Notes on Statistical and Machine Learning

Correspondence Analysis

Chapter: *34* **Prepared by:** *Chenxi Zhou*

This note is prepared based on *Chapter 16, Correspondence Analysis* in Izenman (2009).

I. Introduction

1. **Correspondence Analysis:** *Correspondence analysis* is an exploratory multivariate technique for simultaneously displaying scores representing the row categories and column categories of a two-way contingency table as the coordinates of points in a low-dimensional vector space.

   The objectives are

   (a) to clarify the relationship between the row and column variables of the table, and

   (b) to discover a low-dimensional explanation for possible deviations from independence of those variables.

2. **Categories:**

   (a) For two-way contingency tables, correspondence analysis is known as *simple* correspondence analysis;

   (b) For three-way and higher contingency tables, it is known as *multiple* correspondence analysis.

   We focus on simple correspondence analysis.

3. **Applicability:** Correspondence analysis is applicable

   (a) when the variables are *discrete* with many categories; and

   (b) when the variables are continuous and can be segmented into a finite number of ranges.

   Remark. Discretization of a continuous variable usually entails some loss of information.

II. Simple Correspondence Analysis

1. **Two-way Contingency Table:** A *two-way $r \times c$ contingency table* with r rows (labeled as $A_1, A_2, \cdots, A_r$) and c columns (labeled $B_1, B_2, \cdots, B_c$) has $r \times c$ cells. The

	Column Variable						
Row Variable	B_1	B_2	$\cdots$	B_j	$\cdots$	B_c	Row Total
A_1	$n_{1,1}$	$n_{1,2}$	$\cdots$	$n_{1,j}$	$\cdots$	$n_{1,c}$	$n_{1,\bullet}$
A_2	$n_{2,1}$	$n_{2,2}$	$\cdots$	$n_{2,j}$	$\cdots$	$n_{2,c}$	$n_{2,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_i	$n_{i,1}$	$n_{i,2}$	$\cdots$	$n_{i,j}$	$\cdots$	$n_{i,c}$	$n_{i,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	$n_{r,1}$	$n_{r,2}$	$\cdots$	$n_{r,j}$	$\cdots$	$n_{r,c}$	$n_{r,\bullet}$
Column Total	$n_{\bullet,1}$	$n_{\bullet,2}$	$\cdots$	$n_{\bullet,j}$	$\cdots$	$n_{\bullet,c}$	n

Table 1: Two-way contingency table, showing the observed cell frequencies, row and column marginal totals, and total sample size.

(i,j)-th cell has the entry $n_{i,j}$, representing the observed frequency in row category A_i and column category B_j, for all $i = 1, 2, \cdots, r$ and $j = 1, 2, \cdots, c$.

In addition, a two-way contingency table shows the following quantities

(a) the i-th marginal row total is $n_{i,\bullet} := \sum_{j=1}^{c} n_{i,j}$, for all $i = 1, 2, \cdots, r$;

(b) the j-th marginal column total is $n_{\bullet,j} := \sum_{i=1}^{r} n_{i,j}$, for all $j = 1, 2, \cdots, c$; and

(c) $n := \sum_{i=1}^{r} \sum_{j=1}^{c} n_{i,j}$ is the total sample size.

An example of a two-way contingency table is shown in Table 1.

Remark 1. Such a contingency table is also called a *correspondence table*.

Remark 2. For interpretation purposes, it is important to distinguish

(a) when the n individuals are randomly selected from a very large population, or

(b) when they actually constitute the entire population of interest.

2. **Marginal and Cell Probabilities:** Let

(a) $\pi_{i,j}$ be the probability that an individual has the properties A_i and B_j, for all $i = 1, 2, \cdots, r$ and $j = 1, 2, \cdots, c$;

(b) $\pi_{i,\bullet} := n_{i,\bullet}/n$ be the probability that an individual has the property A_i, for all $i = 1, 2, \cdots, r$; and

(c) $\pi_{\bullet,j} := n_{\bullet,j}/n$ be the probability that an individual has the property B_j, for all $j = 1, 2, \cdots, c$.

