Análisis de Correspondencia (CA)

José A. Perusquía Cortés

Análisis Multivariado Semestre 2023-2

¿De qué va?

- Una técnica multivariada para analizar las asociaciones entre un conjunto de variables categóricas de forma gráfica (reducción de la dimensión).

¿De qué va?

- Una técnica multivariada para analizar las asociaciones entre un conjunto de variables categóricas de forma gráfica (reducción de la dimensión).

- Es una técnica meramente descriptiva conocida desde Hirschfeld (1935) y redescubierta e impulsada por Jean-Paul Benzécri en Francia en los años 60's

¿De qué va?

- Una técnica multivariada para analizar las asociaciones entre un conjunto de variables categóricas de forma gráfica (reducción de la dimensión).

- Es una técnica meramente descriptiva conocida desde Hirschfeld (1935) y redescubierta e impulsada por Jean-Paul Benzécri en Francia en los años 60's.

- Técnica similar a PCA pero para datos categóricos.

Principales Elementos

- Punto en un espacio multidimensional

Principales Elementos

- Punto en un espacio multidimensional
- Un peso (o masa) asignado a cada punto

Principales Elementos

- Punto en un espacio multidimensional
- Un peso (o masa) asignado a cada punto
- Una función de distancia entre puntos: chi-squared distance
  - Para dos renglones i, i'

$$d(i,i') = \sqrt{\sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2 \cdot \frac{1}{f_{\cdot j}}}$$

Principales Elementos

- Punto en un espacio multidimensional
- Un peso (o masa) asignado a cada punto
- Una función de distancia entre puntos: chi-squared distance
  - Para dos renglones i, i'

$$d(i,i') = \sqrt{\sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2 \cdot \frac{1}{f_{\cdot j}}}$$

  - Para dos columnas j, j'

$$d(j,j') = \sqrt{\sum_{i=1}^{n} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)^2 \cdot \frac{1}{f_{i\cdot}}}$$

Ejemplo 1

- Estado de salud por grupo de edades

Grupo\Salud	Muy Bueno	Bueno	Regular	Malo	Muy Malo	Totales Ren.
16-24	243	789	167	18	6	**1223**
25-34	220	809	164	35	6	**1234**
35-44	147	658	181	41	8	**1035**
45-54	90	469	236	50	16	**861**
55-64	53	414	306	106	30	**909**
65-74	44	267	284	98	20	**713**
75+	20	136	157	66	17	**396**
Totales Col.	**817**	**3542**	**1495**	**414**	**103**	**6371**

- Tabla de frecuencias por renglón

Grupo\Salud	Muy Bueno	Bueno	Regular	Malo	Muy Malo	Promedio
16-24	0.199	0.645	0.137	0.015	0.005	**0.192**
25-34	0.178	0.656	0.133	0.028	0.005	**0.194**
35-44	0.142	0.636	0.175	0.040	0.008	**0.162**
45-54	0.105	0.545	0.274	0.058	0.019	**0.135**
55-64	0.058	0.455	0.337	0.117	0.033	**0.143**
65-74	0.062	0.374	0.398	0.137	0.028	**0.112**
75+	0.051	0.343	0.396	0.167	0.043	**0.062**
Promedio	**0.128**	**0.556**	**0.235**	**0.065**	**0.016**	**1.000**

Ejemplo 1

- Puntos en un espacio multidimensional: perfiles por renglón

Grupo\Salud	Muy Bueno	Bueno	Regular	Malo	Muy Malo	Promedio
16-24	0.199	0.645	0.137	0.015	0.005	**0.192**
25-34	0.178	0.656	0.133	0.028	0.005	**0.194**
35-44	0.142	0.636	0.175	0.040	0.008	**0.162**
45-54	0.105	0.545	0.274	0.058	0.019	**0.135**
55-64	0.058	0.455	0.337	0.117	0.033	**0.143**
65-74	0.062	0.374	0.398	0.137	0.028	**0.112**
75+	0.051	0.343	0.396	0.167	0.043	**0.062**
Promedio	**0.128**	**0.556**	**0.235**	**0.065**	**0.016**	**1.000**

