SME0822 Análise Multivariada e Aprendizado Não-Supervisionado

Aula 12a: **Análise de correspondência**

Prof. Cibele Russo

cibele@icmc.usp.br

http://www.icmc.usp.br/~cibele

Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Prentice Hall.

Mingoti, S. A. (2007) Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Editora UFMG.

Análise de correspondência

Objetivo:

A análise de correspondência é um procedimento gráfico para representar associações em uma tabela de frequências ou contagens.

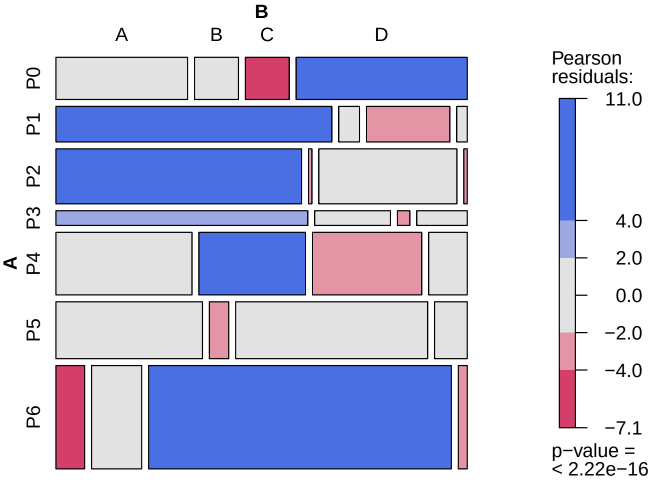
Se a tabela de contingências possui I linhas e J colunas, os gráficos possuem I pontos correspondentes às linhas e J pontos correspondentes às colunas.

Análise de correspondência

Motivação: A tabela a seguir contém frequências de $J = 4$ tipos de cerâmica encontrados em $I = 7$ sítios arqueológicos no sudeste dos Estados Unidos.

Sítio \ Tipo	A	B	C	D	Total
P0	30	10	10	39	89
P1	53	4	16	2	75
P2	73	1	41	1	116
P3	20	6	1	4	31
P4	46	36	37	13	132
P5	45	6	59	10	120
P6	16	28	169	5	218
Total	283	91	333	74	781

Análise de correspondência



Análise de correspondência

Seja X a matriz correspondente à tabela $I \times J$ com frequências x_{ij}.
Considere que $I > J$ e X de posto completo.
Se n é o total de frequências na matriz de dados X, primeiro construímos a matriz de proporções

$P = \{p_{ij}\}$, em que

$$p_{ij} = \frac{x_{ij}}{n}, \; i = 1, 2, \ldots, I, \; j = 1, 2, \ldots, J$$

$$P = \frac{1}{n} X$$

P é chamada a matriz de correspondência.

Análise de correspondência

Para o exemplo de motivação, temos

$$X = \begin{pmatrix} 30 & 10 & 10 & 39 \\ 53 & 4 & 16 & 2 \\ 73 & 1 & 41 & 1 \\ 20 & 6 & 1 & 4 \\ 46 & 36 & 37 & 13 \\ 45 & 6 & 59 & 10 \\ 16 & 28 & 169 & 5 \end{pmatrix} \text{ e } P = \begin{pmatrix} 0.04 & 0.01 & 0.01 & 0.05 \\ 0.07 & 0.01 & 0.02 & 0.00 \\ 0.09 & 0.00 & 0.05 & 0.00 \\ 0.03 & 0.01 & 0.00 & 0.01 \\ 0.06 & 0.05 & 0.05 & 0.02 \\ 0.06 & 0.01 & 0.08 & 0.01 \\ 0.02 & 0.04 & 0.22 & 0.01 \end{pmatrix}$$

Análise de correspondência

Depois, defina os vetores de somas das linhas e colunas como $\underset{\sim}{r}$ e $\underset{\sim}{c}$, respectivamente e D_r e D_c as matrizes diagonais com elementos das linhas e colunas na diagonal principal.

