ME731 - Métodos em Análise Multivariada – Análise de Correspondência I –

Prof. Carlos Trucíos
ctrucios@unicamp.br
ctruciosm.github.io

Instituto de Matemática, Estatística e Computação Científica,
Universidade Estadual de Campinas

Aula 17

Agenda I

Introdução

Introdução

- Até agora vimos ACP e AF, que lidam apenas com variáveis **quantitativas**.

Introdução

- Até agora vimos ACP e AF, que lidam apenas com variáveis **quantitativas**.
- E se o interesse estiver em variáveis **qualitativas**?

Introdução

- Até agora vimos ACP e AF, que lidam apenas com variáveis **quantitativas**.
- E se o interesse estiver em variáveis **qualitativas**?
- Na aula de hoje apresentaremos **Análise de Correspondência**, uma técnica multivariada que nos permite lidar com variáveis categóricas.

Introdução

- Até agora vimos ACP e AF, que lidam apenas com variáveis **quantitativas**.
- E se o interesse estiver em variáveis **qualitativas**?
- Na aula de hoje apresentaremos **Análise de Correspondência**, uma técnica multivariada que nos permite lidar com variáveis categóricas.
- Para muitos, análise de correspondência é uma caixa preta (mas não para nós)!

Motivação

Motivação

Existe associação entre as variáveis?

Tabela 1: Exemplo 1

sex	Adelie	Chinstrap	Gentoo
female	73	34	58
male	73	34	61

Motivação

Existe associação entre as variáveis?

Tabela 1: Exemplo 1

sex	Adelie	Chinstrap	Gentoo
female	73	34	58
male	73	34	61

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 0.048607, df = 2, p-value = 0.976
```

Motivação

Existe associação entre as variáveis?

Tabela 2: Exemplo 2

Profession	Single	Married	Widower	Divorcee	Remarried
Unskilled worker	242	347	108	72	23
Manual labourer	242	660	84	104	71
Technician	109	218	10	44	20
Foreman	144	424	61	70	36
Management	215	623	58	92	64
Employee	603	1247	263	312	127
Other	54	112	17	18	11

Motivação

