Lecture 2: Data Analytics of Narrative

Data Analytics of Narrative: Pattern Recognition in Text, and Text Synthesis, Supported by the Correspondence Analysis Platform. This Lecture is presented in three parts, as follows.

**Part 1** Data analytics of narrative.


**Part 3** Ultrametric embedding.
Lecture 2: Data Analytics of Narrative

Data Analytics of Narrative: Pattern Recognition in Text, and Text Synthesis, Supported by the Correspondence Analysis Platform.

1. A short review of the theory and practical implications of Correspondence Analysis.
4. Towards semantic rating.
“We call distribution of a word the set of all its possible environments” (Z.S. Harris)
“We call distribution of a word the set of all its possible environments” (Z.S. Harris)

Initially, correspondence analysis was proposed as an inductive method for analyzing linguistic data.
“We call distribution of a word the set of all its possible environments” (Z.S. Harris)

Initially, correspondence analysis was proposed as an inductive method for analyzing linguistic data.

“We call distribution of a word the set of all its possible environments” (Z.S. Harris)

Initially, correspondence analysis was proposed as an inductive method for analyzing linguistic data.


“The model should follow the data, not the reverse!” (In J.P. Benzécri, “Statistical analysis as a tool to make patterns emerge from data”, in Methodologies of Pattern Recognition, Ed. Watanable, NY: Academic, 1969.)
”We call distribution of a word the set of all its possible environments” (Z.S. Harris)

Initially, correspondence analysis was proposed as an inductive method for analyzing linguistic data.


”The model should follow the data, not the reverse!” (In J.P. Benzécri, “Statistical analysis as a tool to make patterns emerge from data”, in Methodologies of Pattern Recognition, Ed. Watanable, NY: Academic, 1969.)

So: Description first – priority. Inductive philosophy.
The starting point is a matrix that cross-tabulates the dependencies, e.g. frequencies of joint occurrence, of an observations crossed by attributes matrix.
The starting point is a matrix that cross-tabulates the dependencies, e.g. frequencies of joint occurrence, of an observations crossed by attributes matrix.

By endowing the cross-tabulation matrix with the $\chi^2$ metric on both observation set (rows) and attribute set (columns), we can map observations and attributes into the same space, endowed with the Euclidean metric.
Analysis Chain

- The starting point is a matrix that cross-tabulates the dependencies, e.g. frequencies of joint occurrence, of an observations crossed by attributes matrix.

- By endowing the cross-tabulation matrix with the \( \chi^2 \) metric on both observation set (rows) and attribute set (columns), we can map observations and attributes into the same space, endowed with the Euclidean metric.

- A hierarchical clustering is induced on the Euclidean space, the factor space.
Analysis Chain

- The starting point is a matrix that cross-tabulates the dependencies, e.g. frequencies of joint occurrence, of an observations crossed by attributes matrix.

- By endowing the cross-tabulation matrix with the $\chi^2$ metric on both observation set (rows) and attribute set (columns), we can map observations and attributes into the same space, endowed with the Euclidean metric.

- A hierarchical clustering is induced on the Euclidean space, the factor space.

- Interpretation is through projections of observations, attributes or clusters onto factors. The factors are ordered by decreasing importance.
Correspondence Analysis: Mapping $\chi^2$ Distances into Euclidean Distances

The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i, j) = k(i, j); i \in I, j \in J\}$. 
Correspondence Analysis: Mapping $\chi^2$ Distances into Euclidean Distances

- The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i, j) = k(i, j); i \in I, j \in J\}$.

- $I$ is the set of observation indexes, and $J$ is the set of attribute indexes.

- What we have described here is taking numbers of occurrences into relative frequencies.

- The conditional distribution of $f_J$ knowing $i \in I$, also termed the $j$th profile with coordinates indexed by the elements of $I$, is:
  $$f_iJ = \{f_{ij} = \frac{k_{ij}}{k_i}; f_i > 0; j \in J\}$$ and likewise for $f_J$.
Correspondence Analysis: Mapping $\chi^2$ Distances into Euclidean Distances

- The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i, j) = k(i, j); i \in I, j \in J\}$.

