Interactive CA in a Object Oriented Environment

Jason Bond
George Michailidis
UCLA Statistics

June 26, 1997
1 Exploratory Data Analysis

1. Ability to Handle Categorical Data
2. No Models
3. Few Assumptions
4. Emphasis on Graphical Presentation
## Contingency Tables

### Fishers Hair/Eye Color Dataset

<table>
<thead>
<tr>
<th></th>
<th>Hair</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fair</td>
<td>Red</td>
</tr>
<tr>
<td>Light</td>
<td>688</td>
<td>116</td>
</tr>
<tr>
<td>Blue</td>
<td>326</td>
<td>38</td>
</tr>
<tr>
<td>Med</td>
<td>343</td>
<td>84</td>
</tr>
<tr>
<td>Dark</td>
<td>98</td>
<td>48</td>
</tr>
<tr>
<td>r</td>
<td>1455</td>
<td>286</td>
</tr>
</tbody>
</table>

If the contingency table above, \( N = 5387 \),

\[
D_r = \begin{pmatrix}
1455 & 0 & 0 & 0 & 0 & 0 \\
0 & 286 & 0 & 0 & 0 & 0 \\
0 & 0 & 2137 & 0 & 0 & 0 \\
0 & 0 & 0 & 1391 & 0 & 0 \\
0 & 0 & 0 & 0 & 118 & 0
\end{pmatrix}
\] (1)

and \( D_c \) similariy.
2 Correspondence Analysis - ANACOR

Contributors

1. Pearson (1906)
2. Hirschfeld (1935)
3. Fisher (1940)
4. Guttman (1941)
5. Hayashi (1952)
6. Benzecri (1973)
The Row Problem

The $\chi^2$-distance between row $i$ and row $i'$ of $F$ is

$$
\delta^2_{r_i,r_i'} = N \sum_j \left[ \frac{f_{i,j}}{r_i} - \frac{f_{i',j}}{r_{i'}} \right]^2 / c_j
$$

Let $B = N^{\frac{1}{2}} D_r^{-1} (F - \frac{D_r uu' D_c}{N}) D_c^{-\frac{1}{2}}$, then

$$
\delta^2_{r_i,r_i'} = (b_i - b_{i'})'(b_i - b_{i'}), b_i \text{ the } i^{th} \text{ row of } B.
$$

Goal: Find a $p < l$ dimensional representation of the rows that preserves the row $\chi^2$-distances as well as possible, or equivalently the interpoint euclidean distances between the rows of $B$. 
Think of the rows of $B$ as data points and do PCA, ie, solve

\[
\min_X \|BB' - XX'\|^2
\]  

over $X$ of rank $p < k$ such that $X'D_rX = N\Lambda$
and $u'D_rX = 0$, where $\|\cdot\|^2$ is the sum of squares
norm.
The SVD

\[ N^{-\frac{1}{2}} D_r^{\frac{1}{2}} B = D_r^{\frac{1}{2}} (F - \frac{D_r uu' D_c}{N}) D_c^{\frac{1}{2}} = K \Lambda L' \] (4)

The pair \((\frac{D_r^{\frac{1}{2}} u}{\sqrt{N}}, \frac{D_c^{\frac{1}{2}} u}{\sqrt{N}})\) are a left and right singular vector pair of \(D_r^{-\frac{1}{2}} F D_c^{-\frac{1}{2}}\). The difference \(F - \frac{D_r uu' D_c}{N}\) measures departure of \(F\) from independence.

The Solution

\[ X = N^{\frac{1}{2}} D_r^{-\frac{1}{2}} K_p \Lambda_p \] (5)

\(K_p\) the first \(p\) columns of \(K\). \(\Lambda_p\) the first \(p\) rows and columns of \(\Lambda\).