```
hobbies %>%
  drop_na() %>%
  select(`Marital status`, Profession) %>%
  table() %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 158.01, df = 24, p-value < 2.2e-16
```

Motivação

- Qual categoria da variável `Marital status` está associada com qual categoria da variável `Profession`?
- Se a Tabela de contingüência tivesse mais do que apenas duas entradas? ($4 \times 3 \times 5$, por exemplo?)

Motivação

- Qual categoria da variável `Marital status` está associada com qual categoria da variável `Profession`?
- Se a Tabela de contingência tivesse mais do que apenas duas entradas? ($4 \times 3 \times 5$, por exemplo?)

Nesses casos utilizamos análise de correspondência!.

Motivação

O principal objetivo da Análise de Correpondência é obter índices que permitam mostrar a relação entre as categorias das linhas e colunas de uma tabela de contingência.

Motivação

O principal objetivo da Análise de Correpondência é obter índices que permitam mostrar a relação entre as categorias das linhas e colunas de uma tabela de contingência.

- Isto é feito de uma forma semelhante ao estudado em ACP.

Motivação

O principal objetivo da Análise de Correpondência é obter índices que permitam mostrar a relação entre as categorias das linhas e colunas de uma tabela de contingência.

- Isto é feito de uma forma semelhante ao estudado em ACP.
- Contudo, desta vez, a decomposição será feita sob uma medida de associação (geralmente o valor χ^2 utilizado no teste de independência).

Análise de Correspondência

Análise de Correspondência

Seja n_{ij} o número de elementos que pertencem à categoria i da variável 1 e à categoria j da variável 2. Então,

		Variável 2					
		Cat 1	Cat 2	Cat 3	$\cdots$	Cat q	**Total**
Variável 1	Cat 1	n_{11}	n_{12}	n_{13}	$\cdots$	n_{1q}	$n_{1.}$
	Cat 2	n_{21}	n_{22}	n_{23}	$\cdots$	n_{2q}	$n_{2.}$
	$\vdots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	
	Cat p	n_{p1}	n_{p2}	n_{p3}	$\cdots$	n_{pq}	$n_{p.}$
Total		$n_{.1}$	$n_{.2}$	$n_{.3}$	$\cdots$	$n_{.q}$	$n_{..} = n$

Análise de Correspondência

A forma tradicional de avaliar se existe alguma relação entre ambas as variáveis é através do teste de independência cuja estatística de teste é dada por

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{(n_{ij} - E_{ij})^2}{E_{ij}},$$

em que $E_{ij} = \dfrac{n_{i.}n_{.j}}{n_{..}}$.

Análise de Correspondência

A forma tradicional de avaliar se existe alguma relação entre ambas as variáveis é através do teste de independência cuja estatística de teste é dada por

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - E_{ij})^2}{E_{ij}},$$

em que $E_{ij} = \dfrac{n_{i.} n_{.j}}{n_{..}}$.

Obs: E_{ij} é o número de casos esperados sob H_0, se não existir associação entre as variáveis, as frequências observadas (n_{ij}) e as esperadas (E_{ij}) estarão próximas.

Análise de Correspondência

Alternativamente, se utilizarmos frequência relativas ($f_{ij} = n_{ij}/n$):

		Variável 2					
		Cat 1	Cat 2	Cat 3	$\cdots$	Cat q	**Total**
Variável 1	Cat 1	f_{11}	f_{12}	f_{13}	$\cdots$	f_{1q}	$f_{1.}$
	Cat 2	f_{21}	f_{22}	f_{23}	$\cdots$	f_{2q}	$f_{2.}$
	$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	
	Cat p	f_{p1}	f_{p2}	f_{p3}	$\cdots$	f_{pq}	$f_{p.}$
Total		$f_{.1}$	$f_{.2}$	$f_{.3}$	$\cdots$	$f_{.q}$	$f_{..} = 1$

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{(n_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{(nf_{ij} - nf_{i.}f_{.j})^2}{nf_{i.}f_{.j}} = n\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