Remark. The quantities $\{\pi_{i,j}\}_{i=1,2,\cdots,r; j=1,2,\cdots,c}$, $\{\pi_{i,\bullet}\}_{i=1,2,\cdots,r}$, and $\{\pi_{\bullet,j}\}_{j=1,2,\cdots,c}$ are all population quantities, but *not* the estimates obtained from samples, unless the n individuals constitute the entire population.

3. **Row and Column Dummy Variables:** Let $\mathbf{x}_u := (x_{u,1}, x_{u,2}, \cdots, x_{u,r})^\top \in \mathbb{R}^r$ be a binary vector indicating which of row category the u-th individual belongs to, i.e.,

$$x_{u,i} := \begin{cases} 1, & \text{if the } u\text{-th individual belongs to } A_i \\ 0, & \text{otherwise,} \end{cases}$$

for all $u = 1, 2, \cdots, n$. Similarly, let $\mathbf{y}_v := (y_{v,1}, y_{v,2}, \cdots, y_{v,c})^\top \in \mathbb{R}^c$ be a binary vector indicating which of column category the v-th individual belongs to, i.e.,

$$y_{v,j} := \begin{cases} 1, & \text{if the } v\text{-th individual belongs to } B_j \\ 0, & \text{otherwise,} \end{cases}$$

for all $v = 1, 2, \cdots, n$. Up to a column permutation, these binary vectors can be collected into the following two matrices

$$\mathbf{X} := \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{r \times n},$$

and

$$\mathbf{Y} := \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{c \times n},$$

where each column of $\mathbf{X}$ corresponds to $\mathbf{x}_u$, and each column of $\mathbf{Y}$ corresponds to $\mathbf{y}_v$ for all $u, v = 1, 2, \cdots, n$.

Then, the matrix $\mathbf{X}\mathbf{Y}^\top \in \mathbb{R}^{r \times c}$ reproduces the observed cell frequencies of the contingency table

$$\mathbf{X}\mathbf{Y}^\top = \begin{pmatrix} n_{1,1} & n_{1,2} & \cdots & n_{1,c} \\ n_{2,1} & n_{2,2} & \cdots & n_{2,c} \\ \vdots & \vdots & \ddots & \vdots \\ n_{r,1} & n_{r,2} & \cdots & n_{r,c} \end{pmatrix} =: \mathbf{N}. \tag{1}$$

The matrices $\mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{r \times r}$ and $\mathbf{Y}\mathbf{Y}^\top \in \mathbb{R}^{c \times c}$ are both diagonal, with $\mathbf{X}\mathbf{X}^\top$ having the r marginal row totals as diagonal entries and $\mathbf{Y}\mathbf{Y}^\top$ having the c marginal column totals as diagonal entries; that is,

$$\mathbf{X}\mathbf{X}^\top = \operatorname{diag}(n_{1,\bullet}, n_{2,\bullet}, \cdots, n_{r,\bullet}),$$
$$\mathbf{Y}\mathbf{Y}^\top = \operatorname{diag}(n_{\bullet,1}, n_{\bullet,2}, \cdots, n_{\bullet,c}).$$

4. **Burt Matrix:** We define the following block matrix

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}^\top = \begin{pmatrix} n\mathbf{D}_r & \mathbf{N} \\ \mathbf{N}^\top & n\mathbf{D}_c \end{pmatrix} \in \mathbb{R}^{(r+c)\times(r+c)}, \tag{2}$$

where

$$\mathbf{D}_r := \frac{1}{n}\mathbf{X}\mathbf{X}^\top, \qquad \text{and} \qquad \mathbf{D}_c := \frac{1}{n}\mathbf{Y}\mathbf{Y}^\top.$$

The matrix (2) is called the *Burt matrix* for a two-way contingency table.

Property: Burt matrix is symmetric and positive semi-definite.

5. **Correspondence Matrix:** The matrix $\mathbf{P} := \frac{1}{n}\mathbf{N} \in \mathbb{R}^{r\times c}$ is called a *correspondence matrix.*

Remark. If the n individuals constitute a random sample, the entry, $p_{i,j} := n_{i,j}/n$, in the i-th row and j-th column of $\mathbf{P}$ can be characterized the maximum likelihood estimator of $\pi_{i,j}$.