- Pesos (masas) de cada perfil: promedios por renglón

Grupo\Salud	Muy Bueno	Bueno	Regular	Malo	Muy Malo	Promedio
16-24	0.199	0.645	0.137	0.015	0.005	**0.192**
25-34	0.178	0.656	0.133	0.028	0.005	**0.194**
35-44	0.142	0.636	0.175	0.040	0.008	**0.162**
45-54	0.105	0.545	0.274	0.058	0.019	**0.135**
55-64	0.058	0.455	0.337	0.117	0.033	**0.143**
65-74	0.062	0.374	0.398	0.137	0.028	**0.112**
75+	0.051	0.343	0.396	0.167	0.043	**0.062**
Promedio	**0.128**	**0.556**	**0.235**	**0.065**	**0.016**	**1.000**

- La distancia chi-squared definida por: promedios por columna

Grupo\Salud	Muy Bueno	Bueno	Regular	Malo	Muy Malo	Promedio
16-24	0.199	0.645	0.137	0.015	0.005	**0.192**
25-34	0.178	0.656	0.133	0.028	0.005	**0.194**
35-44	0.142	0.636	0.175	0.040	0.008	**0.162**
45-54	0.105	0.545	0.274	0.058	0.019	**0.135**
55-64	0.058	0.455	0.337	0.117	0.033	**0.143**
65-74	0.062	0.374	0.398	0.137	0.028	**0.112**
75+	0.051	0.343	0.396	0.167	0.043	**0.062**
Promedio	**0.128**	**0.556**	**0.235**	**0.065**	**0.016**	**1.000**

- Definir matrices diagonales $\mathbf{D}_r$ y $\mathbf{D}_c$ con las masas por renglón y columna.

- Definir matrices diagonales $\mathbf{D}_r$ y $\mathbf{D}_c$ con las masas por renglón y columna.

- Obtener la descomposición GSVD de $\mathbf{R} - \mathbf{1c}^T$, i.e.,

$$\mathbf{R} - \mathbf{1c} = \mathbf{N}\Lambda\mathbf{M}^T \qquad \mathbf{N}^T\mathbf{D}_r\mathbf{N} = \mathbf{M}^T\mathbf{D}_c^{-1}\mathbf{M} = \mathbf{I}$$

  - $\mathbf{R}$ es la matriz de perfiles por renglón
  - $\mathbf{c} = \mathbf{D}_c\mathbf{1}$ es el centroide

- Definir matrices diagonales $\mathbf{D}_r$ y $\mathbf{D}_c$ con las masas por renglón y columna.

- Obtener la descomposición GSVD de $\mathbf{R} - \mathbf{1c}^T$, i.e.,

$$\mathbf{R} - \mathbf{1c} = \mathbf{N}\Lambda\mathbf{M}^T \qquad \mathbf{N}^T\mathbf{D}_r\mathbf{N} = \mathbf{M}^T\mathbf{D}_c^{-1}\mathbf{M} = \mathbf{I}$$

  - $\mathbf{R}$ es la matriz de perfiles por renglón
  - $\mathbf{c} = \mathbf{D}_c\mathbf{1}$ es el centroide

- Las primeras dos coordenadas se encuentran con $\mathbf{N}_{(2)}\Lambda_{(2)}$

Ejemplo 1

- Las matrices diagonales son:

$$\mathbf{D}_r = diag(.192,.194,.162,.135,.143,.112,.062)$$

$$\mathbf{D}_c = diag(.128,.556,.235,.065,.016)$$

Ejemplo 1

- Las matrices diagonales son:

$$\mathbf{D}_r = diag(.192,.194,.162,.135,.143,.112,.062) \qquad \mathbf{D}_c = diag(.128,.556,.235,.065,.016)$$

- El centroide es

$$\mathbf{c} = (.128,.556,.235,.065,.016)^T$$

Ejemplo 1

- Las matrices diagonales son:

$$\mathbf{D}_r = diag(.192,.194,.162,.135,.143,.112,.062) \qquad \mathbf{D}_c = diag(.128,.556,.235,.065,.016)$$