Análise de Correspondência

E se tivermos uma forma de fazer uma decomposição dessa estatística de teste?

Análise de Correspondência

E se tivermos uma forma de fazer uma decomposição dessa estatística de teste? AC fará uma decomposição dessa estatística de teste.

Análise de Correspondência

Seja $c_{ij} = \dfrac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}}$, então faremos uma decomposição SVD da matriz

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1q} \\ c_{21} & c_{22} & \cdots & c_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pq} \end{pmatrix}.$$

Análise de Correspondência

Seja $c_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}}$, então faremos uma decomposição SVD da matriz

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1q} \\ c_{21} & c_{22} & \cdots & c_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p1} & c_{p2} & \cdots & c_{pq} \end{pmatrix}.$$

Note que, se não existir associação entre as variáveis, as frequências observadas (n_{ij}) estarão perto das esperadas (E_{ij}) e, consequentemente, c_{ij} será pequeno.

Análise de Correspondência

Definição SVD

Qualquer matriz $\mathbf{M}_{m\times n}$ de posto r pode ser decomposta como:

$$\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^t,$$

em que $\mathbf{U}_{m\times r}$, $\mathbf{V}_{n\times r}$ e $\mathbf{D}_{r\times r} = Diag\{\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_r}\}$. $\lambda_1, \cdots, \lambda_r$ são os autovalores (em ordem decrescente) de $\mathbf{M}\mathbf{M}'$.

- Os elementos D_{ii} são chamdos de valores singulares da matriz $\mathbf{M}$.
- As colunas de $\mathbf{U}$ são os r autovetores (normalizados) associados a $\mathbf{M}\mathbf{M}'$.
- As colunas de $\mathbf{V}$ são os r autovetores (normalizados) associados a $\mathbf{M}'\mathbf{M}$.

Esta decomposição é chamada de SVD (Singular Value Decomposition).

Análise de Correspondência

Seja $r = rank(\mathbf{C}_{q \times p})$, então pela SVD:

$$\mathbf{C} = \Gamma\Lambda\Delta' = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1r} \\ \vdots & \ddots & \vdots \\ \gamma_{q1} & \cdots & \gamma_{qr} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_r} \end{pmatrix} \begin{pmatrix} \delta_{11} & \cdots & \delta_{p1} \\ \vdots & \ddots & \vdots \\ \delta 1r & \cdots & \delta_{pr} \end{pmatrix}$$

Análise de Correspondência

Seja $r = rank(\mathbf{C}_{q\times p})$, então pela SVD:

$$\mathbf{C} = \Gamma\Lambda\Delta' = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1r} \\ \vdots & \ddots & \vdots \\ \gamma_{q1} & \cdots & \gamma_{qr} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_r} \end{pmatrix} \begin{pmatrix} \delta_{11} & \cdots & \delta_{p1} \\ \vdots & \ddots & \vdots \\ \delta 1r & \cdots & \delta_{pr} \end{pmatrix}$$

Então,

$$c_{ij} = \sum_{k=1}^{r} \sqrt{\lambda_k}\gamma_{ik}\delta_{jk} \qquad (1)$$

Análise de Correspondência

Então, utilizando a definição de traço de uma matriz temos que:

$$\sum_{k=1}^{r} \lambda_k = Tr(\mathbf{CC}') = \sum_{i=1}^{p}\sum_{j=1}^{q} c_{ij}^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = \chi^2$$

Análise de Correspondência

Então, utilizando a definição de traço de uma matriz temos que:

$$\sum_{k=1}^{r} \lambda_k = Tr(\mathbf{C}\mathbf{C}') = \sum_{i=1}^{p}\sum_{j=1}^{q} c_{ij}^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = \chi^2$$

