DCONJOINT

A Mu Sigma University Publication, April 2016

Mu Sigma

In This Issue...

Abstract

Correspondence Analysis is a method of decomposing a table of frequencies of two or more categorical variables, into coordinates for rows, columns, etc. With this decomposition the data can be displayed graphically with points for rows, columns, etc. The mathematical technique used is Singular Value Decomposition (SVD) of a matrix. Correspondence Analysis can be considered to be a categorical analog of principal component analysis.

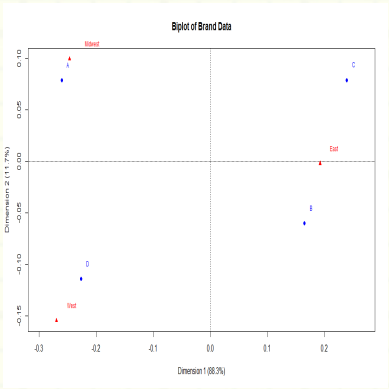
Correspondence Analysis

- By T. Krishnan

An Example

Let us consider an example of a two-way table of frequencies—the sales (number of items sold) of four different brands of refrigerators in three different regions of a country in a month. The data are as follows:

Brand⇓	Region		
	East	Midwest	West
A	51	43	22
B	92	28	21
C	68	22	9
D	34	22	18



One of the objectives of Correspondence Analysis is to represent the relationship between row (brand) elements and column (region) elements in terms of a plot called **biplot**, like the one given here. This is called a biplot since both row and column entities are represented in the graph which shows which brand is strong in which region. This also represents the distance between row entities (along the x-axis) and the distance between column entities (along the y-axis). One could interpret the distances of the row points from the centroid as the amount of contribution to the χ^2; similarly the column points.

The distances of the points in the two-dimensional display are important. These are informative in the sense that row points, which are close to each other are similar with regard to the pattern of relative frequencies across the columns. Along the most important first axis in the plot, brands A and D are close together on the left side of the origin. In these plots, one can only interpret the distances between row points, and the distances between column points, but not the distances between row points and column points.

The Method

Singular Value Decomposition decomposes a matrix into the product of three matrices—(a) a collection of *left singular vectors*, (b) a matrix of *singular values*, and (c) a collection of *right singular vectors*. Let us illustrate it for a two-way table

DCONJOINT

A Mu Sigma University Publication, April 2016

Mu Sigma

In This Issue...

with r rows and c columns. This decomposition begins with a matrix of standardized deviates, computed for each cell in the table, with a frequency of n_{ij} in the i^{th} row and j^{th} column. Let

$$n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}.$$

Let e_{ij} be the expected frequency in the ij^{th} cell under the assumption of row-column independence. Let

$$z_{ij} = \frac{1}{\sqrt{n}}\left(\frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}\right)$$

Notice that the sum of squares of z_{ij}'s is n times the χ^2 statistic for a test of row-column independence. Let $Z = ((z_{ij}))$ the matrix whose ij^{th} element is z_{ij}. And let $S = Z^T Z$. S has t non-zero eigenvalues where $t = \min(r-1, c-1)$. The sum of the eigenvalues of S is $n\chi^2$ and is called the **total inertia**. The matrix S is decomposed using SVD as

$$S = UDV^T,$$

U is a $r \times t$ matrix of row vectors, D is a $t \times t$ diagonal matrix of eigenvalues, and V is a $t \times c$ matrix of column vectors. Let the matrix D_r be the diagonal matrix of row totals of the frequencies and D_c be the diagonal matrix of column totals of frequencies. The standard row coordinates are $D_r^{-1/2}U$ and the standard column coordinates are $D_c^{-1/2}V$. For plotting in two dimensions, the first two coordinates are used of the above standardized versions of U and of V^T.

Total inertia is χ^2/n. Row mass for row i is $n_{i.}/n$ and column mass for column j is $n_{.j}/n$. A method for decomposing the overall Chi-square statistic (or Inertia=Chi-square/Total n) is by identifying a small number of dimensions in which the deviations from the expected values can be represented. This is similar to the objective of Principal Component Analysis, where the total variance is decomposed, so as to arrive at a lower-dimensional representation of the variables that allows one to reconstruct most of the variance/covariance matrix of variables.

A standard Correspondence Analysis output (from R's package "ca") for the above data is given below:

The output contains the eigenvalues and percentages of explained inertia for all possible dimensions. Additionally, values for the rows and columns (masses, chi-squared distances of points to their average, inertias and standard coordinates) are given. However, these values are restricted to two dimensions where applicable, e.g. for the standard coordinates.

DCONJOINT

A Mu Sigma University Publication, April 2016

Mu Sigma

In This Issue...


```
X-squared = 23.995, df = 6, p-value = 0.0005234


Principal inertias (eigenvalues):
           1       2
Value      0.04925 0.006552
Percentage 88.26%  11.74%


 Rows:
                A          B        C          D
Mass     0.269767  0.327907 0.230233  0.172093
ChiDist  0.272039  0.175832 0.251899  0.253868
Inertia  0.019964  0.010138 0.014609  0.011091
Dim. 1  -1.173162  0.744306 1.077977 -1.021352
Dim. 2   0.974480 -0.744621 0.974528 -1.412519


 Columns:
             East    Midwest       West
Mass     0.569767  0.267442  0.162791
ChiDist  0.192754  0.265841  0.310875
Inertia  0.021169  0.018900  0.015733
Dim. 1   0.868500 -1.111222 -1.214172
Dim. 2  -0.028458  1.226506 -1.915371
```

Elements of Output and Interpretation

Mass is the proportion of frequency in the column (or row).

Chidist gives the chisquare distances of points to their average indicating which row (or column) contributes relatively more to the interdependence of rows and columns.

Inertia: The relative inertia represents the proportion of the total inertia accounted for by the respective point, and it is independent of the number of dimensions chosen by the user. Note that a particular solution may represent a point very well, but the same point may not contribute much to the overall inertia (e.g., a row point with a pattern of relative frequencies across the columns that is similar to the average pattern across all rows).

Dim1, Dim2: These are the standardized units which could be plotted for rows and columns.

DCONJOINT

A Mu Sigma University Publication, April 2016

Mu Sigma

In This Issue...

Multiple Correspondence Analysis

Suppose you have a $k_1 \times k_2 \times k_3 \ldots$ table of frequencies. Let $p = k_1 + k_2 + k_3 + \ldots$ and n the total frequency. Z is a $n \times p$ matrix of dummy-coded profiles. Then the matrix $S = Z^T Z$ is constructed, which is rescaled and decomposed with an SVD.

References and Further Reading

- M.J.Greenacre (1993): *Correspondence Analysis in Practice*. London:Academic Press.
- O.Nenadi & M.J.Greenacre (2007): Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package. *Journal of Statistical Software*, **20**, 1–13.