Stat 315c: Transposable Data
Correspondence Analysis

Art B. Owen

Stanford Statistics
Correspondence Analysis

It plots both variables and cases in the same plane. The clearest motivation is for contingency table data. It gets used elsewhere too. Emphasis is on presenting the data themselves as opposed to illuminating an underlying model. This is an old and classical statistical technique pioneered by Jean-Paul Benzécri in the 1960s. The treatment by Greenacre is particularly clear.
Correspondence Analysis

- It plots both variables and cases in the same plane.
Correspondence Analysis

- It plots both variables and cases in the same plane.
- Clearest motivation is for contingency table data. It gets used elsewhere too.
Correspondence Analysis

- It plots both variables and cases in the same plane.
- Clearest motivation is for contingency table data. It gets used elsewhere too.
- Emphasis is on presenting the data themselves as opposed to illuminating an underlying model.
Correspondence Analysis

- It plots both variables and cases in the same plane.
- Clearest motivation is for contingency table data. It gets used elsewhere too.
- Emphasis is on presenting the data themselves as opposed to illuminating an underlying model.
- This is an old and classical statistical technique pioneered by Jean-Paul Benzécri in the 1960s.
Correspondence Analysis

- It plots both variables and cases in the same plane.
- Clearest motivation is for contingency table data. It gets used elsewhere too.
- Emphasis is on presenting the data themselves as opposed to illuminating an underlying model.
- This is an old and classical statistical technique pioneered by Jean-Paul Benzécri in the 1960s.
- The treatment by Greenacre is particularly clear.
Contingency tables

$I \times J$ table of counts

\[
\begin{array}{cccc}
  n_{11} & n_{12} & \cdots & n_{1J} \\
  n_{21} & n_{22} & \cdots & n_{2J} \\
  \vdots & \vdots & \ddots & \vdots \\
  n_{I1} & n_{I2} & \cdots & n_{IJ} \\
\end{array}
\]
Contingency tables

$I \times J$ table of counts

\[
\begin{array}{cccc}
  n_{11} & n_{12} & \cdots & n_{1J} \\
  n_{21} & n_{22} & \cdots & n_{2J} \\
  \vdots & \vdots & \ddots & \vdots \\
  n_{I1} & n_{I2} & \cdots & n_{IJ} \\
\end{array}
\]

Nomenclature

- **Correspondence matrix** $P$  
  \[ p_{ij} = n_{ij}/n_{..} \]

- **Row masses**  
  \[ r_i = p_{i.} = n_{i.}/n_{..} \]

- **Column masses**  
  \[ c_j = p_{.j} = n_{.j}/n_{..} \]

- **Row profiles**  
  \[ \bar{r}_i = (p_{i1}/r_i, \ldots, p_{ij}/r_i)' \in \mathbb{R}^J \]

- **Column profiles**  
  \[ \bar{c}_j = (p_{1j}/c_j, \ldots, p_{Ij}/c_j)' \in \mathbb{R}^I \]
Contingency tables

$I \times J$ table of counts

\[
\begin{array}{cccc}
  n_{11} & n_{12} & \cdots & n_{1J} \\
  n_{21} & n_{22} & \cdots & n_{2J} \\
  \vdots & \vdots & \ddots & \vdots \\
  n_{I1} & n_{I2} & \cdots & n_{IJ} \\
\end{array}
\]

Nomenclature

Correspondence matrix $P$

\[
p_{ij} = \frac{n_{ij}}{n_{\cdot \cdot}}
\]

Row masses

\[
r_i = p_{i \cdot} = \frac{n_{i \cdot}}{n_{\cdot \cdot}}
\]

Column masses

\[
c_j = p_{\cdot j} = \frac{n_{\cdot j}}{n_{\cdot \cdot}}
\]

Row profiles

\[
\bar{r}_i = (p_{i1}/r_i, \ldots, p_{iJ}/r_i)' \in \mathbb{R}^J
\]

Column profiles

\[
\bar{c}_j = (p_{1j}/c_j, \ldots, p_{IJ}/c_j)' \in \mathbb{R}^I
\]

These are conditional and marginal distributions
First moments: centroids

Row centroid

\[
\sum_{i=1}^{I} r_i \bar{r}_i = \sum_{i=1}^{I} r_i \left( \frac{p_{i1}}{r_i}, \ldots, \frac{p_{iJ}}{r_i} \right)' = (c_1, \ldots, c_J)' \equiv c
\]
First moments: centroids

Row centroid

$$\sum_{i=1}^{I} r_i \bar{r}_i = \sum_{i=1}^{I} r_i \left( \frac{p_{i1}}{r_i}, \ldots, \frac{p_{ij}}{r_i} \right)' = (c_1, \ldots, c_J)' \equiv c$$

Column centroid

$$\sum_{j=1}^{J} c_j \bar{c}_j = (r_1, \ldots, r_I)' \equiv r$$
First moments: centroids