- El centroide es

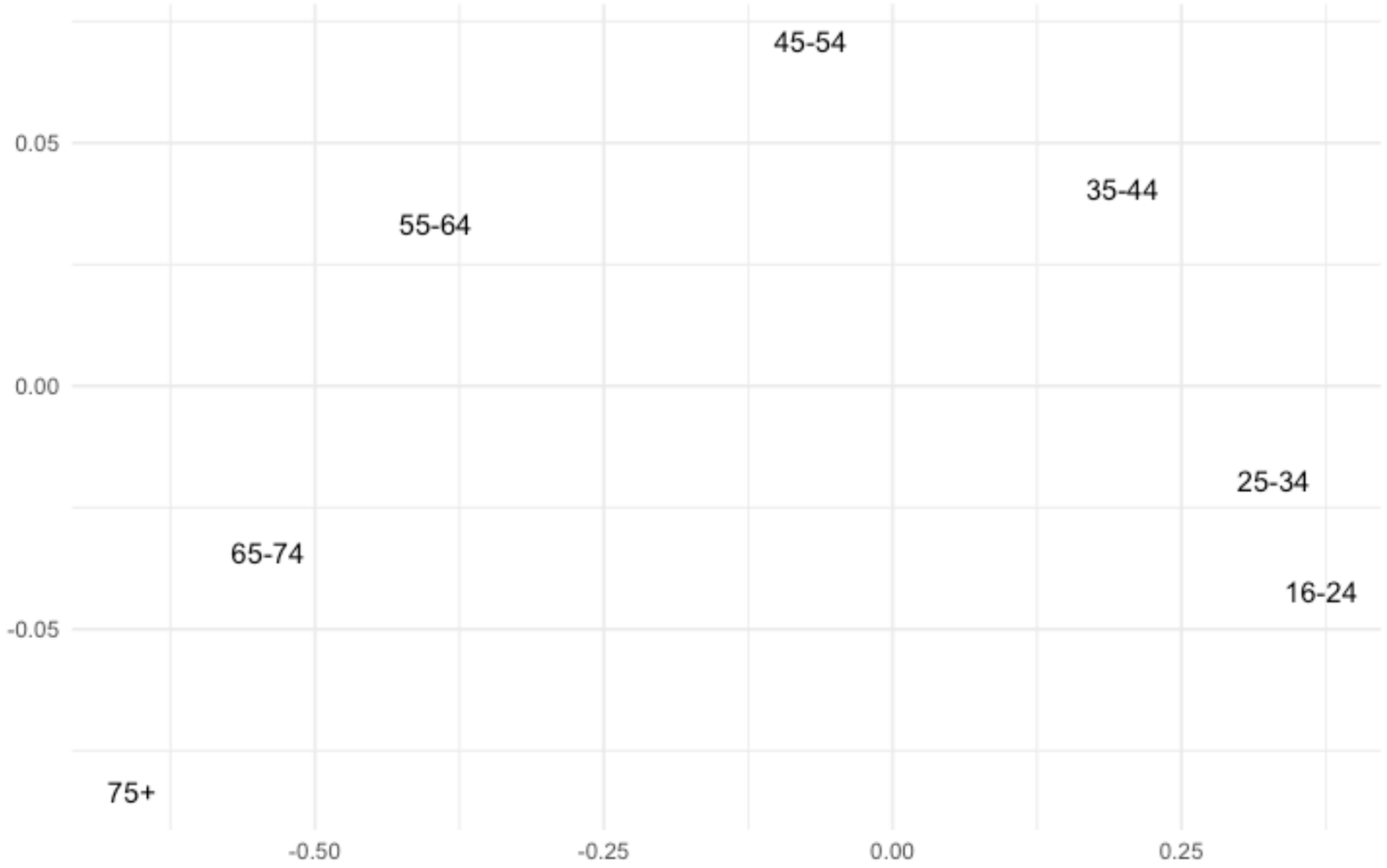
$$\mathbf{c} = (.128,.556,.235,.065,.016)^T$$

- $\mathbf{R} - \mathbf{1}\mathbf{c}^T$ está dado por:

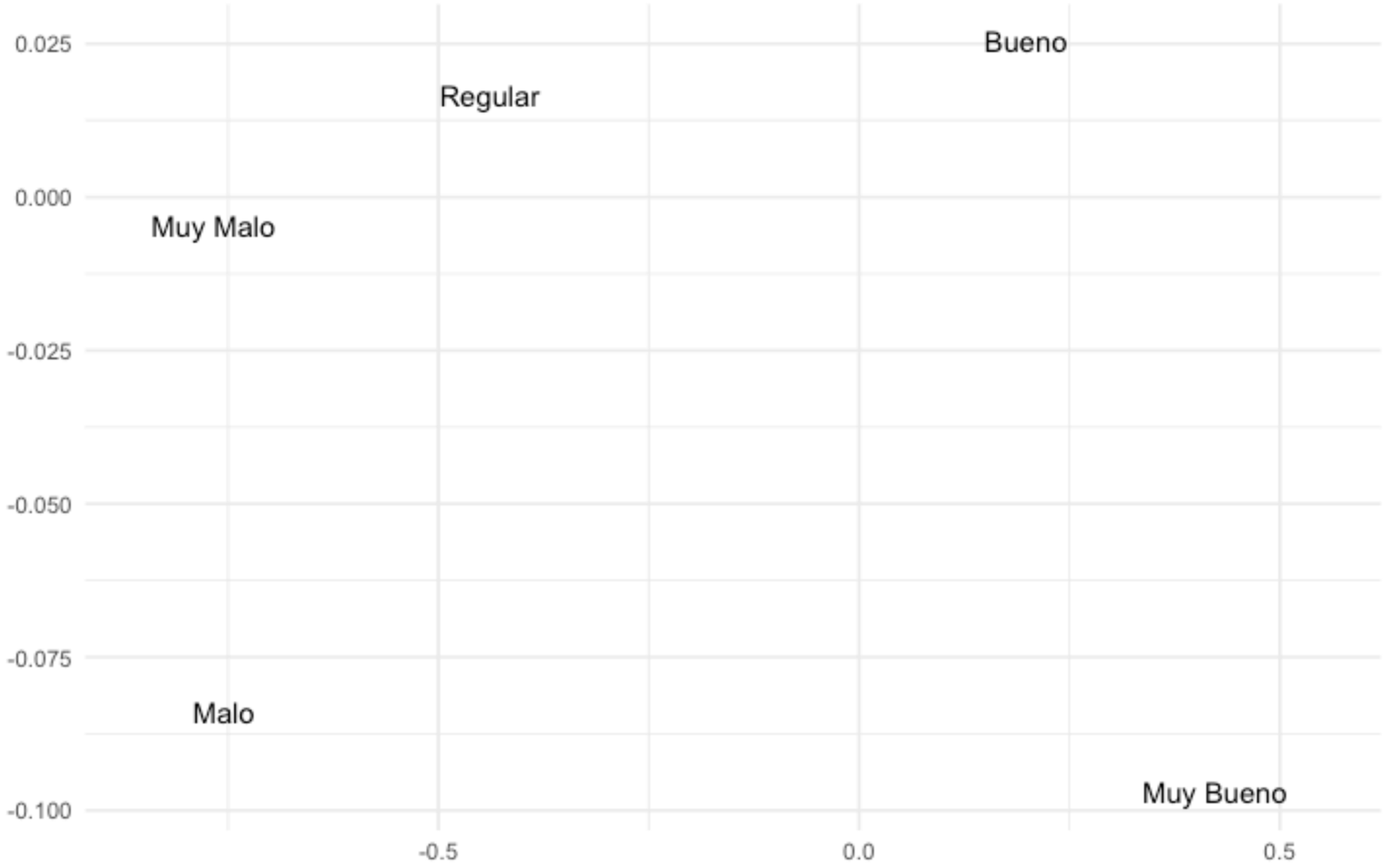
0.071	0.089	-0.098	-0.050	-0.011
0.050	0.100	-0.102	-0.037	-0.011
0.014	0.080	-0.065	-0.025	-0.008
-0.023	-0.011	0.039	-0.007	0.003
-0.070	-0.101	0.102	0.052	0.017
-0.066	-0.182	0.163	0.072	0.012
-0.077	-0.213	0.161	0.102	0.027