- $I$ is the set of observation indexes, and $J$ is the set of attribute indexes.

- We have $k(i) = \sum_{j \in J} k(i, j)$. Analogously $k(j)$ is defined, and $k = \sum_{i \in I, j \in J} k(i, j)$.
Correspondence Analysis: Mapping $\chi^2$ Distances into Euclidean Distances

- The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i, j) = k(i, j); i \in I, j \in J\}$.
- $I$ is the set of observation indexes, and $J$ is the set of attribute indexes.
- We have $k(i) = \sum_{j \in J} k(i, j)$. Analogously $k(j)$ is defined, and $k = \sum_{i \in I, j \in J} k(i, j)$.
- Next, $f_{IJ} = \{f_{ij} = k(i, j)/k; i \in I, j \in J\} \subset \mathbb{R}_{I \times J}$, similarly $f_i$ is defined as $\{f_i = k(i)/k; i \in I, j \in J\} \subset \mathbb{R}_I$, and $f_J$ analogously.
Correspondence Analysis: Mapping $\chi^2$ Distances into Euclidean Distances

- The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i, j) = k(i, j); i \in I, j \in J\}$.
- $I$ is the set of observation indexes, and $J$ is the set of attribute indexes.
- We have $k(i) = \sum_{j \in J} k(i, j)$. Analogously $k(j)$ is defined, and $k = \sum_{i \in I, j \in J} k(i, j)$.
- Next, $f_{IJ} = \{f_{ij} = k(i, j)/k; i \in I, j \in J\} \subset \mathbb{R}_{I \times J}$, similarly $f_i$ is defined as $\{f_i = k(i)/k; i \in I, j \in J\} \subset \mathbb{R}_I$, and $f_J$ analogously.
- What we have described here is taking numbers of occurrences into relative frequencies.
Correspondence Analysis: Mapping $\chi^2$ Distances into Euclidean Distances

- The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i, j) = k(i, j); i \in I, j \in J\}$.
- $I$ is the set of observation indexes, and $J$ is the set of attribute indexes.
- We have $k(i) = \sum_{j \in J} k(i, j)$. Analogously $k(j)$ is defined, and $k = \sum_{i \in I, j \in J} k(i, j)$.
- Next, $f_{IJ} = \{f_{ij} = k(i, j)/k; i \in I, j \in J\} \subset \mathbb{R}_{I \times J}$, similarly $f_i$ is defined as $\{f_i = k(i)/k; i \in I, j \in J\} \subset \mathbb{R}_{I}$, and $f_J$ analogously.
- What we have described here is taking numbers of occurrences into relative frequencies.
- The conditional distribution of $f_J$ knowing $i \in I$, also termed the $j$th profile with coordinates indexed by the elements of $I$, is:

$$f_j^j = \{f_j^i = f_{ij}/f_i = (k_{ij}/k)/(k_i/k); f_i > 0; j \in J\}$$

and likewise for $f_i^j$. 
Input: Cloud of Points Endowed with the Chi Squared Metric

- The cloud of points consists of the couples: (multidimensional) profile coordinate and (scalar) mass. We have $N_J(I) = \{(f^i_j, f_i); i \in I\} \subset \mathbb{R}_J$, and again similarly for $N_I(J)$.
Input: Cloud of Points Endowed with the Chi Squared Metric

- The cloud of points consists of the couples: (multidimensional) profile coordinate and (scalar) mass. We have $N_J(I) = \{(f_j^i, f_i); i \in I\} \subset \mathbb{R}_J$, and again similarly for $N_I(J)$.

- Included in this expression is the fact that the cloud of observations, $N_J(I)$, is a subset of the real space of dimensionality $|J|$ where $|.|$ denotes cardinality of the attribute set, $J$. 
The cloud of points consists of the couples: (multidimensional) profile coordinate and (scalar) mass. We have $N_J(I) = \{(f^i_J, f^i); i \in I\} \subset \mathbb{R}_J$, and again similarly for $N_I(J)$.

Included in this expression is the fact that the cloud of observations, $N_J(I)$, is a subset of the real space of dimensionality $|J|$ where $|.|$ denotes cardinality of the attribute set, $J$.