Row centroid

\[
\sum_{i=1}^{I} r_i \bar{r}_i = \sum_{i=1}^{I} r_i \left( \frac{p_{i1}}{r_i}, \ldots, \frac{p_{iJ}}{r_i} \right)' = (c_1, \ldots, c_J)' \equiv c
\]

Column centroid

\[
\sum_{j=1}^{J} c_j \bar{c}_j = (r_1, \ldots, r_I)' \equiv r
\]

Upshot

‘Mass’ weighted average of row profiles is marginal distribution over columns
**Second moments: inertias**

**Chisquare for independence as weighted Euclidean distance**

\[
X^2 = \sum_i \sum_j \left( \frac{n_{ij} - n_{i\bullet}n_{\bullet j}/n_{\bullet\bullet}}{n_{i\bullet}n_{\bullet j}/n_{\bullet\bullet}} \right)^2
\]

\[
= \sum_i n_{i\bullet} \sum_j \left( \frac{n_{ij}/n_{i\bullet} - n_{\bullet j}/n_{\bullet\bullet}}{n_{\bullet j}/n_{\bullet\bullet}} \right)^2
\]

\[
= n_{\bullet\bullet} \sum_i \frac{n_{i\bullet}}{n_{\bullet\bullet}} \sum_j \left( \frac{n_{ij}/n_{i\bullet} - n_{\bullet j}/n_{\bullet\bullet}}{n_{\bullet j}/n_{\bullet\bullet}} \right)^2
\]

\[
= n_{\bullet\bullet} \sum_i r_i (\bar{r}_i - c)' \text{diag}(c)^{-1} (\bar{r}_i - c)
\]

\[
= n_{\bullet\bullet} \times \text{Inertia}
\]
Second moments: inertias

Chisquare for independence as weighted Euclidean distance

\[ X^2 = \sum_i \sum_j \frac{(n_{ij} - n_i n_j / n_{..})^2}{n_i n_j / n_{..}} \]

\[ = \sum_i n_i \sum_j \frac{(n_{ij} / n_i - n_j / n_{..})^2}{n_j / n_{..}} \]

\[ = n_{..} \sum_i \frac{n_i}{n_{..}} \sum_j \frac{(n_{ij} / n_i - n_j / n_{..})^2}{n_j / n_{..}} \]

\[ = n_{..} \sum_i r_i (\bar{r}_i - c)' \text{diag}(c)^{-1} (\bar{r}_i - c) \]

\[ = n_{..} \times \text{Inertia} \]

This is the total inertia of the row profiles. It equals total inertia of column profiles.
For $J = 3$

12 8 8
7 7 6
8 8 10
6 9 8
9 8 7

We can plot profiles in $\mathbb{R}^3$

Low inertia
Geometry

This example has higher inertia
Geometry

- Still higher inertia.
- $\chi^2$ statistics describes variation of row profiles
- Similarly for col profiles
Rescale

Distances

- Euclidean distance in plot ignores column values
- Replace $\tilde{r}_i$ by $\tilde{\tilde{r}}_i$ with $\tilde{\tilde{r}}_{ij} = \frac{r_{ij}}{\sqrt{c_j}}$
- Euclidean dist between $\tilde{\tilde{r}}_i$ and $\tilde{\tilde{r}}_i'$ is \( \chi^2 \) dist between $\tilde{r}_i$ and $\tilde{r}_i'$. 
Rescale

Distances

- Euclidean distance in plot ignores column values
- Replace $\tilde{r}_i$ by $\tilde{\tilde{r}}_i$ with $\tilde{\tilde{r}}_{ij} = \frac{r_{ij}}{\sqrt{c_j}}$
- Euclidean dist between $\tilde{\tilde{r}}_i$ and $\tilde{\tilde{r}}_i'$ is "$\chi^2$ dist" between $\tilde{r}_i$ and $\tilde{r}_i'$. 
Reason for $\chi^2$

Invariance

- Suppose rows $i$ and $i'$ are proportional
- $n_{ij}/n_{i'j} = \alpha$ all $j = 1, \ldots, J$
- Suppose also that we pool these rows
- New $n_{i*j} = n_{ij} + n_{i'j}$
- and delete originals
Reason for $\chi^2$

Invariance

- Suppose rows $i$ and $i'$ are proportional
  
  \[ \frac{n_{ij}}{n_{i'j}} = \alpha \text{ all } j = 1, \ldots, J \]

- Suppose also that we pool these rows

- New $n_{i*j} = n_{ij} + n_{i'j}$

  and delete originals

Then

- New $\chi^2$ distance between cols $j$ and $j'$ equals old dist

- Principle of distributional equivalence

- Common profile, summed mass
Reason for $\chi^2$

Invariance

- Suppose rows $i$ and $i'$ are proportional
- $n_{ij}/n_{i'j} = \alpha$ all $j = 1, \ldots, J$
- Suppose also that we pool these rows
- New $n_{i^*j} = n_{ij} + n_{i'j}$
- and delete originals

Then

- New $\chi^2$ distance between cols $j$ and $j'$ equals old dist
- Principle of distributional equivalence
- Common profile, summed mass