6. **Row and Column Profiles:** The *row profile* of $\mathbf{N}$, denoted by $\mathbf{P}_r \in \mathbb{R}^{r\times c}$, consists of the rows of $\mathbf{N}$ divided by their corresponding row totals, and can be computed as the regression coefficient matrix of $\mathbf{Y}$ on $\mathbf{X}$; that is,

$$\mathbf{P}_r := (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{Y}^\top = \mathbf{D}_r^{-1}\mathbf{P} = \begin{pmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_r^\top, \end{pmatrix} \in \mathbb{R}^{r\times c},$$

where

$$\mathbf{a}_i^\top := \left(\frac{n_{i,1}}{n_{i,\bullet}}, \frac{n_{i,2}}{n_{i,\bullet}}, \cdots, \frac{n_{i,c}}{n_{i,\bullet}}\right) \in \mathbb{R}^c, \qquad \text{for all } i = 1, 2, \cdots, r.$$

Similarly, the *column profile* of $\mathbf{N}$, denoted by $\mathbf{P}_c \in \mathbb{R}^{c\times r}$, consists of the columns of $\mathbf{N}$ divided by their corresponding column totals, and can be computed as the regression coefficient matrix of $\mathbf{X}$ on $\mathbf{Y}$; that is,

$$\mathbf{P}_c = (\mathbf{Y}\mathbf{Y}^\top)^{-1}\mathbf{Y}\mathbf{X}^\top = \mathbf{D}_c^{-1}\mathbf{P}^\top = \begin{pmatrix} \mathbf{b}_1^\top \\ \mathbf{b}_2^\top \\ \vdots \\ \mathbf{b}_c^\top, \end{pmatrix} \in \mathbb{R}^{c\times r},$$

where

$$\mathbf{b}_j^\top := \left(\frac{n_{1,j}}{n_{\bullet,j}}, \frac{n_{2,j}}{n_{\bullet,j}}, \cdots, \frac{n_{r,j}}{n_{\bullet,j}}\right) \in \mathbb{R}^r, \qquad \text{for all } j = 1, 2, \cdots, c.$$

7. **Row and Column Means:** The *row means* of the contingency table $\mathbf{N}$ are the row sums of $\mathbf{P}$

$$\mathbf{P}\mathbf{1}_c = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_r \end{pmatrix} = \begin{pmatrix} n_{1,\bullet}/n \\ n_{2,\bullet}/n \\ \vdots \\ n_{r,\bullet}/n \end{pmatrix} = \begin{pmatrix} p_{1,\bullet} \\ p_{2,\bullet} \\ \vdots \\ p_{r,\bullet} \end{pmatrix} =: \mathbf{r} \in \mathbb{R}^r.$$

Similarly, the *column means* of $\mathbf{N}$ are the column sums of $\mathbf{P}$, or equivalently, row sums of $\mathbf{P}^\top$,

$$\mathbf{P}^\top\mathbf{1}_r = \begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ \vdots \\ \bar{Y}_c \end{pmatrix} = \begin{pmatrix} n_{\bullet,1}/n \\ n_{\bullet,2}/n \\ \vdots \\ n_{\bullet,c}/n \end{pmatrix} = \begin{pmatrix} p_{\bullet,1} \\ p_{\bullet,2} \\ \vdots \\ p_{\bullet,c} \end{pmatrix} =: \mathbf{c} \in \mathbb{R}^c.$$

Remark 1. The vectors $\mathbf{r}$ and $\mathbf{c}$ can also be formed from the diagonal elements of $\mathbf{D}_r$ and $\mathbf{D}_c$, respectively; that is,

$$\mathbf{D}_r = \operatorname{diag}(\mathbf{r}), \qquad \text{and} \qquad \mathbf{D}_c = \operatorname{diag}(\mathbf{c}).$$

Remark 2. In correspondence analysis, $\mathbf{r}$ is called the *average column profile* and $\mathbf{c}$ is called the *average row profile* of the contingency table.

8. **Row and Columns Masses:** The i-th element of the vector $\mathbf{r} \in \mathbb{R}^r$, $p_{i,\bullet} := n_{i,\bullet}/n$, is called the *$i$-th row mass*, for all $i = 1, 2, \cdots, r$.