Note that

\[ \|D_r^{\frac{1}{2}} B\|^2 = N \sum_{i,j} \left( \frac{f_{i,j} - \frac{r_{i,c,j}}{N}}{\frac{r_{i,c,j}}{N}} \right)^2 \]

\[ = N \text{tr}(K \Lambda^2 K') = N \sum \lambda_i^2 \] (6)
so the total $\chi^2$ for $F$ is approximated by the sum of the lengths of the weighted projections onto the optimal subspace of row quantifications.

**The Column Problem**

Same as the row problem, just transpose $F$.

**The Solution**

Think of the rows of $B$ as data points and do PCA again. The column solution is

$$Y = N^{\frac{1}{2}} D_c^{-\frac{1}{2}} L_p \Lambda_p$$  \hspace{1cm} (7)
Normalization

Define

\[ X = N^{-\frac{1}{2}} D_r^{-\frac{1}{2}} K_p \]
\[ Y = N^{-\frac{1}{2}} D_c^{-\frac{1}{2}} L_p \]

Several possible normalizations

(1) \((X \Lambda_p, Y)\) \hspace{1cm} (3) \((X \Lambda_p^{\frac{1}{2}}, Y \Lambda_p^{\frac{1}{2}})\)

(2) \((X, Y \Lambda_p)\) \hspace{1cm} (4) \((X \Lambda_p, Y \Lambda_p)\)

with properties

1. Using (1), the row points are the center of gravity of the column points \((X = D_r^{-1} FY)\)

2. Using (2), the column points are the center of gravity of the row points \((Y = D_c^{-1} FX)\)

3. For the first 3, the best reconstruction of \(F\) is \[ F = \frac{1}{N} (D_r uu' D_c + D_r X Y' D_c), \] the fourth needs a \(\Lambda^{-1}\).

4. For (4) interpoint distances are equivalent to \(\chi^2\)-distances.
Figure 1: \((X\Lambda_p, Y)\) - Rows Centroids of the Columns
Figure 2: \((X,Y\Lambda_p)\) - Columns Centroids of the Rows
Figure 3: \((X\Lambda_p^{\frac{1}{2}}, Y\Lambda_p^{\frac{1}{2}})\) - Symmetric Treatment
Equivalent Approaches

Let $G_r$, $G_c$ be the indicator matrices for the two variables. $G_r$ has $N$ rows and $k_1$ columns. $G_c$ has $N$ rows and $k_2$ columns. Then $F = G'_r G_c$, and $q_r = G_r x$, $q_c = G_c y$ are the quantified variables.

Maximize Correlation Between Quantified Variables

Requiring $x' r = y' c = 0$, we maximize

$$\text{Max}_{x, y} \rho = \frac{y' F x}{\sqrt{x' D_x y y' D_y y}}$$  \hspace{1cm} (8)

Using Lagrange multipliers and maximizing (8), we get

$$x = D_r^{-1} F' y / \rho$$ \hspace{1cm} (9)

$$y = D_c^{-1} F x / \rho$$ \hspace{1cm} (10)

Linearizing Bivariate Regressions With Maximum Correlation

Equations (9) and (10).
Maximize Correlation Ratio

\[ \max_x \eta^2 = \frac{SS_b}{SS_t} = \frac{x'[F'D_r^{-1}F - \frac{D_r uu'D_r}{N}]x}{x'[D_c - \frac{D_c uu'D_c}{N}]x} \quad (11) \]

or

\[ \max_w w'D_c^{-\frac{1}{2}}F'D_r^{-1}F D_c^{-\frac{1}{2}}w \quad (12) \]

subject to \( w'w = N \).
3 MCA of the Burt Matrix

What if we have more than 2 variables? Set up the same as in Homogeneity Analysis (MCA) (Gifi, 1991).

Let $G_j$ be the indicator matrix for the $j^{th}$ variable ($j = 1, \ldots, m$) having $k_j$ categories, $G = [G_1, \ldots, G_m]$ the super indicator matrix, $D_j = G'_j G_j$, $D = diag(D_1, \ldots D_m)$. 
MCA finds

1. optimal scores, $X$, satisfying
   
   $X'X = I, u'X = 0$, for objects (rows of the data matrix).

