i – student, v – subject)

**Find:** talent score $z_i$ and subject loading $c_v$

\[ x_{iv} = c_v z_i + e_{iv} \]

\[ L^2 = \sum_{i \in I} \sum_{v \in V} e_{iv}^2 = \sum_{i \in I} \sum_{v \in V} (x_{iv} - c_v z_i)^2 \]

**Solution:** \( X^T z^* = \mu c^* \), \( Xc^* = \mu z^* \), \( \max \mu \)

**Properties:**

**P1:** $z^*$ is lc of $X$ columns

**P2:** \( T(X) = \mu^2 + L^2 \), \( T(X) = \sum x_{iv}^2 \) – data scatter
PCA as a data model

Data Model:

\[ y_{iv} = \sum_{k=1}^{K} c_{kv}z_{ik} + e_{iv}, \]

minimising \( L^2 \) over \( c \) and \( z \)

Properties:

\[ [Z, M, C] = \text{svd}(Y), \]

\[ \text{Thus } z \text{ and } c \text{ are lc of } X \]

\[ \text{Can be done sequentially, one by one} \]

\[ T(Y) = \mu_1^2 + \mu_2^2 + \ldots + \mu_K^2 + L^2 \]
Extension of PCA to clustering

\[ y_{iv} = \sum_{k=1}^{K} c_{kv} z_{ik} + \varepsilon_{iv}, \]

\[ \sum_{i=1}^{N} \sum_{v=1}^{V} y_{iv}^2 = \sum_{v=1}^{V} \sum_{k=1}^{K} c_{kv}^2 N_k + \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v=1}^{V} (y_{iv} - c_{kv})^2 \]

- \( y \) – data entry
- \( z \) – 1/0 membership
- \( c \) – cluster centroid
- \( N \) – cardinality
- \( i \) – entity
- \( v \) – feature /category
- \( k \) – cluster
Representing a partition

Cluster $k$:

Centroid

$C_{kv}$ (ν - feature)

Binary 1/0 membership

$Z_{ik}$ (i - entity)
Standardisation of features

\[ Y_{ik} = \frac{(X_{ik} - A_k)}{B_k} \]

- **X** - original data
- **Y** - standardised data
- **i** - entities
- **k** - features
- **A_k** - shift of the origin, typically, the average
- **B_k** - rescaling factor, traditionally the standard deviation, but range seems better in clustering
No standardisation
Z-scoring (scaling by std)
Standardising by range & weight
Fitting the model with Straight K-Means Partitioning

Start:
* Presenting cases as multidimensional points
* Putting initial centroids (seeds)

Reiterated until no change:
* Collecting points into clusters around centroids
* Recalculating centroids as cluster prototypes
Advantages of K-Means

- **Conventional:**
  - Models typology building
  - Computationally effective
  - Can be incremental, `on-line`

- **Unconventional:**
  - Associates feature salience with feature scales and correlation/association
  - Applicable to mixed scale data
Drawbacks of K-Means

• No advice on:
  • Data pre-processing
  • Number of clusters
  • Initial setting
• Instability of results
• Criterion can be inadequate
• Insufficient interpretation aids
Initial Centroids: Correct

Two cluster case
Initial Centroids: Correct
Different Initial Centroids:
Wrong, even though in different clusters
Principal Cluster Analysis: 
One cluster at a time

\[ y_{iv} = c_v z_i + e_{iv}, \]

where \( z_i = 1 \) if \( i \in S \), \( z_i = 0 \) if \( i \not\in S \)

With Euclidean distance squared

\[
\sum_{i=1}^{N} \sum_{v=1}^{V} y_{iv}^2 = \sum_{v=1}^{V} c_{Sv}^2 N_S + \sum_{i \in S} \sum_{v=1}^{V} (y_{iv} - c_{Sv})^2
\]

\[
\sum_{i=1}^{N} d(i,0) = d(c_S,0) N_S + \sum_{i \in S} d(i, c_S)
\]

\( c_S \) must be anomalous, that is, interesting
Principal Cluster Analysis:
One cluster at a time

\[ \sum_{i=1}^{N} \sum_{v=1}^{V} y_{iv}^2 = \sum_{v=1}^{V} c_{sv}^2 N_S + \sum_{i \in S} \sum_{v=1}^{V} (y_{iv} - c_{sv})^2 \]

\[ \sum_{i=1}^{N} d(i,0) = d(c_s,0) N_S + \sum_{i \in S} d(i,c_s) \]