   Similarly, the j-th element of the $\mathbf{c} \in \mathbb{R}^c$, $p_{\bullet,j} := n_{\bullet,j}/n$, is called the *$j$-th column mass*, for all $j = 1, 2, \cdots, c$.

   Under random sampling,

   - $p_{i,\bullet}$ is an estimate of the unconditional probability of belonging to A_i, $\pi_{i,\bullet}$; and
   - $p_{\bullet,j}$ is an estimate of the unconditional probability of belonging to B_j, $\pi_{\bullet,j}$.

9. **Row and Column Centroids:** The vector $\mathbf{c} \in \mathbb{R}^c$ is also referred to as the *row centroid*, because it can be expressed as the weighted average of the row profiles, that is,

$$\mathbf{c} = \sum_{i=1}^{r} p_{i,\bullet}\mathbf{a}_i,$$

where the weights are the row masses.

Similarly, the vector $\mathbf{r} \in \mathbb{R}^r$ is referred to as the *column centroid*, because it can be expressed as the weighted average of the column profiles

$$\mathbf{r} = \sum_{j=1}^{c} p_{\bullet,j}\mathbf{b}_j,$$

where the weights are the column masses.

10. **Relationship between r and c:** The relationship between $\mathbf{r}$ and $\mathbf{c}$ is given by

$$\mathbf{r} = \mathbf{P}\mathbf{D}_c^{-1}\mathbf{c}, \qquad \text{and} \qquad \mathbf{c} = \mathbf{P}^\top\mathbf{D}_r^{-1}\mathbf{r}.$$

Proof. The results are obvious by noting $\mathbf{D}_c^{-1}\mathbf{c} = \mathbf{1}_c$ and $\mathbf{D}_r^{-1}\mathbf{r} = \mathbf{1}_r$. ∎

11. **Centered Row and Column Profiles:**

(a) *Centered Row Profile:* Let $\mathbf{c} \in \mathbb{R}^c$ be the row centroid. The *centered row profile matrix* is

$$\mathbf{P}_r - \mathbf{1}_r\mathbf{c}^\top \in \mathbb{R}^{r\times c}, \tag{3}$$

where $\mathbf{P}_r := \mathbf{D}_r^{-1}\mathbf{P}$. In particular, the i-th row of (3) is $(\mathbf{a}_i - \mathbf{c})^\top$.

(b) *Centered Column Profile:* Let $\mathbf{r} \in \mathbb{R}^r$ be the column centroid. The *centered column profile matrix* is

$$\mathbf{P}_c - \mathbf{1}_c\mathbf{r}^\top \in \mathbb{R}^{c\times r}, \tag{4}$$

where $\mathbf{P}_c := \mathbf{D}_c^{-1}\mathbf{P}^\top$. In particular, the j-th row of (4) is $(\mathbf{b}_j - \mathbf{r})^\top$.

12. **Row Distances:**

(a) *Squared χ^2-distance Between Two Row Profiles:* Consider the i-th and i'-th row profiles, $\mathbf{a}_i \in \mathbb{R}^c$ and $\mathbf{a}_{i'} \in \mathbb{R}^c$, respectively. Note that the j-th entry of $\mathbf{a}_i - \mathbf{a}_{i'}$ is

$$\frac{n_{i,j}}{n_{i,\bullet}} - \frac{n_{i',j}}{n_{i',\bullet}}.$$

The *squared χ^2-distance* between $\mathbf{a}_i$ and $\mathbf{a}_{i'}$ is defined as the quadratic form

$$\begin{aligned} d^2(\mathbf{a}_i, \mathbf{a}_{i'}) &:= (\mathbf{a}_i - \mathbf{a}_{i'})^\top\mathbf{D}_c^{-1}(\mathbf{a}_i - \mathbf{a}_{i'}) \\ &= \sum_{j=1}^{c} \frac{n}{n_{\bullet,j}}\left(\frac{n_{i,j}}{n_{i,\bullet}} - \frac{n_{i',j}}{n_{i',\bullet}}\right)^2. \end{aligned}$$

Remark. Note that the inverse of the j-th column mass, $n/n_{\bullet,j}$, enters the squared χ^2-distance above. Hence, the categories having fewer observations contribute more to the inter-row profile distances.