Aplicar SVD em $\mathbf{C}$ decompoe a estatística de teste χ^2.

Análise de Correspondência

Por outro lado, as projeções das linhas e colunas de $\mathbf{C}$ nos hiperplanos com vetor direção dados pelas colunas de $\mathbf{U}$ e $\mathbf{V}$ são dadas por

$$\mathbf{C}\gamma_k \quad e \quad \mathbf{C}'\delta_k.$$

Análise de Correspondência

Por outro lado, as projeções das linhas e colunas de **C** nos hiperplanos com vetor direção dados pelas colunas de **U** e **V** são dadas por

$$\mathbf{C}\gamma_k \quad e \quad \mathbf{C}'\delta_k.$$

Existe um Teorema que nos ajudará a relacionar ambas as projeções

Análise de Correspondência

Por outro lado, as projeções das linhas e colunas de $\mathbf{C}$ nos hiperplanos com vetor direção dados pelas colunas de $\mathbf{U}$ e $\mathbf{V}$ são dadas por

$$\mathbf{C}\gamma_k \quad e \quad \mathbf{C}'\delta_k.$$

Existe um Teorema que nos ajudará a relacionar ambas as projeções

Teorema: Relações de Dualidade

Seja r o posto de $\mathbf{M}$. Para $k \leq r$, os autovalores λ_k de $\mathbf{M}'\mathbf{M}$ e $\mathbf{M}\mathbf{M}'$ são os mesmos e os autovetores u_k e v_k tem a seguinte relação

$$u_k = \frac{1}{\sqrt{\lambda_k}}\mathbf{M}'v_k \quad e \quad v_k = \frac{1}{\sqrt{\lambda_k}}\mathbf{M}u_k.$$

Análise de Correspondência

Para $k = 1, \cdots, r$, pelo Teorema de relações de dualidade,

$$\gamma_k = \frac{1}{\sqrt{\lambda_k}} \mathbf{C}' \delta_k \quad e \quad \delta_k = \frac{1}{\sqrt{\lambda_k}} \mathbf{C} \gamma_k$$

Análise de Correspondência

Para $k = 1, \cdots, r$, pelo Teorema de relações de dualidade,

$$\gamma_k = \frac{1}{\sqrt{\lambda_k}}\mathbf{C}'\delta_k \quad e \quad \delta_k = \frac{1}{\sqrt{\lambda_k}}\mathbf{C}\gamma_k$$

Então, as projeções

$$\mathbf{C}\gamma_k \quad e \quad \mathbf{C}'\delta_k$$

podem (utilizando o Teorema de relações de dualidade) ser escritas como

$$\mathbf{C}\gamma_k = \sqrt{\lambda_k}\delta_k \quad e \quad \mathbf{C}'\delta_k = \sqrt{\lambda_k}\gamma_k \tag{2}$$

Análise de Correspondência

Agora suponha que λ_1 é muito maior do que os outros autovalores, então

$$c_{ij} = \sum_{k=1}^{r} \sqrt{\lambda_k}\gamma_{ik}\delta_{jk} \approx \sqrt{\lambda_1}\gamma_{i1}\delta_{j1}$$

Análise de Correspondência

Agora suponha que λ_1 é muito maior do que os outros autovalores, então

$$c_{ij} = \sum_{k=1}^{r} \sqrt{\lambda_k}\gamma_{ik}\delta_{jk} \approx \sqrt{\lambda_1}\gamma_{i1}\delta_{j1}$$

- Se γ_{i1} e δ_{j1} tiverem o mesmo sinal e forem grandes, então c_{ij} também será grande e positivo, indicando uma associação positiva entre a i-éssima linha e j-éssima coluna da Tabela de contingência.
- Se γ_{i1} e δ_{j1} tiverem diferente sinal e forem grandes, então c_{ij} também será grande e negativo, indicando uma associação negativa entre a i-éssima linha e j-éssima coluna da Tabela de contingência.

Análise de Correspondência

Na maioria dos casos, λ_1 e λ_2 são suficientes para obtermos uma boa aproximação para c_{ij}

- Assim, (2) e os autovetores (γ_1, γ_2) podem ser utilizados para obter uma boa representação gráfica das linhas da Tabela de contigência.
- Equivalentemente, (2) e os autovetores (δ_1, δ_2) podem ser utilizados para obter uma boa representação gráfica das colunas da Tabela de contigência.

Análise de Correspondência

Na maioria dos casos, λ_1 e λ_2 são suficientes para obtermos uma boa aproximação para c_{ij}