2. optimal quantifications, $Y$, for categories of variables

such that the departure from homogeneity

\[
\sigma(X, Y_1, \ldots, Y_m) = \sum_{j=1}^{m} \|X - G_jY_j\|^2 \tag{13}
\]

is minimized.
MCA minimizes (13) by computing $K_p$, $L_p$, and $\Lambda_p$ by alternating least squares methods where

$$m^{-\frac{1}{2}}GD^{-\frac{1}{2}} = K\Lambda L'$$  \hspace{1cm} (14)

with solution

$$X = \sqrt{n}K_p$$  \hspace{1cm} (15)

$$Y = \sqrt{n}D^{-\frac{1}{2}}L_p\Lambda_p$$  \hspace{1cm} (16)

$K_p$, $L_p$, the first $p$ columns of $K$, $L$, and $\Lambda_p$ the first $p$ rows and columns of $\Lambda$. 
**Goal**: MCA of the Burt matrix computes just the optimal category quantifications, $Y$.

**The Burt matrix**: $C = G'G$ where $G$ is the super indicator matrix. $D$ is the block diagonal matrix with $D_j$ on the diagonals.

**The SVD**

$$m^{-1}D^{-\frac{1}{2}}(G'G - Duu'D/n)D^{-\frac{1}{2}} = LL' \quad (17)$$

Computational load much lower here than in MCA but we only get the category quantifications.

**The Solution**

As before

$$Y = \sqrt{n}D^{-\frac{1}{2}}L\Lambda \quad (18)$$
4 Analysis of Profiles - ANAPROF

Goal: Same as that of MCA

Here we require that $q << n$, the number of unique rows of the dataset is much less than the number of observations.

Define $H$ as the data matrix. $G_p$ is the $n \times q$ profile matrix with a one in the $t^{th}$ column of the $i^{th}$ row if the $i^{th}$ object has has the $t^{th}$ profile. Let $S$ be the $q \times \sum k_j$ matrix that indicates whether a category shows up in a particular profile, so that each row sum of $S$ will be $m$. Note $G = G_p S$. 
Simple Example

\[
H = \begin{pmatrix}
1 & 1 & 1 \\
2 & 1 & 2 \\
1 & 1 & 1 \\
1 & 2 & 2
\end{pmatrix}
\quad G_p = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1
\end{pmatrix}
\] (19)

\[
S = \begin{pmatrix}
1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 1
\end{pmatrix}
\] (20)

The contingency table \( F = G_p' G_p S \) has row marginals \( D_r = G_p' G_p \) and column marginals \( D_c = D \).
The SVD

\[ D_r^{-\frac{1}{2}} (F - D_r uu' D_c) D_c^{-\frac{1}{2}} = K^* \Lambda L' \]  

(21)

where \( K^* = (G'_p G_p)^{-\frac{1}{2}} G'_p K \).

The Solution

\[ X = G_p (G'_p G_p)^{-\frac{1}{2}} K^* \]  

(22)

\[ Y = \sqrt{n} D^{-\frac{1}{2}} L \Lambda \]  

(23)
but for plotting we just care about unique $X$, and can set $X = I_p(G_p^r G_p)^{-\frac{1}{2}} K^*$. When $q \approx n$ then any computational savings are lost since we compute the entire SVD. Better to use MCA. Other (free) programs for MCA

1. An LISP-STAT version. Available at http://www.stat.ucla.edu/journals/jss/

5 The Common SVD

All three problems produce an SVD such as

\[ D_r^{-\frac{1}{2}} (F - D_r uu'D_c) D_c^{-\frac{1}{2}} = K\Lambda L' \]  \hspace{1cm} (24)

So only one routine is needed. We could use ALS techniques, as does homogenaity analysis. In all 3 cases, the number of rows and columns of the matrix of interest should be much smaller than \( N \), so an explicit SVD can be used.
6 Stability

Analytic Stability for MCA (Gifi, 1991). We use the bootstrap here to measure replication stability. A resample is obtained by

1. **ANACOR** Sampling from a multinomial distribution with parameters $N$ and $\hat{p}_{i,j}$, the grand total of $F$ and $\frac{f_{i,j}}{N}$.