Or, with

Euclidean distance squared \( d(, ,) \)
Initial setting with Anomalous Pattern (AP) clustering
AP clustering: Iterate

- DombeySon
- OliverTwist
- GreatExpectations
- TomSawyer
- Yankee
- HuckFinn
- WarPeace
- AnnaKarenin
iK-Means with Anomalous Single Clusters

1. AnnaKarenin
   - WarPeace

2. DombeySon
   - OliverTwist
   - GreatExpectations

3. TomSawyer
   - HuckFinn
   - Yankee
Decomposing Data scatter

The sum of standardised entries squared

\[ D^2 = \sum_{i=1}^{N} \sum_{v=1}^{V} y_{iv}^2 \]

The sum of contributions of features

Proportional to the summary variance
Contribution of a feature $F$ to a partition

$Contrib(F) = \sum_{v \in F} \sum_{k=1}^{K} c_{kv}^2 N_k$

Proportional to

- correlation ratio $\eta^2$ if $F$ is quantitative
- a contingency coefficient if $F$ is nominal
  - Pearson chi-square (Poisson normalised)
  - Goodman-Kruskal tau-b (Range normalised)
Contribution of a quantitative feature to a partition

\[ N \eta^2 = N \sum_{k=1}^{K} \left( \sigma^2 - p_k \sigma_k^2 \right) / \sigma^2 \]

Proportional to

- correlation ratio \( \eta^2 \) if F is quantitative
Contribution of a nominal feature to a partition

\[ N X^2 = N \sum_{k=1}^{K} \left( p_{ij} - p_i p_j \right)^2 / p_i B_j^2 \]

Proportional to a contingency coefficient

- Pearson chi-square (Poisson normalised)
  \[ B_j = \sqrt{p_j} \]

- Goodman-Kruskal tau-b (Range normalised)
Pythagorean Decomposition of data scatter for interpretation

Praised as in Table 2.0 according to author based clusters.

<table>
<thead>
<tr>
<th>Title</th>
<th>LenS</th>
<th>LenD</th>
<th>NChar</th>
<th>FCon</th>
<th>Pers</th>
<th>Obje</th>
<th>Dire</th>
<th>Cntr</th>
<th>Cntr,%</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTwist</td>
<td>-0.18</td>
<td>0.29</td>
<td>0.29</td>
<td>1.46</td>
<td>0.23</td>
<td>0.02</td>
<td>0.10</td>
<td>2.21</td>
<td>6.31</td>
</tr>
<tr>
<td>DombyS</td>
<td>0.36</td>
<td>0.06</td>
<td>0</td>
<td>1.46</td>
<td>0.23</td>
<td>0.02</td>
<td>0.10</td>
<td>2.22</td>
<td>6.34</td>
</tr>
<tr>
<td>GExpectations</td>
<td>0.08</td>
<td>0.12</td>
<td>0</td>
<td>1.46</td>
<td>-0.14</td>
<td>-0.03</td>
<td>0.10</td>
<td>1.58</td>
<td>4.51</td>
</tr>
</tbody>
</table>

| Cl. 1 Cntr       | 0.26  | 0.47  | 0.29  | 4.38 | 0.32 | 0.01 | 0.29 | 6.01 | 17.17  |
| TomSoyer         | 0.48  | 0.44  | 0.58  | 0.52 | -0.03| -0.14| 0.10 | 1.95 | 5.57   |
| HuckFinn         | -0.38 | 0.83  | 0     | 0.52 | 0.02 | 0.23 | 0.10 | 1.32 | 3.77   |
| YankeeA          | 1.22  | 1.21  | 0.58  | 0.52 | 0.02 | 0.23 | 0.10 | 3.88 | 11.09  |

| Cl. 2 Cntr       | 1.31  | 2.48  | 1.17  | 1.58 | 0.01 | 0.32 | 0.29 | 7.15 | 20.43  |
| WarPeace         | 0.14  | -0.23 | 1.31  | 0.52 | 0.18 | 0.18 | 0.88 | 2.97 | 8.49   |
| Akarenina        | 0.47  | 1.42  | 2.62  | 0.52 | 0.18 | 0.18 | 0.88 | 6.26 | 17.89  |