- Assim, (2) e os autovetores (γ_1, γ_2) podem ser utilizados para obter uma boa representação gráfica das linhas da Tabela de contigência.
- Equivalentemente, (2) e os autovetores (δ_1, δ_2) podem ser utilizados para obter uma boa representação gráfica das colunas da Tabela de contigência.

A proximidade dos pontos representando as linhas e colunas devem ser interpretados como categorias relacionadas entre si.

Análise de Correspondência

Em análise de correspondência projetamos as linhas e colunas de **C** mas ponderadas. Assim, se fizermos

$$\mathbf{A} = \begin{pmatrix} n_{1.} & 0 & \cdots & 0 \\ 0 & n_{2.} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_{p.} \end{pmatrix} \quad e \quad \mathbf{B} = \begin{pmatrix} n_{.1} & 0 & \cdots & 0 \\ 0 & n_{.2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_{.q} \end{pmatrix}$$

Análise de Correspondência

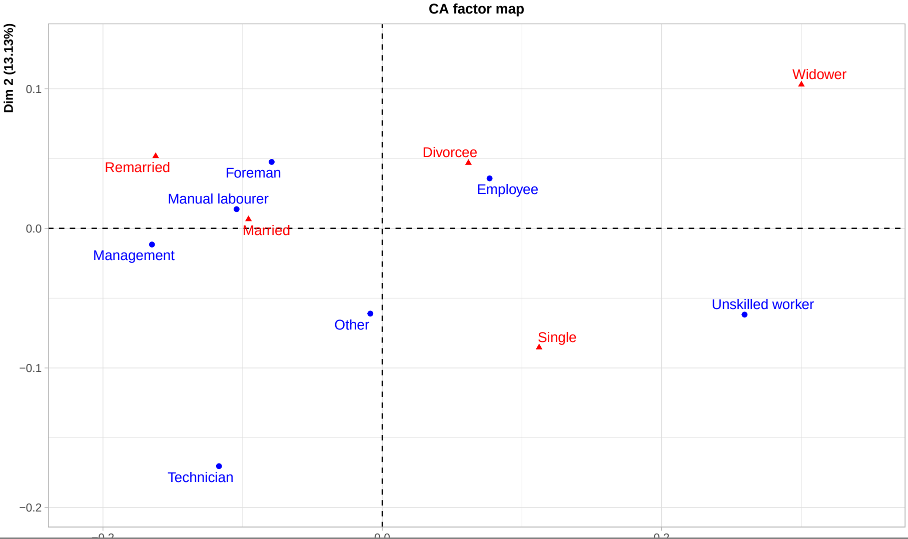
Em análise de correspondência projetamos as linhas e colunas de **C** mas ponderadas. Assim, se fizermos

$$\mathbf{A} = \begin{pmatrix} n_{1.} & 0 & \cdots & 0 \\ 0 & n_{2.} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_{p.} \end{pmatrix} \quad e \quad \mathbf{B} = \begin{pmatrix} n_{.1} & 0 & \cdots & 0 \\ 0 & n_{.2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_{.q} \end{pmatrix}$$

As projeções serão:

$$r_k = \mathbf{A}^{-1/2}\mathbf{C}\delta_k = \sqrt{\lambda_k}\mathbf{A}^{-1/2}\gamma_k \quad e \quad s_k = \mathbf{B}^{-1/2}\mathbf{C}\gamma_k = \sqrt{\lambda_k}\mathbf{B}^{-1/2}\delta_k.$$

Análise de Correspondência



Análise de Correspondência

O que, de fato, estamos fazendo ao projetar $\mathbf{A}^{-1/2}\mathbf{C}$ e $\mathbf{B}^{-1/2}\mathbf{C}$?

Análise de Correspondência

O que, de fato, estamos fazendo ao projetar $\mathbf{A}^{-1/2}\mathbf{C}$ e $\mathbf{B}^{-1/2}\mathbf{C}$?

- Estamos interessados em representar as frequências relativas por linhas (colunas) em um espaço de dimensão pequena que permita apreciar as distâncias relativas entre os pontos (linhas/colunas).

Análise de Correspondência

O que, de fato, estamos fazendo ao projetar $\mathbf{A}^{-1/2}\mathbf{C}$ e $\mathbf{B}^{-1/2}\mathbf{C}$?

- Estamos interessados em representar as frequências relativas por linhas (colunas) em um espaço de dimensão pequena que permita apreciar as distâncias relativas entre os pontos (linhas/colunas).
- A frequenca relativa de cada linha(coluna) é diferente. Ou seja, as linhas (colunas) não tem o mesmo peso, pois algumas contém mais dados do que outras. Quando representarmos cada uma das linhas (colunas) devemos levas isto em consideração.

Análise de Correspondência

O que, de fato, estamos fazendo ao projetar $\mathbf{A}^{-1/2}\mathbf{C}$ e $\mathbf{B}^{-1/2}\mathbf{C}$?

- Estamos interessados em representar as frequências relativas por linhas (colunas) em um espaço de dimensão pequena que permita apreciar as distâncias relativas entre os pontos (linhas/colunas).
- A frequenca relativa de cada linha(coluna) é diferente. Ou seja, as linhas (colunas) não tem o mesmo peso, pois algumas contém mais dados do que outras. Quando representarmos cada uma das linhas (colunas) devemos levas isto em consideração.
- Para saber quão próximas são as linhas (colunas), precisamos de uma medida de distância entre elas.

Análise de Correspondência

Pense na seguinta Tabela de frequências relativas

