Correspondence Analysis

Correspondence Analysis (CA) may have been evolved from methods used in multidimensional scaling, however Correspondence Analysis deals with a different type of data. While multidimensional scaling aims for continuous p dimensional data, correspondence analysis deals with categorical variables represented in contingency tables with count data.

The method of correspondence analysis can also apply to non-negative data matrix. Historically the approach has been discovered and rediscovered several times, with other names such as reciprocal averaging and dual scaling.

Correspondence analysis is usually used on two-way contingency tables. The generalization to three-way or high-way contingency tables is called multiple correspondence analysis.

1 Data type and objective of CA

Objectives of correspondence analysis

Correspondence analysis is a method for graphically displaying both the rows and columns of a two-way contingency table. Its graphical procedure aims to present associations between the row variable and the column variable, both are categorical.

Data of correspondence analysis

The following is a two-dimensional contingency table.

	col-variable level 1	col-level 2	$\cdots$	col-level j	$\cdots$	col-level J
Row-variable level 1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}
Row-variable level 2	n_{21}	n_{22}	$\cdots$	n_{2k}	$\cdots$	n_{2J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Row-variable level i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Row-variable level I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}

- Two categorical variables involved are of I and J categories respectively. The two variables are represented in the table as row and column variables.
- The (i, j)th entry n_{ij} in the table is the count of items simultaneously belonging to category i of the row variable and category j of the column variable.
- Usually the row variable and the column variables are in a equal position. The table can be presented in a transposed matter with row and column switched.

- In presentation, the data in the contingency table are often formatted as an $I \times J$ matrix,

$$\begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1j} & \cdots & n_{1J} \\ n_{21} & n_{22} & \cdots & n_{2j} & \cdots & n_{2J} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{i1} & n_{i2} & \cdots & n_{ij} & \cdots & n_{iJ} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{I1} & n_{I2} & \cdots & n_{Ij} & \cdots & n_{IJ} \end{bmatrix}$$

  Note that this is not the one-observation-per-row type of data.

- A common data format used on computer software is at the cell-item level. The dataset consists of $I \times J$ rows, each row is for one individual item with its corresponding categories in X and Y.

row category	*column category*	*counts*
1	1	n_{11}
$\vdots$	$\vdots$	$\vdots$
1	J	n_{1J}
2	1	n_{21}
$\vdots$	$\vdots$	$\vdots$
2	J	n_{2J}
3	1	n_{31}
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
I	J	n_{IJ}

2 Formulation and representation of CA

Correspondence Analysis yields graphical, map-like representation of the contingency table, especially when row and columns are symmetrically interesting.

Notations and profiles

To make row and column variable comparable, the data need to be scaled in several ways.

- $X = [n_{ij}]_{IJ} : \mathbb{R}^J \to \mathbb{R}^I$ (row space to column space) is the original data matrix.
  Usually X is a a **contingency table**.
- $P = \dfrac{X}{n} = \left[\dfrac{n_{ij}}{n}\right]$, $n = IJ$ is the **cell percentage matrix** with cell sum $= 1$.
- r_i, c_j are **row margin** and **column margin** of cell percentage matrix P.

$$r_i = \frac{\sum_{j=1}^{J} n_{ij}}{n} = \frac{n_{i\bullet}}{n}, \qquad c_j = \frac{\sum_{i=1}^{I} n_{ij}}{n} = \frac{n_{\bullet j}}{n}$$

- Two diagonal matrices are formed for scalling, such as producing the row and column profiles below.

$$D_r = diag(r_1, \cdots, r_I), \qquad D_c = diag(c_1, \cdots, c_J).$$

- $P_r = \left[\frac{p_{ij}}{r_i}\right] = \left[\frac{n_{ij}}{n_{i\bullet}}\right] = D_r^{-1}P$ is the **row profile matrix** with row sum = 1.

  $P_c = \left[\frac{p_{ij}}{c_j}\right] = \left[\frac{n_{ij}}{n_{\bullet j}}\right] = PD_c^{-1}$ is the **column profile matrix** with column sum = 1.

Distance between rows and columns

The row and column profiles provide a way to compare rows and columns.

The rows in the original contingency table X are not directly comparable. First of all, the row totals may be quite different. Same problem exist in the percentage matrix P. The row profile matrix is a good candidate.

When compare the components of two rows by their difference $\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}}$, the size of the difference is affected by the column percentage c_j. The larger the c_j, likely the larger $\frac{p_{ij}}{r_i}$ and $\frac{p_{i'j}}{r_{i'}}$.

Same issues need to be addressed when comparing columns as well.