The overall inertia is as follows:

$$M^2(N_J(I)) = M^2(N_I(J)) = \| f_{ij} - f_i f_j \|^2_{f_i f_j}$$

$$= \sum_{i \in I, j \in J} (f_{ij} - f_i f_j)^2 / f_i f_j \quad (1)$$
The term $\| f_{IJ} - f_I f_J \|_{f_I f_J}^2$ is the $\chi^2$ metric between the probability distribution $f_{IJ}$ and the product of marginal distributions $f_I f_J$, with as center of the metric the product $f_I f_J$. 

Decomposing the moment of inertia of the cloud $N_J(I)$ – or of $N_I(J)$ since both analyses are inherently related – furnishes the principal axes of inertia, defined from a singular value decomposition.
The term $\| f_{IJ} - f_I f_J \|^2_{f_I f_J}$ is the $\chi^2$ metric between the probability distribution $f_{IJ}$ and the product of marginal distributions $f_I f_J$, with as center of the metric the product $f_I f_J$.

Decomposing the moment of inertia of the cloud $N_J(I)$ – or of $N_I(J)$ since both analyses are inherently related – furnishes the principal axes of inertia, defined from a singular value decomposition.
The $\chi^2$ distance with center $f_j$ between observations $i$ and $i'$ is written as follows in two different notations:

$$d(i, i')^2 = \| f_j - f'_j \|_{f_j}^2 = \sum_j \frac{1}{f_j} \left( \frac{f_{ij}}{f_i} - \frac{f_{ij'}}{f_i'} \right)^2$$  \hspace{1cm} (2)
Output: Cloud of Points Endowed with the Euclidean Metric in Factor Space

- The $\chi^2$ distance with center $f_j$ between observations $i$ and $i'$ is written as follows in two different notations:

$$d(i, i')^2 = \parallel f_j - f'_j \parallel^2_{f_j} = \sum_j \frac{1}{f_j} \left( \frac{f_{ij}}{f_i} - \frac{f'_{ij}}{f_i'} \right)^2$$ (2)

- In the factor space this pairwise distance is identical. The coordinate system and the metric change. For factors indexed by $\alpha$ and for total dimensionality $N$ ($N = \min \{ |I| - 1, |J| - 1 \}$; the subtraction of 1 is since the $\chi^2$ distance is centered and hence there is a linear dependency which reduces the inherent dimensionality by 1) we have the projection of observation $i$ on the $\alpha$th factor, $F_\alpha$, given by $F_\alpha(i)$:

$$d(i, i')^2 = \sum_{\alpha=1..N} (F_\alpha(i) - F_\alpha(i'))^2$$ (3)
Output: Cloud of Points Endowed with the Euclidean Metric in Factor Space

▶ The $\chi^2$ distance with center $f_J$ between observations $i$ and $i'$ is written as follows in two different notations:

$$d(i, i')^2 = \| f_J - f'_J \|_f^2 = \sum_j \frac{1}{f_j} \left( \frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}} \right)^2 \quad (2)$$

▶ In the factor space this pairwise distance is identical. The coordinate system and the metric change. For factors indexed by $\alpha$ and for total dimensionality $N$ ($N = \min \{|I| - 1, |J| - 1\}$; the subtraction of 1 is since the $\chi^2$ distance is centered and hence there is a linear dependency which reduces the inherent dimensionality by 1) we have the projection of observation $i$ on the $\alpha$th factor, $F_\alpha$, given by $F_\alpha(i)$:

$$d(i, i')^2 = \sum_{\alpha=1..N} (F_\alpha(i) - F_\alpha(i'))^2 \quad (3)$$

▶ Invariance of distance in equations 2 and 3: Parseval relation.
In Correspondence Analysis the factors are ordered by decreasing moments of inertia. The factors are closely related, mathematically, in the decomposition of the overall cloud, $N_J(I)$ and $N_I(J)$, inertias. These are the dual spaces.
In Correspondence Analysis the factors are ordered by decreasing moments of inertia. The factors are closely related, mathematically, in the decomposition of the overall cloud, $N_J(I)$ and $N_I(J)$, inertias. These are the dual spaces.