```
##           A    B    C    D
## Zona A 0.03 0.06 0.15 0.06
## Zona B 0.07 0.14 0.35 0.14
```

Análise de Correspondência

Pense na seguinta Tabela de frequências relativas

```
##             A    B    C    D
## Zona A 0.03 0.06 0.15 0.06
## Zona B 0.07 0.14 0.35 0.14
```

Se calcularmos a distância euclideana entre ambas as linhas, parecerá que as frequências relativas são muito distintas. Contudo, basta condicionarmos por linha para perceber que as frequências não são distintas.

Análise de Correspondência

Pense na seguinte Tabela de frequências relativas

```
##            A    B    C    D
## Zona A 0.03 0.06 0.15 0.06
## Zona B 0.07 0.14 0.35 0.14
```

Se calcularmos a distância euclideana entre ambas as linhas, parecerá que as frequências relativas são muito distintas. Contudo, basta condicionarmos por linha para perceber que as frequências não são distintas.

```
##          A   B   C   D
## Zona A 0.1 0.2 0.5 0.2
## Zona B 0.1 0.2 0.5 0.2
```

Ou seja, em lugar de olhar para a matriz de frequências relativas $\mathbf{F}$, podemos olhar para $\mathbf{R} = \mathbf{D}_f^{-1}\mathbf{F}$ ($\mathbf{D}_f = Diag\{f_{1.}, \cdot, f_{p.}\}$).

Análise de Correspondência

Mas isso não é suficiente para podermos comparar as linhas apropriadamente. Pense no seguinte caso:

Análise de Correspondência

Mas isso não é suficiente para podermos comparar as linhas apropriadamente. Pense no seguinte caso:

```
##         Loiro Ruivo Marrom C Marrom O  Preto
## Verdes  0.435 0.073    0.369    0.119 0.0030
## Azuis   0.454 0.053    0.336    0.153 0.0040
## Marrons 0.193 0.047    0.512    0.232 0.0150
## Preto   0.075 0.037    0.307    0.518 0.0065
```

Os loiros tem uma diferença entre cor de olhos azul e verdes de $0.454 - 0.435 = 0.019$. Por outro lado, pessoas de cabelo preto tem ma diferença entre cor de olhos Marrons e Azuis de $0.015 - 0.004 = 0.011$.

A simples vista, diriamos que diferença do primeiro é maior do que do segundo. Contudo, no segundo caso 0.015 é $\approx 4\times$ 0.004.

Análise de Correspondência

Precisamos levar em consideração a frequência relativa da categoria que estudamos.

- Em categorias raras, pequenas diferenças absolutas podem ser grandes diferenças relativas.
- Em categorias com frequências maiores, a mesma diferença absoluta será menos importante.

Análise de Correspondência

Precisamos levar em consideração a frequência relativa da categoria que estudamos.

- Em categorias raras, pequenas diferenças absolutas podem ser grandes diferenças relativas.
- Em categorias com frequências maiores, a mesma diferença absoluta será menos importante.

Uma forma de se fazer isto é ponderar as diferências de forma inversamente proporcional à categoria de interesse. Assim, em lugar de fazermos $(f_{ij}/f_{i.} - f_{kj}/f_{k.})^2$, faremos $(f_{ij}/f_{i.} - f_{kj}/f_{k.})^2/f_{.j}$

Análise de Correspondência

Podemos comparar cada linha (coluna) w.r.t valores médios. Isto é exatamente o que $\mathbf{A}^{-1/2}\mathbf{C}$ e $\mathbf{B}^{-1/2}\mathbf{C}$ fazem!

Análise de Correspondência

Podemos comparar cada linha (coluna) w.r.t valores médios. Isto é exatamente o que $\mathbf{A}^{-1/2}\mathbf{C}$ e $\mathbf{B}^{-1/2}\mathbf{C}$ fazem!

Note que os elementos de $\mathbf{A}^{-1/2}\mathbf{C}$ e $\mathbf{B}^{-1/2}\mathbf{C}$ são, respectivamente:

$$\frac{n_{ij} - \frac{n_{i.}n_{.j}}{n}}{\sqrt{n_{i.}}\sqrt{\frac{n_{i.}n_{.j}}{n}}} = \frac{\frac{n_{ij}}{n_{i.}} - \frac{n_{.j}}{n}}{\sqrt{\frac{n_{.j}}{n}}} = \frac{f_{ij}/f_{i.} - f_{.j}}{\sqrt{f_{.j}}} \quad e \quad \frac{f_{ij}/f_{.j} - f_{i.}}{\sqrt{f_{i.}}}$$

Implementação

Implementação