Ejemplo 1

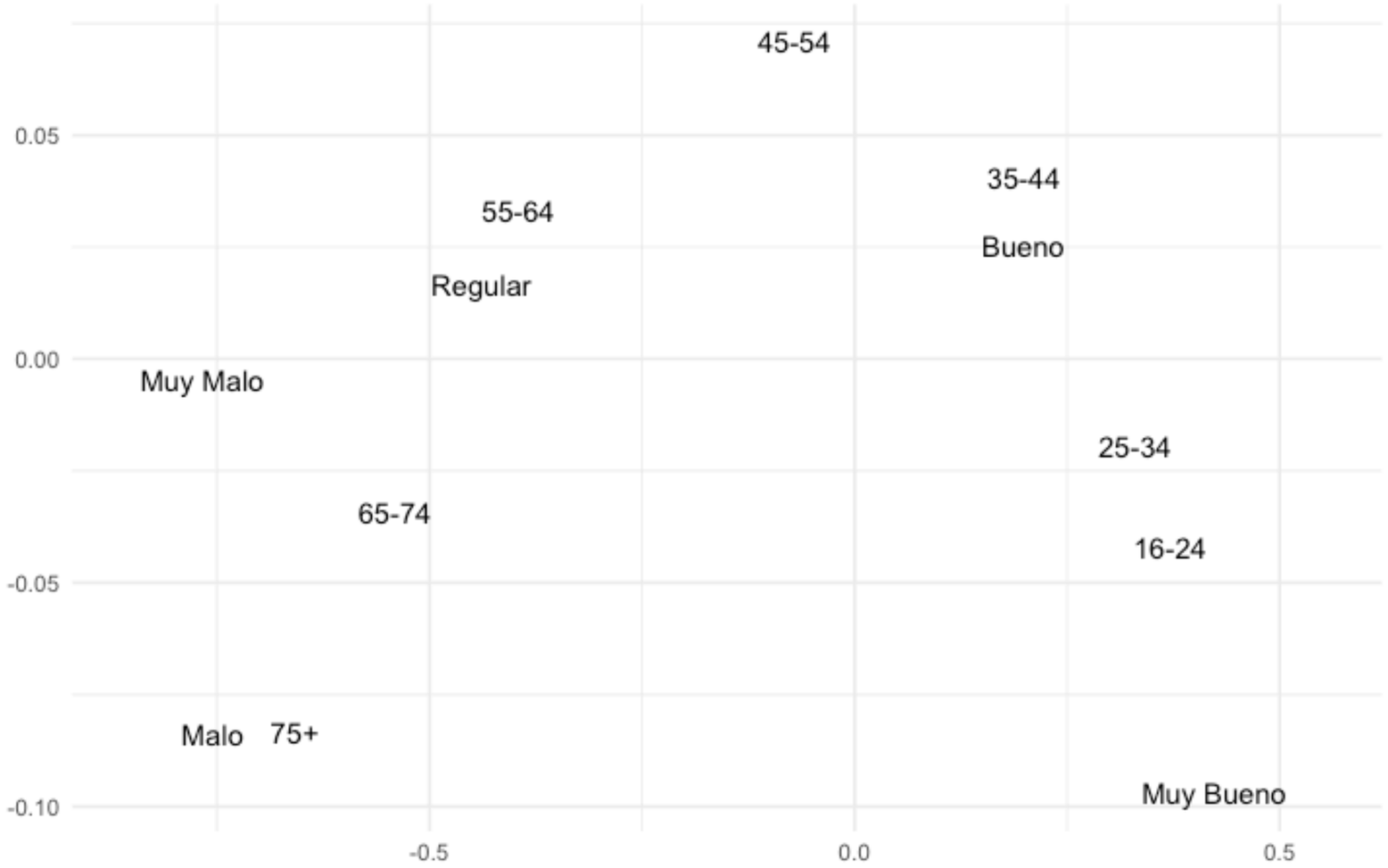
- Tenemos una representación en $\mathbb{R}^2$ de los grupos de edad



- Usamos la transpuesta de la tabla de contingencia y repetimos.



- Graficamos las dos variables al mismo tiempo



- Calcular la matriz de correspondencia $\mathbf{P} = \frac{\mathbf{N}}{n}$

Algoritmo 2

- Calcular la matriz de correspondencia $\mathbf{P} = \frac{\mathbf{N}}{n}$

- Definir matrices diagonales $\mathbf{D}_r$ y $\mathbf{D}_c$ con las sumas por renglón y columna.

Algoritmo 2

- Calcular la matriz de correspondencia $\mathbf{P} = \frac{\mathbf{N}}{n}$

- Definir matrices diagonales $\mathbf{D}_r$ y $\mathbf{D}_c$ con las sumas por renglón y columna.

- Obtener la descomposición SVD de

$$\mathbf{D}_r^{-\frac{1}{2}} \left(\mathbf{P} - \mathbf{r}\mathbf{c}^T \right) \mathbf{D}_c^{-\frac{1}{2}} = \mathbf{U}\Lambda\mathbf{V}^T$$

Algoritmo 2

- Calcular la matriz de correspondencia $\mathbf{P} = \frac{\mathbf{N}}{n}$

- Definir matrices diagonales $\mathbf{D}_r$ y $\mathbf{D}_c$ con las sumas por renglón y columna.