Role of $\chi^2$ in statistical significance is not considered important in this literature
Dimension reduction

Now we have a plot

- With rows and cols both in $\mathbb{R}^{J-1}$
Dimension reduction

Now we have a plot
- With rows and cols both in $\mathbb{R}^{J-1}$

If $J$ is too big
- reduce dimension
- by principal components of
  $$\frac{\overline{R}_{ij} - c_j}{\sqrt{c_j}}$$
- plot in reduced dimension
- along with images of corners
Duality

- Rows lie in $\min(I - 1, J - 1)$ dimensional space
- So do columns
- In PC of row profiles . . . columns are outside
- In PC of column profiles . . . rows are outside
- Symmetric correspondence analysis overlap the points after rescaling

$$D_r = \text{diag}(r_1, \ldots, r_I)$$

$$D_c = \text{diag}(c_1, \ldots, c_J)$$
Duality

- Rows lie in $\min(I - 1, J - 1)$ dimensional space
- So do columns
- In PC of row profiles . . . columns are outside
- In PC of column profiles . . . rows are outside
- Symmetric correspondence analysis overlap the points after rescaling

More notation

$$D_r = \text{diag}(r) = \text{diag}(r_1, \ldots, r_I)$$
$$D_c = \text{diag}(c) = \text{diag}(c_1, \ldots, c_J)$$
Symmetric analysis

- Uses SVD $S = U \Sigma V'$ where
  $$s_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}$$

- Total inertia is $\|S\|_F^2$
- 'principal inertias' are $\lambda_j^2$
Symmetric analysis

- Uses SVD $S = U\Sigma V'$ where
  
  $$s_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}$$

- Total inertia is $\|S\|_F^2$
- 'principal inertias' are $\lambda_j^2$

Coordinates

- Rows (1st k cols of)
  
  $$F = (D_r^{-1} P - 1r')D_c^{-1}V_{1:k} = D_r^{-1}U\Sigma$$

- Columns (1st k cols of)
  
  $$G = (D_c^{-1} P - 1c')D_r^{-1}U_{1:k} = D_c^{-1}V\Sigma'$$
### Symmetric analysis

#### Interpretation is tricky/controversial

- $r_i$ near $r_i'$ \(\checkmark\)
- $c_j$ near $c_j'$ \(\checkmark\)
- $r_i$ near $c_j$ ??
  Rows and columns are not in the same space

#### Biplots

- Due to Gabriel (1971) Biometrika
- For matrix $X_{ij}$
  - plot rows as $u_i \in \mathbb{R}^2$
  - cols as $v_j \in \mathbb{R}^2$
  - with $u_i'v_j = X_{ij}$

A biplot interpretation applies to asymmetric plots
Some finer points

Ghost points

- Apply projection to point not in table
- E.G. hypothetical row entity,
  1. impute a president’s ‘senate voting record’
  2. compare a state’s economy to those of countries
- Treat as fixed profile with mass ↓ 0
Some finer points

Ghost points
- Apply projection to point not in table
- E.G. hypothetical row entity,
  1. impute a president’s ’senate voting record’
  2. compare a state’s economy to those of countries
- Treat as fixed profile with mass \(\downarrow 0\)

Merged points
- Add linear combination or sum of rows, E.G.
  1. pool columns for math and statistics into “math sciences”
  2. pool rows for EU countries into an EU point
Data types

- Counts are straightforward
- Other 'near measures' are reasonable
  - rainfalls, heights, volumes, temperatures Kelvin
  - dollars spent
  - parts per million
- Reweight cols to equalize inertia \( \approx \) standardizing to equalize variance
  Requires iteration
Data types

- Counts are straightforward
- Other 'near measures' are reasonable
  - rainfalls, heights, volumes, temperatures Kelvin
  - dollars spent
  - parts per million
- Reweight cols to equalize inertia \( \approx \) standardizing to equalize variance
  Requires iteration

Fuzzy coding

\( x \in \mathbb{R} \) becomes two columns

- \((1, 0)\) for small \( x \) say \( x < L \)
- \((0, 1)\) for large \( x \) say \( x > U \)
- \((1 - t, t)\) for intermediate \( x \) \( t = (x - L)/(U - L) \)

Generalizations to \( > 2 \) columns
Puzzlers

- Does it scale? (eg $10^8$ points in the plane)
- Is there a tensor version? (Beyond all pairs of two way versions)
- Distributional equivalence vs Poisson models
Puzzlers

- Does it scale? (e.g., 10^8 points in the plane)
- Is there a tensor version? (Beyond all pairs of two-way versions)
- Distributional equivalence vs Poisson models

Further reading

- “Correspondence Analysis in Practice” M.J. Greenacre, 1993
  Emphasizes geometry with examples
- “Theory and Applications of Correspondence Analysis” M.J. Greenacre, 1984
  Good coverage of theory with examples
- “Correspondence Analysis and Data Coding with Java and R” F. Murtagh, 2005
  Code and worked examples