(b) *Squared χ^2-distance to Row Centroid:* Let $\mathbf{c} \in \mathbb{R}^c$ be the row centroid defined earlier. The *squared χ^2-distance* between $\mathbf{a}_i$ and $\mathbf{c}$ is

$$\begin{aligned} d^2(\mathbf{a}_i, \mathbf{c}) &= (\mathbf{a}_i - \mathbf{c})^\top\mathbf{D}_c^{-1}(\mathbf{a}_i - \mathbf{c}) \\ &= \frac{1}{n_{i,\bullet}}\sum_{j=1}^{c} \frac{n}{n_{i,\bullet}n_{\bullet,j}}\left(n_{i,j} - \frac{n_{i,\bullet}n_{\bullet,j}}{n}\right)^2. \end{aligned}$$

(c) *Connection with Pearson's χ^2-statistic:* If we sum $d^2(\mathbf{a}_i, \mathbf{c})$ over all $i = 1, 2, \cdots, r$ with the weight $np_{i,\bullet}$, we have

$$n \sum_{i=1}^{r} p_{i,\bullet} d^2(\mathbf{a}_i, \mathbf{c}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(n_{i,j} - \frac{n_{i,\bullet} n_{\bullet,j}}{n}\right)^2 \bigg/ \left(\frac{n_{i,\bullet} n_{\bullet,j}}{n}\right),$$

which is the *Pearson's χ^2 statistics*

$$\chi^2 := \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}},$$

with

- $O_{i,j} := n_{i,j}$ being the observed cell frequency, and
- $E_{i,j} := n_{i,\bullet} n_{\bullet,j}/n$ being the expected cell frequency (assuming the independence of row and column variables),

for all $i = 1, 2, \cdots, r$ and $j = 1, 2, \cdots, s$.

Approximate Distribution: Under random sampling, for large n, χ^2 has approximately the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

13. Column Distances:

(a) *Squared χ^2-distance Between Two Column Profiles:* Define the squared χ^2-distance between the j-th and j'-th column profiles, $\mathbf{b}_j$ and $\mathbf{b}_{j'}$, as the following quadratic form

$$\begin{aligned} d^2(\mathbf{b}_j, \mathbf{b}_{j'}) &= (\mathbf{b}_j - \mathbf{b}_{j'})^\top \mathbf{D}_r^{-1} (\mathbf{b}_j - \mathbf{b}_{j'}) \\ &= \sum_{i=1}^{r} \frac{n}{n_{i,\bullet}} \left(\frac{n_{i,j}}{n_{\bullet,j}} - \frac{n_{i,j'}}{n_{\bullet,j'}}\right)^2. \end{aligned}$$

(b) *Squared χ^2-distance to Column Centroid:* The *squared χ^2-distance between the j-th column profile and the column centroid* is

$$\begin{aligned} d^2(\mathbf{b}_j, \mathbf{r}) &= (\mathbf{b}_j - \mathbf{r})^\top \mathbf{D}_r^{-1} (\mathbf{b}_j - \mathbf{r}) \\ &= \frac{1}{n_{\bullet,j}} \sum_{i=1}^{r} \frac{n}{n_{i,\bullet} n_{\bullet,j}} \left(n_{i,j} - \frac{n_{i,\bullet} n_{\bullet,j}}{n}\right)^2. \end{aligned}$$

(c) *Connection with Pearson's χ^2-statistic:* If we sum $d^2(\mathbf{b}_j, \mathbf{r})$ over all $j = 1, 2, \cdots, c$ with the weight $np_{\bullet,j}$, we have

$$n \sum_{j=1}^{c} p_{\bullet,j} d^2(\mathbf{b}_j, \mathbf{r}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(n_{i,j} - \frac{n_{i,\bullet} n_{\bullet,j}}{n}\right)^2 \bigg/ \left(\frac{n_{i,\bullet} n_{\bullet,j}}{n}\right),$$

which is again Pearson's chi-squared statistic.

14. **Test of Independence in a Contingency Table:** We are interested in testing whether row and column variables in a two-way contingency table are independent or not.