- A reasonable **distance for row variables** is the χ^2-distance $d_{ii'}$ between rows i and i':

$$d_{ii'}^2 = \sum_{j=1}^{J} \frac{1}{c_j}\left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}}\right)^2$$

- Similarly, a reasonable **distance for column variables** is the χ^2-distance $d_{jj'}$ between columns j and j':

$$d_{jj'}^2 = \sum_{i=1}^{I} \frac{1}{r_i}\left(\frac{p_{ij}}{c_j} - \frac{p_{ij'}}{c_{j'}}\right)^2$$

The distances are compatible to be used on the same plot.

Expected values under independence

The row and column variables have associations when the variables are not independent.

One way to evaluate variable association is to examine the amount of deviation from independence between the row and column variables.

For an $I \times J$ contingency table with given row and column margins, the expected counts in cell (i, j) is

$$E_{ij} = nr_ic_j = \frac{n_{i\bullet}n_{\bullet j}}{n}, \qquad i = 1, \cdots, I,\ j = 1, \cdots, J,$$

where $n_{i\bullet} = \sum_{j=1}^{J} n_{ij}$ and $n_{\bullet j} = \sum_{i=1}^{I} n_{ij}$ are row and column totals defined earlier.

There are many test statistics to examine the independence between row and column variables in a contingency table.

The hypothesis test on the independence is

$$\begin{cases} H_o: & \text{rwo variable and column variable are independent.} \\ H_a: & \text{rwo variable and column variable are not independent.} \end{cases}$$

In correspondence analysis, the chi-squared statistic is used for testing independence.

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(I-1)(J-2)} \qquad under \quad H_o$$

Inertia of a contingency table

Total inertia for a contigency table is defined as

$$\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} - r_ic_j)^2}{r_ic_j} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = \frac{\chi^2}{n}$$

which is a measure of the amount of association between the row and column variables.

Standardized residual matrix

The **residual matrix**

$$P - rc^T$$

of dimension $I \times J$ is the difference between the observed and expected cell percentage, where the expected percentages are under the null assumption of independence between row and column variables.

The difference at cell (i, j) would be large if the row marginal percentage r_i or column marginal percentage c_j are large. Therefore it is reasonable to standardize the residuals.

The (Pearson) **standardized residual matrix** is

$$R = D_r^{-1/2}(P - rc^T)D_c^{-1/2} = \left[\frac{p_{ij} - r_ic_j}{\sqrt{r_i}\sqrt{c_j}}\right]_{i=1,\cdots,I;\ j=1,\cdots,J}$$

Large values in R indicates associations between the row and column variables.

View from row centroid and column centroid

The column proportions can be expressed as

$$(c_1, c_2, \cdots, c_J) = \sum_{i=1}^{I}\left(\frac{n_{i1}}{n},\ \frac{n_{i2}}{n},\ \cdots, \frac{n_{iJ}}{n}\right) = \sum_{i=1}^{I}\frac{n_{i\bullet}}{n}\left(\frac{n_{i1}}{n_{i\bullet}},\ \frac{n_{i2}}{n_{i\bullet}},\ \cdots, \frac{n_{iJ}}{n_{i\bullet}}\right)$$

which is an weighted average of row profiles $\left(\frac{n_{i1}}{n_{i\bullet}},\ \frac{n_{i2}}{n_{i\bullet}},\ \cdots, \frac{n_{iJ}}{n_{i\bullet}}\right)$, where the weights are row-percentage $\frac{n_{i\bullet}}{n}$.

Thus $(c_1, c_2, \cdots, c_J)$ can be considered as row centroid.

The χ^2 distance between rows can be written as

$$d_{ii'}^2 = \sum_{j=1}^{J}\frac{1}{c_j}\left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}}\right)^2 = \sum_{j=1}^{J}\frac{n}{n_{\bullet j}}\left(\frac{n_{ij}}{n_{i\bullet}} - \frac{n_{i'j}}{n_{i\bullet}}\right)^2$$

The χ^2 distance of row i and the row centroid $(c_1, c_2, \cdots, c_J)$ is

$$\sum_{j=1}^{J}\frac{n}{n_{\bullet j}}\left(\frac{n_{ij}}{n_{i\bullet}} - c_j\right)^2 = \sum_{j=1}^{J}\frac{n}{n_{\bullet j}}\left(\frac{n_{ij}}{n_{i\bullet}} - \frac{n_{\bullet j}}{n}\right)^2 = \frac{1}{n_{i\bullet}}\sum_{j=1}^{J}\frac{n}{n_{i\bullet}n_{\bullet j}}\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2$$

Then

$$\sum_{i=1}^{I} n_{i\bullet}\sum_{j=1}^{J}\frac{n}{n_{\bullet j}}\left(\frac{n_{ij}}{n_{i\bullet}} - c_j\right)^2 = \chi^2$$

3 Methods of CA

Correspondence analysis can be viewed as the weighted least squares problem of selecting $\hat{P} = [\hat{p}_{ij}]_{I,J}$ to minimize

$$trace\left[(D_r^{-1/2}(P-\hat{P})D_c^{-1/2})(D_r^{-1/2}(P-\hat{P})D_c^{-1/2})'\right] = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-\hat{p}_{ij})^2}{r_i c_j}$$

where the rank of $\hat{P}$ is specified low rank, in order to produce a low dimensional display.