- Obtener la descomposición SVD de

$$\mathbf{D}_r^{-\frac{1}{2}} \left(\mathbf{P} - \mathbf{r}\mathbf{c}^T \right) \mathbf{D}_c^{-\frac{1}{2}} = \mathbf{U}\Lambda\mathbf{V}^T$$

- Obtener las coordenadas estándar

$$\mathbf{X} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U} \qquad \mathbf{Y} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}$$

Algoritmo 2

- Calcular la matriz de correspondencia $\mathbf{P} = \frac{\mathbf{N}}{n}$

- Definir matrices diagonales $\mathbf{D}_r$ y $\mathbf{D}_c$ con las sumas por renglón y columna.

- Obtener la descomposición SVD de

$$\mathbf{D}_r^{-\frac{1}{2}} \left(\mathbf{P} - \mathbf{r}\mathbf{c}^T \right) \mathbf{D}_c^{-\frac{1}{2}} = \mathbf{U}\Lambda\mathbf{V}^T$$

- Obtener las coordenadas estándar

$$\mathbf{X} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U} \qquad \mathbf{Y} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}$$

- Obtener las coordenadas principales

$$\mathbf{F} = \mathbf{X}\Lambda \qquad \mathbf{G} = \mathbf{Y}\Lambda$$

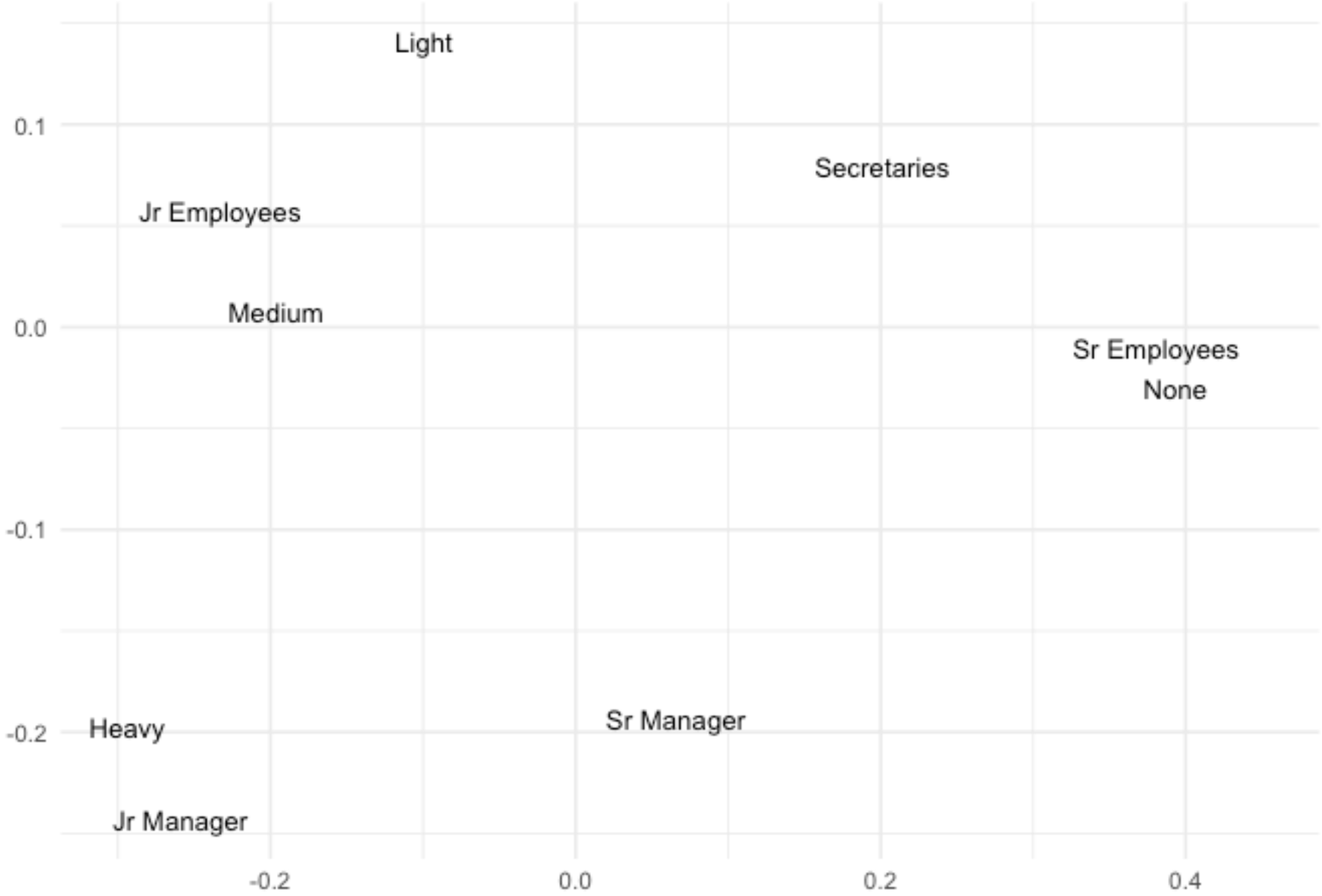
Ejemplo 2

- Encuesta a trabajadores de una empresa sobre sus hábitos de fumar

Staff\Nivel	None	Light	Medium	Heavy	Totales Ren.
Sr Managers	4	2	3	2	**11**
Jr Managers	4	3	7	4	**18**
Sr Employees	25	10	12	4	**51**
Jr Employees	18	24	33	13	**88**
Secretaries	10	6	7	2	**25**
Totales Col.	**61**	**45**	**62**	**25**	**193**

Ejemplo 2

- Graficamos las dos variables al mismo tiempo



- Como PCA se busca que explicar la mayor cantidad de varianza definida como:

$$\text{Inercia} = \sum_{i,j} \frac{\left(p_{ij} - r_i c_j\right)^2}{(r_i c_j)}$$

- Como PCA se busca que explicar la mayor cantidad de varianza definida como:

$$\text{Inercia} = \sum_{i,j} \frac{\left(p_{ij} - r_i c_j\right)^2}{(r_i c_j)}$$