```
N <- hobbies %>% drop_na() %>%
  select(`Marital status`, Profession) %>%
  table() %>% as.matrix() %>% t()
N
```

```
##                    Marital status
## Profession          Single Married Widower Divorcee Remarried
##   Unskilled worker     242     347     108       72        23
##   Manual labourer      242     660      84      104        71
##   Technician           109     218      10       44        20
##   Foreman              144     424      61       70        36
##   Management           215     623      58       92        64
##   Employee             603    1247     263      312       127
##   Other                 54     112      17       18        11
```

Implementação

```
tot_linha <- rowSums(N)
tot_coluna <- colSums(N)
E <- tot_linha %o% tot_coluna / sum(N) # outer multiplication
round(E, 4)
```

```
##                    Single   Married  Widower Divorcee Remarrie
## Unskilled worker 184.5515  416.4739  68.9344  81.6660   40.37
## Manual labourer  270.5357  610.5128 101.0516 119.7150   59.18
## Technician        93.4408  210.8662  34.9024  41.3486   20.44
## Foreman          171.2694  386.5004  63.9732  75.7886   37.46
## Management       245.1366  553.1951  91.5644 108.4756   53.62
## Employee         594.6659 1341.9713 222.1219 263.1461  130.09
## Other             49.4001  111.4804  18.4521  21.8601   10.80
```

Implementação

```
C <- (N - E) / sqrt(E)
svd_decomposition <- svd(C)
Gama <- svd_decomposition$u
Lambda <- diag(svd_decomposition$d)
Delta <- svd_decomposition$v
A <- diag(tot_linha)
B <- diag(tot_coluna)
r_1 <- Lambda[1,1] * sqrtm(solve(A)) %*% matrix(Gama[,1], ncol =
r_2 <- Lambda[2,2] * sqrtm(solve(A)) %*% matrix(Gama[,2], ncol =
s_1 <- Lambda[1,1] * sqrtm(solve(B)) %*% matrix(Delta[,1], ncol =
s_2 <- Lambda[2,2] * sqrtm(solve(B)) %*% matrix(Delta[,2], ncol =
```

Implementação

```
correspondencia <- FactoMineR::CA(N, ncp = 2, graph = FALSE)
cbind(correspondencia$col$coord, s_1, s_2)
```

```
##                 Dim 1        Dim 2
## Single     0.11229699 -0.085297255 -0.11229699 -0.085297255
## Married   -0.09579981  0.006535596  0.09579981  0.006535596
## Widower    0.30009996  0.103029272 -0.30009996  0.103029272
## Divorcee   0.06169403  0.046876693 -0.06169403  0.046876693
## Remarried -0.16228113  0.051749256  0.16228113  0.051749256
```

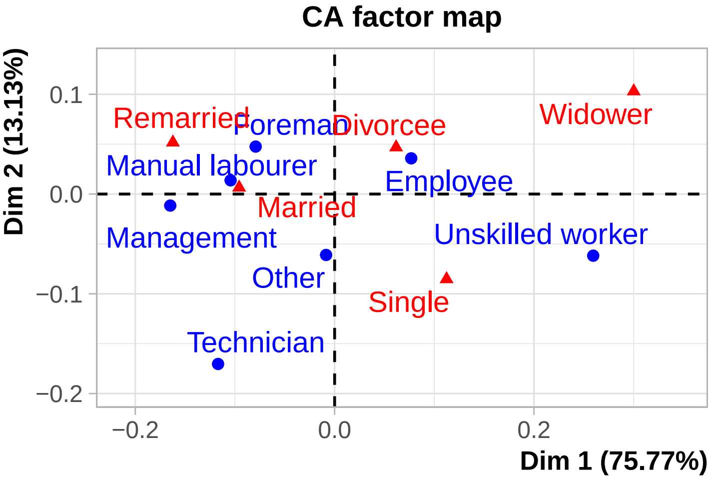
Implementação