Além disso, para $i = 1, \ldots, I$ e $j = 1, \ldots, J$:

$$r_i = \sum_{j=1}^{J} p_{ij} = \sum_{j=1}^{J} \frac{x_{ij}}{n}$$

$$c_j = \sum_{i=1}^{I} p_{ij} = \sum_{i=1}^{I} \frac{x_{ij}}{n}$$

Análise de correspondência

No exemplo,

$$\underset{\sim}{r} = \begin{pmatrix} 0.114 \\ 0.096 \\ 0.147 \\ 0.040 \\ 0.169 \\ 0.155 \\ 0.278 \end{pmatrix} \text{ e } \underset{\sim}{c} = \begin{pmatrix} 0.362 \\ 0.117 \\ 0.425 \\ 0.095 \end{pmatrix}$$

Análise de correspondência

Temos

$$\underset{\sim}{r} = P\underset{\sim}{1}_J$$
$$\underset{\sim}{c} = P^\top\underset{\sim}{1}_I$$
$$D_r = \text{diag}\{r_1, \dots r_I\}$$
$$D_c = \text{diag}\{c_1, \dots, c_J\}$$
$$D_r^{1/2} = \text{diag}\{\sqrt{r_1}, \dots \sqrt{r_I}\}$$
$$D_c^{1/2} = \text{diag}\{\sqrt{c_1}, \dots, \sqrt{c_J}\}$$

Análise de correspondência

A análise de correspondência pode ser formulada como um problema de mínimos quadrados para selecionar

$$\widehat{P} = \{\widehat{p_{ij}}\}$$

uma matriz de posto reduzido para minimizar

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-\widehat{p_{ij}})^2}{r_i c_j} = tr\left\{[D_r^{-1/2}(P-\widehat{P})D_c^{-1/2}][D_r^{-1/2}(P-\widehat{P})D_c^{-1/2}]^\top\right\}\star$$

já que $\dfrac{(p_{ij}-\widehat{p_{ij}})}{\sqrt{r_i c_j}}$ é o i-ésimo elemento de $D_r^{-1/2}(P-\widehat{P})D_c^{-1/2}$.

Análise de correspondência

Decomposição em valores singulares (SVD)

Para uma matriz A de dimensão $p \times q$, digamos $p > q$ sem perda de generalidade, a decomposição em valores singulares (SVD) de A é dada por

$$A = U\Lambda V^\top$$

com

- Λ uma matriz retangular diagonal $p \times q$, com os autovalores de $A^\top A$ na diagonal principal da submatriz $q \times q$ de Λ e zero no restante das entradas
- U é uma matriz $p \times p$ que contém autovetores da matriz $AA^\top$ em suas colunas
- V é uma matriz $q \times q$ que contém autovetores da matriz $A^\top A$ em suas colunas

Análise de correspondência

Decomposição em valores singulares (SVD)

Ver mais, por exemplo, em

`https://en.wikipedia.org/wiki/Singular_value_decomposition`

Análise de correspondência

Resultado

O termo $\underset{\sim}{r}\underset{\sim}{c}^\top$ *é comum para aproximar* $\widehat{P}$, *não importa qual a matriz* $P_{I\times J}$. *A aproximação de posto reduzido para* P *que minimiza a soma de quadrados* $\star$ *é dada por*

$$P = \sum_{k=1}^{s} \tilde{\lambda}_k (D_r^{-1/2}\underset{\sim}{\tilde{u}}_k)(D_c^{-1/2}\underset{\sim}{\tilde{v}}_k)^\top = \underset{\sim}{r}\underset{\sim}{c}^\top + \sum_{k=2}^{s} \tilde{\lambda}_k (D_r^{-1/2}\underset{\sim}{\tilde{u}}_k)(D_c^{-1/2}\underset{\sim}{\tilde{v}}_k)^\top$$

em que $\tilde{\lambda}_k$ *são os valores singulares e* $\underset{\sim}{\tilde{u}}_k$ *e* $\underset{\sim}{\tilde{v}}_k$ *são os vetores singulares correspondentes da matriz* $D_r^{-1/2} P D_c^{-1/2}$. *O valor mínimo de* $\star$ *é* $\sum_{k=s+1}^{J} \tilde{\lambda}_k^2$.

Análise de correspondência

Resultado

O termo $\underset{\sim}{r}\underset{\sim}{c}^\top$ *é comum para aproximar* $\widehat{P}$*, não importa qual a matriz* $P_{I\times J}$*.*
A aproximação de posto reduzido para P *que minimiza a soma de quadrados* $\star$ *é dada por*

$$P = \sum_{k=1}^{s} \tilde{\lambda}_k (D_r^{-1/2}\underset{\sim}{\tilde{u}}_k)(D_c^{-1/2}\underset{\sim}{\tilde{v}}_k)^\top = \underset{\sim}{r}\underset{\sim}{c}^\top + \sum_{k=2}^{s} \tilde{\lambda}_k (D_r^{-1/2}\underset{\sim}{\tilde{u}}_k)(D_c^{-1/2}\underset{\sim}{\tilde{v}}_k)^\top$$

em que $\tilde{\lambda}_k$ *são os valores singulares e* $\underset{\sim}{\tilde{u}}_k$ *e* $\underset{\sim}{\tilde{v}}_k$ *são os vetores singulares correspondentes da matriz* $D_r^{-1/2} P D_c^{-1/2}$*. O valor mínimo de* $\star$ *é*
$\displaystyle\sum_{k=s+1}^{J} \tilde{\lambda}_k^2$*.*