    (a) *Intuition:* If row and column variables are indeed independent, we expect

    $$n_{i,j} \approx n_{i,\bullet} \times n_{\bullet,j}, \qquad \text{for all } i = 1, 2, \cdots, r \text{ and } j = 1, 2, \cdots, c.$$

    Otherwise, we expect to see large deviation between $n_{i,j}$ and the product of $n_{i,\bullet}$ and $n_{\bullet,j}$.

    (b) *Hypothesis Statement:* We formulate the problem of interest as

    $$H_0 : \text{ Row and column variables are independent}$$

    against

    $$H_1 : \text{ Row and column variables are } \textit{not} \text{ independent.}$$

    (c) *Test Statistic and Asymptotic Distribution:* We use Pearson's chi-squared statistic, χ^2. For large n, χ^2 approximately follows a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.
    We reject H_0 if $\chi^2 > \chi^2_{(r-1)(c-1),1-\alpha}$, where $\chi^2_{(r-1)(c-1),1-\alpha}$ is the $(1-\alpha) \cdot 100\%$ percentile of a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

15. **Matrix of Residuals:** Consider the observed cell frequency matrix $\mathbf{N} \in \mathbb{R}^{r \times c}$ defined in (1). Define the *matrix of residuals*, denoted by $\widetilde{\mathbf{N}}$, as

    $$\widetilde{\mathbf{N}} := \mathbf{N} - n\mathbf{r}\mathbf{c}^\top.$$

    In particular, note that the (i,j)-th entry of $\widetilde{\mathbf{N}}$, denoted by $\tilde{n}_{i,j}$, is given by

    $$\tilde{n}_{i,j} := [\widetilde{\mathbf{N}}]_{i,j} = n_{i,j} - E_{i,j} = n_{i,j} - \frac{n_{i,\bullet} n_{\bullet,j}}{n}.$$

    Remark 1. The standard assumption of the contingency table analysis is that the row and column totals are considered fixed and the cell frequencies in $\mathbf{N}$ are allowed to vary within those constraints. Hence, $\widetilde{\mathbf{N}}$ is a centered version of $\mathbf{N}$ by centering the elements of the latter at the values we expect them to have under independence.

    Remark 2. The matrix $\widetilde{\mathbf{N}}$ is called the *matrix of residuals* because its (i,j)-th entry, $\tilde{n}_{i,j} = O_{i,j} - E_{i,j}$, shows the difference between the observed cell frequency $(O_{i,j})$ and its expected cell frequency $(E_{i,j})$, assuming independence between row and column variables, for all $i = 1, 2, \cdots, r$ and $j = 1, 2, \cdots, s$.

    Remark 3. Since

    $$\widetilde{\mathbf{N}}\mathbf{1}_c = (\mathbf{N} - n\mathbf{r}\mathbf{c}^\top)\mathbf{1}_c = \mathbf{N}\mathbf{1}_c - n\mathbf{r}\mathbf{c}^\top\mathbf{1}_c = n\mathbf{r} - n\mathbf{r} = \mathbf{0}_r,$$

    the rank of $\widetilde{\mathbf{N}}$ is at most $c - 1$.

16. **Relative Frequency Matrix:** Define the *relative frequency matrix* as

$$\widetilde{\mathbf{P}} := \frac{1}{n}\widetilde{\mathbf{N}} = \frac{1}{n}\mathbf{X}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y}^\top = \mathbf{P} - \mathbf{r}\mathbf{c}^\top,$$

where $\mathbf{J}_n \in \mathbb{R}^{n\times n}$ is the matrix with all entries equal to 1.

Remark. Similar to $\widetilde{\mathbf{N}}$, the rank of $\widetilde{\mathbf{P}}$ is at most $c-1$ as well.