2. **ANAPROF** Sampling from a multinomial distribution with parameters $N$ and $1'(G_p'G_p)/N$. The resampled diagonal of $G_p'G_p$ is the only quantity needed for the formation of the matrix to be decomposed (assuming all profiles show up at least once).

3. **MCA of the Burt Matrix** Resampling from either the rows of the original data matrix or equivalently the rows of $G$. 
ANACOR and ANAPROF are both relatively inexpensive to bootstrap, each iteration requiring only a $k \times l$ and $q$ dimensional deviate, respectively, from a multinomial distribution. A resample from $G$ can be much more expensive, as $G$ has $N$ rows.
7 Implementation

Considerations

1. Free
2. Ease of Use (menu driven, no need to know the language)
3. Strong Graphical Display Abilities
4. OOP (code reuse)

Language choice: LISP-STAT

Plotting Enhancements

2. Point Separation for overlapping points.
3. Ability to Separate Bootstrap replications for a given solution point.
5. Adding more plots not difficult.
Inheritance

*Object*

Hardware Object Proto

Anacor Proto

Square-Table Score-Plot Proto

Category Quantification Plot Proto

Proto

Graph Overlay Proto

Zoom Overlay Proto

Window Proto

Graph Window Proto

Graph Proto

Scatterplot Proto

Anacor Plot Proto

Spin Proto

Anacor-2D Plot Proto

Anacor-3D Plot Proto

Anacor-Output Dialog-Proto

Ananprof-Output Dialog-Proto

Anacor-Plot Dialog-Proto

Ananprof-Plot Dialog-Proto

Corresp-Dialog-Proto

Corresp-Output Dialog-Proto

Corresp-Plot Dialog-Proto

Corresp-Dailog-Proto

Burt Dialog Proto

Burt-Output Dialog-Proto

Burt-Plot Dialog-Proto
Example Data from NELS:88. Longitudinal Study of N=27394 students. Variables are occupation student expects to have at age 30 and how far in school father wants student to go.
Decomposition of total inertia along principal axes

<table>
<thead>
<tr>
<th>AXES</th>
<th>INERTIA</th>
<th>%of INERTIA</th>
<th>Cum %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.13658</td>
<td>80.967</td>
<td>80.967</td>
</tr>
<tr>
<td>2</td>
<td>0.01205</td>
<td>7.1452</td>
<td>88.112</td>
</tr>
<tr>
<td>3</td>
<td>0.00711</td>
<td>4.2175</td>
<td>92.33</td>
</tr>
</tbody>
</table>

Total  0.16868

Decomposition of total inertia along principal axes for 20 Bootstrap Samples

<table>
<thead>
<tr>
<th>AXES</th>
<th>INERTIA BS MEANS</th>
<th>INERTIA BS SDs</th>
<th>%of Cum</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.13792</td>
<td>0.00455</td>
<td>75.633</td>
<td>75.6</td>
</tr>
<tr>
<td>2</td>
<td>0.01580</td>
<td>0.00272</td>
<td>8.666</td>
<td>84.3</td>
</tr>
<tr>
<td>3</td>
<td>0.01037</td>
<td>0.00139</td>
<td>5.688</td>
<td>90.0</td>
</tr>
</tbody>
</table>

Total  0.18236
Figure 4: First Two Dimensions
Figure 5: Zoom of First Two Dimensions
Figure 6: Solution With Passive Points
(a) Bootstrap Replications

(b) Bootstrap Selection Dialog