The eigenvalues associated with the factors, identically in the space of observations indexed by set $I$, and in the space of attributes indexed by set $J$, are given by the eigenvalues associated with the decomposition of the inertia.
In Correspondence Analysis the factors are ordered by decreasing moments of inertia. The factors are closely related, mathematically, in the decomposition of the overall cloud, $N_J(I)$ and $N_I(J)$, inertias. These are the **dual spaces**.

The eigenvalues associated with the factors, identically in the space of observations indexed by set $I$, and in the space of attributes indexed by set $J$, are given by the eigenvalues associated with the decomposition of the inertia.

The decomposition of the inertia is a principal axis decomposition, which is arrived at through a singular value decomposition.
Important Consequences

- Given the inherent (mathematical) relationship between the dual spaces of observations and attributes, the eigen-reduction or decomposition of the cloud in terms of moments of inertia, is carried out in the lower dimensional of the dual spaces.

- The principle of distributional equivalence allows for aggregation of input data (observations, or attributes) with no effect on the analysis beyond the aggregated data. (Hence a type of scale-invariance principle.)

- Supplementary elements are observations or attributes retrospectively projected into the factor space.

- Further topics, not covered here: Data Coding. Multiple Correspondence Analysis.

  A Social Critique of the Judgment of Taste.
Important Consequences

- Given the inherent (mathematical) relationship between the dual spaces of observations and attributes, the eigen-reduction or decomposition of the cloud in terms of moments of inertia, is carried out in the lower dimensional of the dual spaces.

- The principle of distributional equivalence allows for aggregation of input data (observations, or attributes) with no effect on the analysis beyond the aggregated data. (Hence a type of scale-invariance principle.)
Important Consequences

- Given the inherent (mathematical) relationship between the dual spaces of observations and attributes, the eigen-reduction or decomposition of the cloud in terms of moments of inertia, is carried out in the lower dimensional of the dual spaces.

- The **principle of distributional equivalence** allows for aggregation of input data (observations, or attributes) with no effect on the analysis beyond the aggregated data. (Hence a type of scale-invariance principle.)

- Supplementary elements are observations or attributes retrospectively projected into the factor space.

Further topics, not covered here: Data Coding. Multiple Correspondence Analysis.

Important Consequences

- Given the inherent (mathematical) relationship between the dual spaces of observations and attributes, the eigen-reduction or decomposition of the cloud in terms of moments of inertia, is carried out in the lower dimensional of the dual spaces.

- The principle of distributional equivalence allows for aggregation of input data (observations, or attributes) with no effect on the analysis beyond the aggregated data. (Hence a type of scale-invariance principle.)

- Supplementary elements are observations or attributes retrospectively projected into the factor space.

- Further topics, not covered here: Data Coding. Multiple Correspondence Analysis.
Important Consequences

- Given the inherent (mathematical) relationship between the dual spaces of observations and attributes, the eigen-reduction or decomposition of the cloud in terms of moments of inertia, is carried out in the lower dimensional of the dual spaces.

- The **principle of distributional equivalence** allows for aggregation of input data (observations, or attributes) with no effect on the analysis beyond the aggregated data. (Hence a type of scale-invariance principle.)

- Supplementary elements are observations or attributes retrospectively projected into the factor space.

- Further topics, not covered here: Data Coding. Multiple Correspondence Analysis.

Contributions, Correlations

- Contributions

...
Contributions, Correlations

- Contributions
- Contribution of $i$ to moment $\alpha$: CTR: $f_i F_\alpha(i)^2 / \lambda_\alpha$

Cosine squared of angle between $i$ and factor $\alpha$.

$$\cos^2 \alpha = f_i F_\alpha(i)^2 / \rho(i)^2$$

where $\rho(i)^2 = \|f_i^T - f_J\|^2$
Contributions, Correlations

- Contributions
- Contribution of $i$ to moment $\alpha$: CTR: $f_i F_\alpha(i)^2/\lambda_\alpha$
- Correlations

\[
\cos^2 \alpha = F_\alpha(i)^2/\rho(i)^2
\]

where $\rho(i)^2 = \|f_i J - f_J\|^2$ and $f_J = \sum_{j \in J} (f_{ij} - f_j)^2/f_j$. 