17. **An Alternative Expression of Pearson's χ^2 Statistic:** Define the following matrix

$$\mathbf{R} := \mathbf{D}_c^{-\frac{1}{2}}\widetilde{\mathbf{P}}^\top\mathbf{D}_r^{-1}\widetilde{\mathbf{P}}\mathbf{D}_c^{-\frac{1}{2}}. \tag{5}$$

The (j, j')-th entry of $\mathbf{R}$, where $j \neq j'$, is given by

$$\frac{1}{\sqrt{n_{\bullet,j}n_{\bullet,j'}}}\sum_{i=1}^{r}\frac{1}{n_{i,\bullet}}\left(n_{i,j} - \frac{n_{i,\bullet}n_{\bullet,j}}{n}\right)\left(n_{i,j'} - \frac{n_{i,\bullet}n_{\bullet,j'}}{n}\right),$$

and the j-th diagonal element of $\mathbf{R}$ is

$$\frac{1}{n_{\bullet,j}}\sum_{i=1}^{r}\frac{1}{n_{i,\bullet}}\left(n_{i,j} - \frac{n_{i,\bullet}n_{\bullet,j}}{n}\right)^2.$$

The trace of $\mathbf{R}$, which is also the sum of eigenvalues of $\mathbf{R}$, is

$$\sum_{j=1}^{c}\lambda_j = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{1}{n_{i,\bullet}n_{\bullet,j}}\left(n_{i,j} - \frac{n_{i,\bullet}n_{\bullet,j}}{n}\right)^2 = \frac{\chi^2}{n}, \tag{6}$$

where $\lambda_1, \lambda_2, \cdots, \lambda_c$ are eigenvalues of $\mathbf{R}$, and χ^2 is the Pearson's chi-squared statistic.

18. **Total Inertia:** The quantity χ^2/n is referred to as the amount of *total inertia* in the contingency table.

Moreover, the eigenvalues of $\mathbf{R}$ form a decomposition of the total inertia. The accumulated contribution of the first t principal inertias is given by

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_t}{\lambda_1 + \lambda_2 + \cdots + \lambda_c},$$

which is an analogue of the percentage of total variance explained by the first t principal components.

19. **Decomposition of R:** We can decompose the matrix $\mathbf{R}$ in (5) as

$$\mathbf{R} = \mathbf{M}^\top\mathbf{M},$$

where $\mathbf{M} := \mathbf{D}_r^{-\frac{1}{2}}\widetilde{\mathbf{P}}\mathbf{D}_c^{-\frac{1}{2}} \in \mathbb{R}^{r\times c}$ with its (i, j)-th entry being *Pearson's residual*

$$m_{i,j} := [\mathbf{M}]_{i,j} = \frac{1}{\sqrt{n_{i,\bullet}n_{\bullet,j}}}\left(n_{i,j} - \frac{n_{i,\bullet}n_{\bullet,j}}{n}\right),$$

for all $i = 1, 2, \cdots, r$ and $j = 1, 2, \cdots, c$.

Remark 1. From (6), the total inertia χ^2/n is the sum of squares of all rc Pearson residuals in the contingency table.

Remark 2. Since $\mathrm{rank}(\mathbf{P}) \le c - 1$, it follows that $\mathbf{M}$ also has rank at most $c - 1$.

20. Singular Value Decomposition of M and Consequences: The singular value decomposition of $\mathbf{M}$ is given by

$$\mathbf{M} = \mathbf{U}\mathbf{D}_\lambda\mathbf{V}^\top,$$

where

- $\mathbf{U} \in \mathbb{R}^{r\times c}$ satisfies $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_c$, with columns being the eigenvectors corresponding to the matrix
$$\mathbf{M}\mathbf{M}^\top = \mathbf{D}_r^{-\frac{1}{2}}\widetilde{\mathbf{P}}\mathbf{D}_c^{-1}\widetilde{\mathbf{P}}^\top\mathbf{D}_r^{-\frac{1}{2}} =: \mathbf{R}_1,$$
- $\mathbf{V} \in \mathbb{R}^{c\times c}$ satisfies $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_c$, with columns being the eigenvectors corresponding to the matrix $\mathbf{R}$, and
- $\mathbf{D}_\lambda := \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_c}) \in \mathbb{R}^{c\times c}$ is a diagonal matrix with its principal diagonal having the singular values.

Using the notation above, we have

$$\widetilde{\mathbf{P}} = (\mathbf{D}_r^{\frac{1}{2}}\mathbf{U})\mathbf{D}_\lambda(\mathbf{V}^\top\mathbf{D}_c^{\frac{1}{2}}) = \mathbf{A}\mathbf{D}_\lambda\mathbf{B}^\top, \tag{7}$$

where $\mathbf{A} := \mathbf{D}_r^{\frac{1}{2}}\mathbf{U}$ and $\mathbf{B} := \mathbf{D}_c^{\frac{1}{2}}\mathbf{V}$.