By the Spectral Theorem, $\hat{P} = rc^T = rc'$ is the first order approximation of P (see section 12.7 in the textbook by Johnson and Wichern for a proof). Hence the approximation is reduce to that of

$$trace(RR') = tr\left[(D_r^{-1/2}(\hat{P}-rc^T)D_c^{-1/2})(D_r^{-1/2}(\hat{P}-rc^T)D_c^{-1/2})'\right] = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-r_i c_j)^2}{r_i c_j}$$

by the definition of the standardized residual matrix R.

RR' is an $I \times I$ symmetric semi-definite matrix with rank r, $r \leq \min\{I, J\}$. The non-zero eigenvalues of RR' can be denoted as $\lambda_k^2, k = 1, \cdots, r$. By the definition of matrix trace,

$$trace(RR') = \sum_{k=1}^{k}\lambda_k^2$$

By convention,

$$\lambda_1^2 \geq \cdots \geq \lambda_r^2 > 0.$$

Conduct a Singular Value Decomposition (SVD) on the standardized residual matrix R,

$$R = U\Sigma V', \qquad with \qquad U'U = I_r \ (or\ I_I), \qquad VV' = I_r \ (or\ I_J)$$

Then λ_k's must be the singular values of R on the kth diagonal of Σ, and

$$R'RV = V\Sigma'\Sigma, \qquad RR'U = U\Sigma\Sigma'.$$

Note that there are different versions of singular value decompositions in terms of dimensions of the matrices U, V, and Σ. Here if we use the decomposition where the singular value matrix Σ is an $r \times r$ matrix, then $\Sigma' = \Sigma$.

By the singular value decomposition, V consists of orthonormal eigenvectors of $R'R$, each eigenvector v_k is a linear combination of the columns of $R'R$, and U consists of orthonormal eigenvectors of RR', each eigenvector u_k is a linear combination of the rows of $R'R$.

$$(R'R)v_k = \lambda_k^2 v_k, \qquad (RR')u_k = \lambda_k^2 u_k, \qquad k = 1, \cdots, r.$$

Recall the construction of principal components for a data matrix with observations as rows and variables as columns. Then the v_k's can be viewed as the principal component directions of R, when the columns in R are treated as variables and row as items of observations; while the u_k's can be viewed as the principal component directions of R', when the rows in R are treated as variables and columns as observations.

Correspondence analysis aims to produce a plot with the row variable and column variable on an equal footing. The plots have several versions.

CA plots

In correspondence analysis plots, the row and column variables are plotted in two aligned coordinate systems formed in the spaces of (v_1, v_2) and (u_1, u_2) in the same plot. In order for the variables to be comparable, the eigenvectors

are rescaled. Different type of CA plots scale the (v_1, v_2) and (u_1, u_2) differently. Examples of common scalings are *symmetric bi-plot*, *row principal* axes, or *column principal* axes. The scalings are obtained from the following steps.

$$\begin{aligned} U \quad &\rightarrow \quad D_r^{-1/2}U \quad \rightarrow \quad D_r^{-1/2}U\Sigma = F \\ V \quad &\rightarrow \quad D_c^{-1/2}V \quad \rightarrow \quad D_c^{-1/2}V\Sigma' = G \end{aligned}$$

For example, for a symmetric bi-plot, the CA coordinates for row variables come from the first two columns of $F = D_r^{-1/2}U\Sigma$, which are from the first two column vectors in U; and the CA coordinates for column variables come from the first two columns of $G = D_c^{-1/2}V\Sigma'$, which are from first two column vectors in V.

By the definition of inertia,

$$\text{Total inertia } = \frac{\chi^2}{n} = n\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-rc^T)^2}{r_i c_j} \propto \sum_{k=1}^{r}\lambda_k^2$$

Recall that total inertia is used as a measure of the amount of association between the row and column variables. The larger the inertia, the more strongly associated the row and column variables. In the plot, the percentages of association or variation "explained" by the first and second axes are computed as

$$\frac{\lambda_1^2}{\sum_{k=1}^{r}\lambda_k^2}, \quad \frac{\lambda_2^2}{\sum_{k=1}^{r}\lambda_k^2}$$

respectively.

Note: Relevant section in Johnson and Wichern: Section 12.7.