Contributions determine the factor space, correlations illustrate it.
Contributions, Correlations

- Contributions
- Contribution of $i$ to moment $\alpha$: $\text{CTR}: f_i F_{\alpha}(i)^2 / \lambda_{\alpha}$
- Correlations
- Cosine squared of angle between $i$ and factor $\alpha$.
Contribution of $i$ to moment $\alpha$: CTR: $f_i F_\alpha(i)^2 / \lambda$$\alpha$

Correlations

Cosine squared of angle between $i$ and factor $\alpha$.

$\cos^2 a = F_\alpha(i)^2 / \rho(i)^2$ where $\rho(i)^2 = \|f_j^i - f_j\|^2_{f_j} = \sum_{j \in J} (f_j^i - f_j)^2 / f_j$
Contributions, Correlations

- **Contributions**
  - Contribution of $i$ to moment $\alpha$: CTR: $f_i F_\alpha(i)^2 / \lambda_\alpha$

- **Correlations**
  - Cosine squared of angle between $i$ and factor $\alpha$.
  - $\cos^2 a = F_\alpha(i)^2 / \rho(i)^2$ where $\rho(i)^2 = \|f_i^i - f_j\|^2_{f_j} = \sum_{j \in J} (f_i^j - f_j)^2 / f_j$

- Contributions *determine* the factor space, correlations *illustrate* it.
Hierarchical Clustering

- Consider the projection of observation $i$ onto the set of all factors indexed by $\alpha$, $\{F_\alpha(i)\}$ for all $\alpha$, which defines the observation $i$ in the new coordinate frame.
Hierarchical Clustering

- Consider the projection of observation $i$ onto the set of all factors indexed by $\alpha$, $\{F_\alpha(i)\}$ for all $\alpha$, which defines the observation $i$ in the new coordinate frame.
- This new factor space is endowed with the (unweighted) Euclidean distance, $d$. 
Hierarchical Clustering

- Consider the projection of observation $i$ onto the set of all factors indexed by $\alpha$, $\{F_\alpha(i)\}$ for all $\alpha$, which defines the observation $i$ in the new coordinate frame.
- This new factor space is endowed with the (unweighted) Euclidean distance, $d$.
- We seek a hierarchical clustering that takes into account the observation sequence, i.e. observation $i$ precedes observation $i'$ for all $i, i' \in I$. We use the linear order on the observations.
Sequence-Constrained Hierarchical Clustering

- Consider each text in the sequence of texts as constituting a singleton cluster. Determine the closest pair of adjacent texts, and define a cluster from them.

- Determine and merge the closest pair of adjacent clusters, \( c_1 \) and \( c_2 \), where closeness is defined by 
  \[
  d(c_1, c_2) = \max\{d(i, i') : i \in c_1, i' \in c_2\}.
  \]

- Repeat this merge step until only one cluster remains.

- Here we use a complete link criterion which additionally takes account of the adjacency constraint imposed by the sequence of texts in set \( I \).

- It can be shown that the closeness value, given by \( d \), at each agglomerative step is strictly non-decreasing.

- That is, if cluster \( c_3 \) is formed earlier in the series of agglomerations compared to cluster \( c_4 \), then the corresponding distances will satisfy
  \[
  d(c_3) \leq d(c_4).
  \]

- \( d \) here is as determined in the merge step of the algorithm above.
Consider each text in the sequence of texts as constituting a singleton cluster. Determine the closest pair of adjacent texts, and define a cluster from them.

Determine and merge the closest pair of adjacent clusters, \( c_1 \) and \( c_2 \), where closeness is defined by
\[
d(c_1, c_2) = \max \{d_{ii'} \text{ such that } i \in c_1, i' \in c_2\}.
\]

Repeat this merge step until only one cluster remains.

Here we use a complete link criterion which additionally takes account of the adjacency constraint imposed by the sequence of texts in set \( I \).