- Equivalentemente, $\text{Inercia} = \frac{\chi^2}{n}$, donde χ^2 es el estadístico de Pearson y n el total de observaciones

- Como PCA se busca que explicar la mayor cantidad de varianza definida como:

$$\text{Inercia} = \sum_{i,j} \frac{\left(p_{ij} - r_i c_j\right)^2}{(r_i c_j)}$$

- Equivalentemente, $\text{Inercia} = \frac{\chi^2}{n}$, donde χ^2 es el estadístico de Pearson y n el total de observaciones

- Los valores singulares al cuadrado $\lambda_1^2, \lambda_2^2, \ldots$ son las inercias principales y explican la inercia total.

Ejemplo 2

- La inercia total:

$$\text{Inercia} = 0.08518986$$

- Las inercias principales:

$$\lambda_1^2 = 0.07475911; \qquad \lambda_2^2 = 0.01001718; \qquad \lambda_3^2 = 0.0004135741$$

- Porcentaje explicado acumulado:

$$0.8775588 \qquad 0.9951453 \qquad 1$$

- Las coordenadas de los renglones $\mathbf{F}$ y la de las columnas $\mathbf{G}$ están relacionadas

$$\mathbf{F} = \mathbf{RG}\Lambda^{-1} \qquad \mathbf{G} = \mathbf{CF}\Lambda^{-1}$$

- Las coordenadas de los renglones $\mathbf{F}$ y la de las columnas $\mathbf{G}$ están relacionadas

$$\mathbf{F} = \mathbf{R}\mathbf{G}\Lambda^{-1} \qquad \mathbf{G} = \mathbf{C}\mathbf{F}\Lambda^{-1}$$

- Nos da una forma de añadir perfiles suplementarios para columnas y renglones

Ejemplo 2

- Imaginar que se tiene un promedio nacional de fumadores

Staff\Nivel	None	Light	Medium	Heavy
Sr Managers	4	2	3	2
Jr Managers	4	3	7	4
Sr Employees	25	10	12	4
Jr Employees	18	24	33	13
Secretaries	10	6	7	2
Promedio	**42%**	**29%**	**20%**	**9%**

Ejemplo 2

- Imaginar que se tiene un promedio nacional de fumadores

Staff\Nivel	None	Light	Medium	Heavy
Sr Managers	4	2	3	2
Jr Managers	4	3	7	4
Sr Employees	25	10	12	4
Jr Employees	18	24	33	13
Secretaries	10	6	7	2
Promedio	**42%**	**29%**	**20%**	**9%**