```r
round(cbind(correspondencia$row$coord, r_1, r_2), 4)
```

```
##                    Dim 1   Dim 2
## Unskilled worker  0.2594 -0.0618 -0.2594 -0.0618
## Manual labourer  -0.1043  0.0138  0.1043  0.0138
## Technician       -0.1169 -0.1704  0.1169 -0.1704
## Foreman          -0.0792  0.0476  0.0792  0.0476
## Management       -0.1649 -0.0116  0.1649 -0.0116
## Employee          0.0768  0.0358 -0.0768  0.0358
## Other            -0.0085 -0.0611  0.0085 -0.0611
```

Implementação: Gráfico de correspondência

```
plot(correspondencia)
```



Interpretação

Interpretação

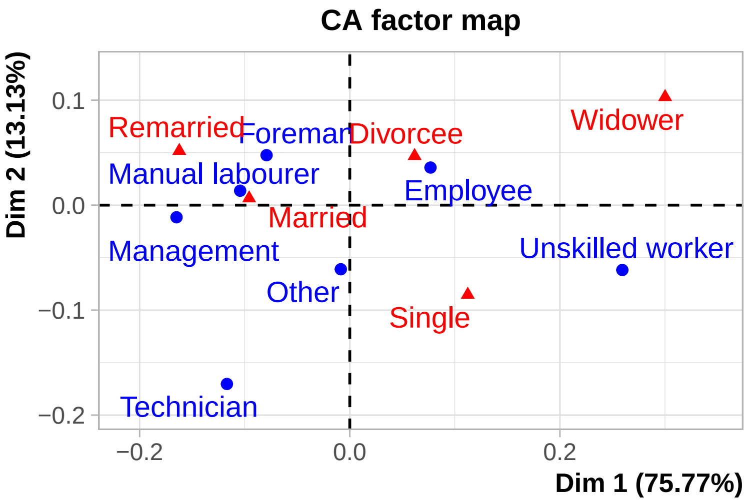
Interpretaç ao do Gráfico de correspondência:

- Pontos linha (ou seja, aquela obtidos dos (r_1, r_2)) que estão próximos, indicam que essas linhas da tabela de contingência tem perfis[1] semelhantes.
- Pontos coluna (ou seja, aquela obtidos dos (s_1, s_2)) que estão próximos, indicam que essas colunas da tabela de contingência tem perfis[2] semelhantes.
- Pontos linha e pontos coluna que estão próximos entre si (mas afastados da origem) indicam que essas categorias estão associadas.
- Como a origem é o centro dos fatores, pontos projetados perto da origem indicam perfis próximos do *perfil médio*

[1] Perfis linha são obtidos pelo cociente $n_{i,j}/n_{i.}$

[2] Perfis coluna são obtidos pelo cociente $n_{i,j}/n_{.j}$

Interpretação



Interpretação

Pode-se provar que:

- $\bar{r}_k = 0$ e $\bar{s}_k = 0$
- $Var(r_k) = \dfrac{\lambda_k}{n}$ e $Var(s_k) = \dfrac{\lambda_k}{n}$

Interpretação

Pode-se provar que:

- $\bar{r}_k = 0$ e $\bar{s}_k = 0$
- $Var(r_k) = \dfrac{\lambda_k}{n}$ e $Var(s_k) = \dfrac{\lambda_k}{n}$

Isto implica que,

$$Var(r_k)/Var(r_1 + \cdots + r_r) = \frac{\lambda_k}{\lambda_1 + \cdots + \lambda_r} = Var(s_k)/Var(s_1 + \cdots + s_r)$$

.

- Esta quantidade pode ser interpretado como a proporção de variância explicada pelo k-éssimo *fator*.
- Note que essa quantidade também nos diz quanto da estatística χ^2 foi recuperada pelo k-éssimo *fator*.

Referências

Referências

- Härdle, W. K., & Simar, L. (2019). Applied Multivariate Statistical Analysis. Fifth Editon. Springer Nature. Capítulo 15.
- Jhonson, R. & Wichern, D. (2007). Applied Multivariate Statistical Analysis. Sixth Edition. Person. Capítulo 12.7
- Mingoti, S. (2007). Análise de Dados Através de Métodos de Estatística Multivariada: Uma abordagem aplicada. Editora UFMG. Capítulo 8.