| Cl. 3 Cntr       | 0.61  | 1.19  | 3.94  | 1.05 | 0.35 | 0.35 | 1.75 | 9.23 | 26.37  |

| Explained        | 2.18  | 4.14  | 5.40  | 7.00 | 0.67 | 0.67 | 2.33 | 22.39| 63.97  |
| Unexplained      | 4.82  | 2.86  | 1.60  | 0    | 1.66 | 1.67 | 0.00 | 12.61| 36.03  |
| Total            | 7.00  | 7.00  | 7.00  | 7.00 | 2.33 | 2.33 | 2.33 | 35.00| 100.00 |
Contribution based description of clusters

- C. Dickens: \( FCon = 0 \)
- M. Twain: \( LenD < 28 \)
- L. Tolstoy: \( \text{NumCh} > 3 \) or \( Direct = 1 \)
Simulation study of Number-of clusters methods (joint work with Mark Chiang):

- Variance based:
  - Hartigan (HK)
  - Calinski & Harabasz (CH)
  - Jump Statistic (JS)

- Structure based:
  - Silhouette Width (SW)

- Consensus based:
  - Consensus Distribution area (CD)
  - Consensus Distribution mean (DD)

- Sequential extraction of APs:
  - Least Square (LS)
  - Least Moduli (LM)
Data generation for the experiment

- Gaussian Mixture (6, 7, 9 clusters) with:
  - Cluster spatial size:
    - Constant (spherical)
    - k-proportional
    - $k^2$-proportional
  - Cluster spread (distance between centroids)

<table>
<thead>
<tr>
<th>Spread</th>
<th>Spherical</th>
<th>PPCA model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>k-proport.</td>
</tr>
<tr>
<td>Large</td>
<td>2 (1)</td>
<td>10 (2)</td>
</tr>
<tr>
<td>Small</td>
<td>0.2 (4)</td>
<td>0.5 (5)</td>
</tr>
</tbody>
</table>
Evaluation of results:
Estimated clustering versus that generated

- Number of clusters
- Distance between centroids
- Similarity between partitions
Distance between estimated centroids (o) and those generated (o )
Distance between estimated centroids (\(o\)) and those generated (\(o'\))

Final Assignment

\[G_1(p_1)\]

\[G_2(p_2)\]

\[G_3(p_3)\]

\[e_1(q_1)\]

\[e_2(q_2)\]

\[e_3(q_3)\]

\[e_4(q_4)\]

\[e_5(q_5)\]

\[g_1------e_2, e_1\]

\[g_2------e_4, e_3\]

\[g_3------e_5\]
Distance between centroids: quadratic and city-block

1. Assignment

\[ g_1(p_1)-----e_1(q_1), \ e_2(q_2) \]
\[ g_2(p_2)-----e_3(q_3), \ e_4(q_4) \]
\[ g_3(p_3)-----e_5(q_5) \]

2. Distancing

\[ d_1 = \frac{q_1 \cdot d(g_1,e_1) + q_2 \cdot d(g_1,e_2)}{q_1 + q_2} \]
\[ d_2 = \frac{q_3 \cdot d(g_2,e_3) + q_4 \cdot d(g_2,e_4)}{q_3 + q_4} \]
\[ d_3 = \frac{q_5 \cdot d(g_3,e_5)}{q_5} \]
Distance between centroids: quadratic and city-block

\[ p_1d_1 + p_2d_2 + p_3d_3 \]

1. Assignment
2. Distancing
3. Averaging
Similarity between partitions according to their confusion table

- Relative distance (Mirkin-Cherny 1970)
- Tchouprov coefficient (Cramer 1943)
- Adjusted Rand Index (Arabie-Hubert, 1985)
- Average Overlap (Mirkin 2005)
Results
at 9 clusters, 1000 entities, 20 features generated

<table>
<thead>
<tr>
<th></th>
<th>Estimated number of clusters</th>
<th>Distance between Centroids</th>
<th>Adjust Rand Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Large spread</td>
<td>Small spread</td>
<td>Large spread</td>
</tr>
<tr>
<td>HK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SW</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Extending PCA to ITEX

Iterative Extraction Elements:

- Data X format: at PCA, entity-to-feature
- Structure to extract; at t-th step set D(t): at PCA, a pair z and c;
- Criterion to minimise, $\Phi(\varepsilon)$: at PCA, $L2$
- Relation between D(t) and D(t+1): at PCA, same
- Method for minimising, at step t, $\Phi(|X(t) - s|)$ over $s \in D(t)$ where $X(t) = X(t-1) - s(t-1)$, $X(0) = X$: at PCA, svd or AP clustering

Result: $X = \sum_t s(t) + \varepsilon$, along with Pythagorean decomposition of T(X)

Proof of (finite) convergence (Mirkin (1990, 1998))
ITEX examples:

- Hierarchical clustering for conventional and spatial data
- Similarity clustering with additive clustering
- Similarity clustering with boxes (“plaid clustering”)
- Contingency data clustering and aggregation
Hierarchical clustering for conventional and spatial data

- **Model**: Same

- **Cluster structure**: 3-valued $z$’s

- **A split** $S = S_1 + S_2$ of a node $S$ in children $S_1,$ $S_2$:

  $z_i = 0$ if $i \not\in S,$  
  $= a$ if $i \in S_1$  
  $= -b$ if $i \in S_2$

If $a$ and $b$ taken to $z$ being centred, the node vectors for a hierarchy form **orthogonal base** (an analogue to SVD)

$$y_{iv} = \sum_{k=1}^{K} c_{kv} z_{ik} + e_{iv},$$
Figure 7. Compression and decompression of the boxed data with hierarchies A and B from Figure 6.
Similarity additive (and hierarchical) clustering

Observed similarity matrix

\[ B = \lambda_1 z_1 z_1^T + \lambda_2 z_2 z_2^T + \lambda_K z_K z_K^T + E \]

Problem: given \( B \), find \( \lambda \)s and \( z \)s to minimize \( E \), the differences between \( B \) and summary clusters

\[ ||E||^2 \Rightarrow \min_A \]
Additive clusters: ITEX

Doubly greedy strategy

**OUTER LOOP:** One cluster at a time

Find real $\lambda$ (intensity) and binary $z$ (membership) to minimize $L(B, \lambda, z)$.

Update $B \leftarrow B - \lambda z z^T$; and reiterate!

After $K$ iterations, clusters $S_k$ of cardinality $N_k$.

$$T(B) = \lambda_1^2 N_1^2 + \lambda_2^2 N_2^2 + \ldots + \lambda_K^2 N_K^2 + L^2$$

**INNER LOOP:** maximise $\lambda_k N_k$
Algorithm: ADDI-S (Mirkin JoC 1987), a data approximation technique

To maximize Contribution to Data Scatter,
Average within-cluster similarity $\lambda$ multiplied by the cluster’s size $\#S$

Algorithm ADDI-S:
- Take $S=\{ j \}$ for arbitrary $j$
- Given $S$, find $\lambda = c(S)$ and similarities $b(i,S)$ to $S$ for all entities $i$ in and out of $S$;
- Check the differences $b(i,S) - \lambda / 2$. If they are consistent, change the state of a most contributing entity. Else, stop and output $S$.

Resulting $S$: a tightness property.

Algorithm: ADDI-S a data approximation techniques

Number of clusters: Depends on similarity shift threshold $b$

$$b(ij) \leftarrow b(ij) - b$$
Domain knowledge: Function is known at some HPFs

- 287 pairs of HPFs with known function of which 86 are SYNONYMOUS (same function)

Two values:
- Min error
- No non-synonymous
Hierarchical similarity clusters

Spectral clustering
Similarity clustering with boxes

Plaid clustering
Contingency data clustering and aggregation

- $P(I,J) = (p_{ij})$ non-negative and summable

- Correspondence Analysis rather than PCA

- Quetelet coefficients rather than $p_{ij}$

\[
q_{ij} = \frac{p_{ij}}{(p_{i+} + p_{j+}) - 1} = \frac{[p(i/j) - p(i)]}{p(i)}
\]

Let A partitions I and B partitions J: $P(A,B)$ by summing up $p_{ij}$ within blocks to approximate $q_{ab}$ by the $p_{i+} p_{j+}$ weighted least-squares $L^2$:

- Pythagorean

\[
X^2(I,J) = X^2(A,B) + L^2
\]
Conclusion

Looking forward to hear of further ideas for combining clustering and visualisation à la PCA