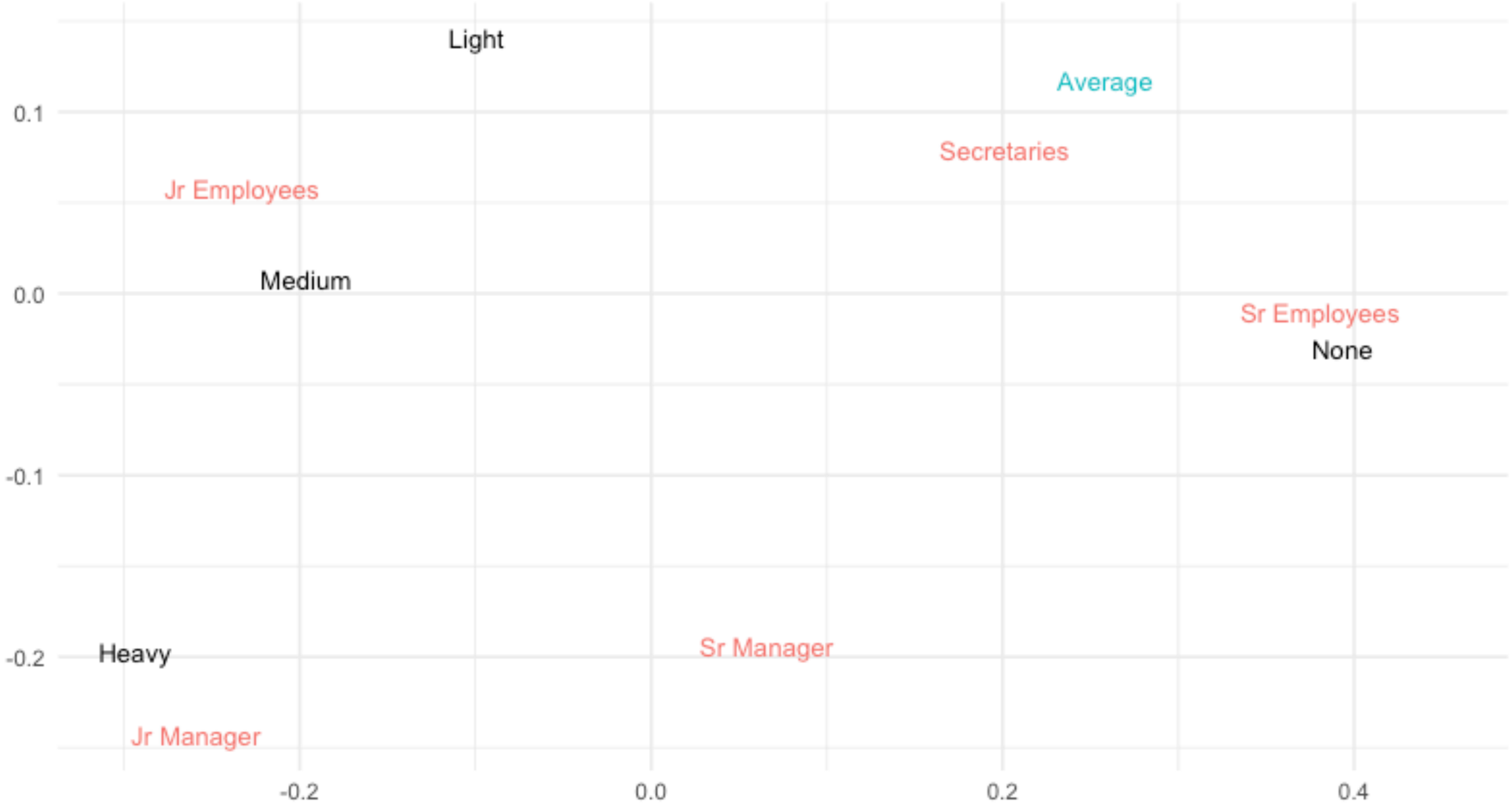
- Encontramos su representación como:

$$f_{11}* = \frac{(.42 * -0.39330845) + (.29 * 0.09945592) + (.2 * 0.19632096) + (.09 * 0.29377599)}{.2734211} = .258$$

$$f_{12}* = \frac{(.42 * -0.030492071) + (.29 * 0.141064289) + (.2 * 0.007359109) + (.09 * -0.197765656)}{0.1000859} = .118$$

Ejemplo 2

- Lo graficamos



Ejemplo 2

- De forma similar podemos añadir columnas

Staff\Nivel	None	Light	Medium	Heavy	Drinking	Not Drinking
Sr Managers	4	2	3	2	0	11
Jr Managers	4	3	7	4	1	17
Sr Employees	25	10	12	4	5	46
Jr Employees	18	24	33	13	10	78
Secretaries	10	6	7	2	7	18
Promedio	**42%**	**29%**	**20%**	**9%**		

- **Obs:** Ya **no** es una tabla de contingencia

- Lo graficamos