Análise de correspondência

No exemplo,

$$\underset{\sim}{r}\underset{\sim}{c}^\top = \begin{pmatrix} 0.04 & 0.01 & 0.05 & 0.01 \\ 0.03 & 0.01 & 0.04 & 0.01 \\ 0.05 & 0.02 & 0.06 & 0.01 \\ 0.01 & 0.00 & 0.02 & 0.00 \\ 0.06 & 0.02 & 0.07 & 0.02 \\ 0.06 & 0.02 & 0.07 & 0.01 \\ 0.10 & 0.03 & 0.12 & 0.03 \end{pmatrix}$$

Análise de correspondência

A aproximação de posto reduzido $K > 1$ para $P - \underset{\sim}{r}\underset{\sim}{c}^\top$ é

$$P - \underset{\sim}{r}\underset{\sim}{c}^\top = \sum_{k=1}^{K} \lambda_k (D_r^{1/2}\underset{\sim}{u}_k)(D_c^{1/2}\underset{\sim}{v}_k)^\top$$

em que λ_k são valores singulares e os vetores $\underset{\sim}{u}_k$ e $\underset{\sim}{v}_k$ são os vetores singulares correspondentes da matriz

$$D_r^{-1/2}(P - \underset{\sim}{r}\underset{\sim}{c}^\top)D_c^{-1/2}.$$

Aqui, $\lambda_k = \tilde{\lambda}_{k+1}, \quad \underset{\sim}{u}_k = \tilde{\underset{\sim}{u}}_{k+1} \quad \underset{\sim}{v}_k = \tilde{\underset{\sim}{v}}_{k+1}$ para $k = 1, \ldots, J-1$.
Prova em Johnson (2007), p; 720.

Análise de correspondência

Obs: Note que os vetores $D_r^{1/2}\underset{\sim}{u}_k$ e $D_c^{1/2}\underset{\sim}{v}_k$ em

$$P - \underset{\sim}{r}\underset{\sim}{c}^\top = \sum_{k=1}^{K} \lambda_k (D_r^{1/2}\underset{\sim}{u}_k)(D_c^{1/2}\underset{\sim}{v}_k)^\top$$

não precisam ser de tamanho 1 mas devem satisfazer

$$(D_r^{1/2}\underset{\sim}{u}_k)^\top D_r^{-1}(D_r^{1/2}\underset{\sim}{u}_k) = \underset{\sim}{u}_k^\top \underset{\sim}{u}_k = 1 \text{ e}$$
$$(D_c^{1/2}\underset{\sim}{v}_k)^\top D_r^{-1}(D_c^{1/2}\underset{\sim}{v}_k) = \underset{\sim}{v}_k^\top \underset{\sim}{v}_k = 1.$$

Análise de correspondência

Considere a matriz $\tilde{P} = P - \underset{\sim}{r}\underset{\sim}{c}^\top$.
Decompondo $\tilde{P}$ em valores singulares

$$\tilde{P} = A\Lambda B^\top$$

em que

- $A = D_r^{1/2} U$,
- $B = D_c^{1/2} V$,
- Λ contém os autovalores de $\tilde{P}^\top \tilde{P}$ (ordenados de forma decrescente),
- U contém os autovetores de $\tilde{P}\tilde{P}^\top$,
- V contém os autovetores de $\tilde{P}^\top \tilde{P}$.

Análise de correspondência

Coordenadas principais das linhas

As coordenadas principais das linhas da matriz $\tilde{P}$ são definidas como

$$Y = D_r^{-1} A \Lambda$$

Coordenadas principais das colunas

As coordenadas principais das colunas da matriz $\tilde{P}$ são definidas como

$$Z = D_c^{-1} B \Lambda$$

É comum considerar as duas primeiras coordenadas principais para representar a associação entre as quantidades originais nas linhas e colunas da matriz de dados.

Análise de correspondência

Inércia total

A inércia total, ou variação total existente nos dados é representada por

$$\sum_{k=1}^{K} \lambda_k^2.$$

Pode-se mostrar que a inércia total está relacioanda com a estatística qui-quadrado da seguinte forma

$$\sum_{k=1}^{K} \lambda_k^2 = \frac{\chi^2}{n} = \frac{1}{n} \sum_{i,j} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}.$$

em que n_{ij} representa a frequência observada na linha i e coluna j e E_{ij} a frequência esperada correspondente.

Análise de correspondência