It can be shown that the closeness value, given by \( d \), at each agglomerative step is strictly non-decreasing. That is, if cluster \( c_3 \) is formed earlier in the series of agglomerations compared to cluster \( c_4 \), then the corresponding distances will satisfy \( d_{c_3} \leq d_{c_4} \). (\( d \) here is as determined in the merge step of the algorithm above.)
Sequence-Constrained Hierarchical Clustering

- Consider each text in the sequence of texts as constituting a singleton cluster. Determine the closest pair of adjacent texts, and define a cluster from them.
- Determine and merge the closest pair of adjacent clusters, $c_1$ and $c_2$, where closeness is defined by $d(c_1, c_2) = \max \{d_{ii'} \text{ such that } i \in c_1, i' \in c_2\}$.
- Repeat this merge step until only one cluster remains.

Here we use a complete link criterion which additionally takes account of the adjacency constraint imposed by the sequence of texts in set $I$.

It can be shown that the closeness value, given by $d$, at each agglomerative step is strictly non-decreasing. That is, if cluster $c_3$ is formed earlier in the series of agglomerations compared to cluster $c_4$, then the corresponding distances will satisfy $d_{c_3} \leq d_{c_4}$.
Consider each text in the sequence of texts as constituting a singleton cluster. Determine the closest pair of adjacent texts, and define a cluster from them.

Determine and merge the closest pair of adjacent clusters, $c_1$ and $c_2$, where closeness is defined by $d(c_1, c_2) = \max \left\{ d_{ii'} \text{ such that } i \in c_1, i' \in c_2 \right\}$.

Repeat this merge step until only one cluster remains.

Here we use a complete link criterion which additionally takes account of the adjacency constraint imposed by the sequence of texts in set $I$. 

It can be shown that the closeness value, given by $d$, at each agglomerative step is strictly non-decreasing. That is, if cluster $c_3$ is formed earlier in the series of agglomerations compared to cluster $c_4$, then the corresponding distances will satisfy $d_{c_3} \leq d_{c_4}$. ($d$ here is as determined in the merge step of the algorithm above.)
Consider each text in the sequence of texts as constituting a singleton cluster. Determine the closest pair of adjacent texts, and define a cluster from them.

Determine and merge the closest pair of adjacent clusters, $c_1$ and $c_2$, where closeness is defined by $d(c_1, c_2) = \max \{d_{ii'} \text{ such that } i \in c_1, i' \in c_2\}$.

Repeat this merge step until only one cluster remains.

Here we use a complete link criterion which additionally takes account of the adjacency constraint imposed by the sequence of texts in set $I$.

It can be shown that the closeness value, given by $d$, at each agglomerative step is strictly non-decreasing.
Sequence-Constrained Hierarchical Clustering

- Consider each text in the sequence of texts as constituting a singleton cluster. Determine the closest pair of adjacent texts, and define a cluster from them.

- Determine and merge the closest pair of adjacent clusters, $c_1$ and $c_2$, where closeness is defined by $d(c_1, c_2) = \max \{d_{ii'} \text{ such that } i \in c_1, i' \in c_2\}$.

- Repeat this merge step until only one cluster remains.

- Here we use a complete link criterion which additionally takes account of the adjacency constraint imposed by the sequence of texts in set $I$.

- It can be shown that the closeness value, given by $d$, at each agglomerative step is strictly non-decreasing.

- That is, if cluster $c_3$ is formed earlier in the series of agglomerations compared to cluster $c_4$, then the corresponding distances will satisfy $d_{c_3} \leq d_{c_4}$. ($d$ here is as determined in the merge step of the algorithm above.)
Example of Hierarchy Without and With Inversion

- Inversions in the sequence of agglomerations.
- That is, \(i\) and \(j\) merge, and the distance of the this new cluster to another cluster is smaller than the dening distance of the \(i; j\) merger.
- Hence, there is non-monotonic change in the level index, given by the distance dening the merger agglomeration.
Figure: Hierarchical clustering using the Ward minimum variance agglomerative criterion.
Hierarchy (not sequence-constrained, 30 terms)

Figure: Median agglomerative criterion. (For each agglomeration, minimize the median of the pairwise dissimilarities.)