Noticing that

$$\mathbf{A}^\top\mathbf{D}_r^{-1}\mathbf{A} = \mathbf{I}_c, \qquad \text{and} \qquad \mathbf{B}^\top\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{I}_c,$$

we call (7) the *generalized singular value decomposition* of $\widetilde{\mathbf{P}}$ in the matrices $\mathbf{D}_r^{-1}$ and $\mathbf{D}_c^{-1}$. The columns of $\mathbf{A}$ and $\mathbf{B}$ are called the *principal axes* of the row and column profiles, respectively.

21. Principal Coordinates of Row and Column Profiles:

(a) *Principal Coordinates of Row Profiles:* The squared χ^2-distance (in the metric $\mathbf{D}_c^{-1}$) between the centered row profile matrix $\mathbf{P}_r - \mathbf{1}_r\mathbf{c}^\top$ and $\mathbf{B}$ is given by

$$\begin{aligned}
\mathbf{G}_P^\top :&= (\mathbf{P}_r - \mathbf{1}_r\mathbf{c}^\top)\mathbf{D}_c^{-1}\mathbf{B} \\
&= (\mathbf{D}_r^{-1}\widetilde{\mathbf{P}}\mathbf{D}_c^{-1})\mathbf{B} \\
&= \mathbf{D}_r^{-1}\mathbf{A}\mathbf{D}_\lambda\mathbf{B}^\top\mathbf{D}_c^{-1}\mathbf{B} \\
&= \mathbf{D}_r^{-1}\mathbf{A}\mathbf{D}_\lambda \\
&= \mathbf{D}_r^{-1}\mathbf{D}_r^{\frac{1}{2}}\mathbf{U}\mathbf{D}_\lambda \\
&= \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}\mathbf{D}_\lambda.
\end{aligned}$$

The columns of $\mathbf{G}_P^\top$ is called the *principal coordinates of the row profiles*.

(b) *Principal Coordinates of Column Profiles:* The squared χ^2-distance (in the metric $\mathbf{D}_r^{-1}$) between the centered column profile matrix $\mathbf{P}_c - \mathbf{1}_c\mathbf{r}^\top$ and $\mathbf{A}$ is given by

$$\mathbf{H}_P^\top := (\mathbf{P}_c - \mathbf{1}_c\mathbf{r}^\top)\mathbf{D}_r^{-1}\mathbf{A} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}\mathbf{D}_\lambda,$$

by a similar derivation. The columns of $\mathbf{H}_P^\top$ are called the *principal coordinates of the column profiles.*

(c) *Relationships between $\mathbf{G}_P^\top$ and $\mathbf{H}_P^\top$:* Using the notation above, we have

$$\mathbf{G}_P^\top = \mathbf{D}_r^{-1}\mathbf{P}\mathbf{H}_P^\top\mathbf{D}_\lambda^{-1}, \qquad \text{and} \qquad \mathbf{H}_P^\top = \mathbf{D}_c^{-1}\mathbf{P}^\top\mathbf{G}_P^\top\mathbf{D}_\lambda^{-1}.$$

22. Correspondence Map:

(a) *Procedure:*

(1) Make a scatterplot of each of the r rows of the first two (or three) columns of $\mathbf{G}_P^\top$;

(2) On the same scatterplot, plot each of the c rows of the first two (or three) columns of $\mathbf{H}_P^\top$.

The resulting scatterplot consisting of $(r + c)$ points is called a *correspondence map.*

(b) *Recommendations:*

i. For clearer interpretation, different symbols should be used for the row points and column points;

ii. It is useful to identify each point in the plot by a tag showing its corresponding category name;

iii. If row or column categories are ordered in some way, it is visually helpful to connect those category points to indicate such order dependence.

(c) *Interpretation:*

i. If row points are close, then those rows have similar conditional distributions across columns;

ii. If column points are close, then those columns have similar conditional distributions across rows;

iii. If a row point is close to a column point, then that configuration suggests a particular deviation from independence.

References

Izenman, Alan J (Mar. 2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* en. Springer Science & Business Media.