Análisis de correspondencias

José A. Perusquía Cortés

Análisis Multivariado Semestre 2025-II

Motivación

- 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.

Elementos

- Puntos en un espacio multidimensional y un peso (o masa) asignado a cada punto
- Un centroide
- 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: Estado de salud

- Estado de salud por grupo de edades

Grupo\Salud	Muy Bueno	Bueno	Regular	Malo	Muy Malo	Total
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**
Total	**817**	**3542**	**1495**	**414**	**103**	**6371**

Ejemplo: Estado de salud

- Tabla de frecuencias por renglón

Grupo\Salud	Muy Bueno	Bueno	Regular	Malo	Muy Malo	Pesos Ren.
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**
Pesos Col.	**0.128**	**0.556**	**0.235**	**0.065**	**0.016**	**1.000**

Ejemplo: Estado de salud

- Puntos en un espacio multidimensional: perfiles por renglón

Grupo\Salud	Muy Bueno	Bueno	Regular	Malo	Muy Malo	Pesos Ren.
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**
Pesos Col.	**0.128**	**0.556**	**0.235**	**0.065**	**0.016**	**1.000**

Ejemplo: Estado de salud

- Pesos (masas) de cada perfil: peso renglón

Grupo\Salud	Muy Bueno	Bueno	Regular	Malo	Muy Malo	Pesos Ren.
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**
Pesos Col.	**0.128**	**0.556**	**0.235**	**0.065**	**0.016**	**1.000**

Ejemplo: Estado de salud

- El centroide: peso columna

Grupo\Salud	Muy Bueno	Bueno	Regular	Malo	Muy Malo	Pesos Ren.
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**
Pesos Col.	**0.128**	**0.556**	**0.235**	**0.065**	**0.016**	**1.000**

Algoritmo

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

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

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

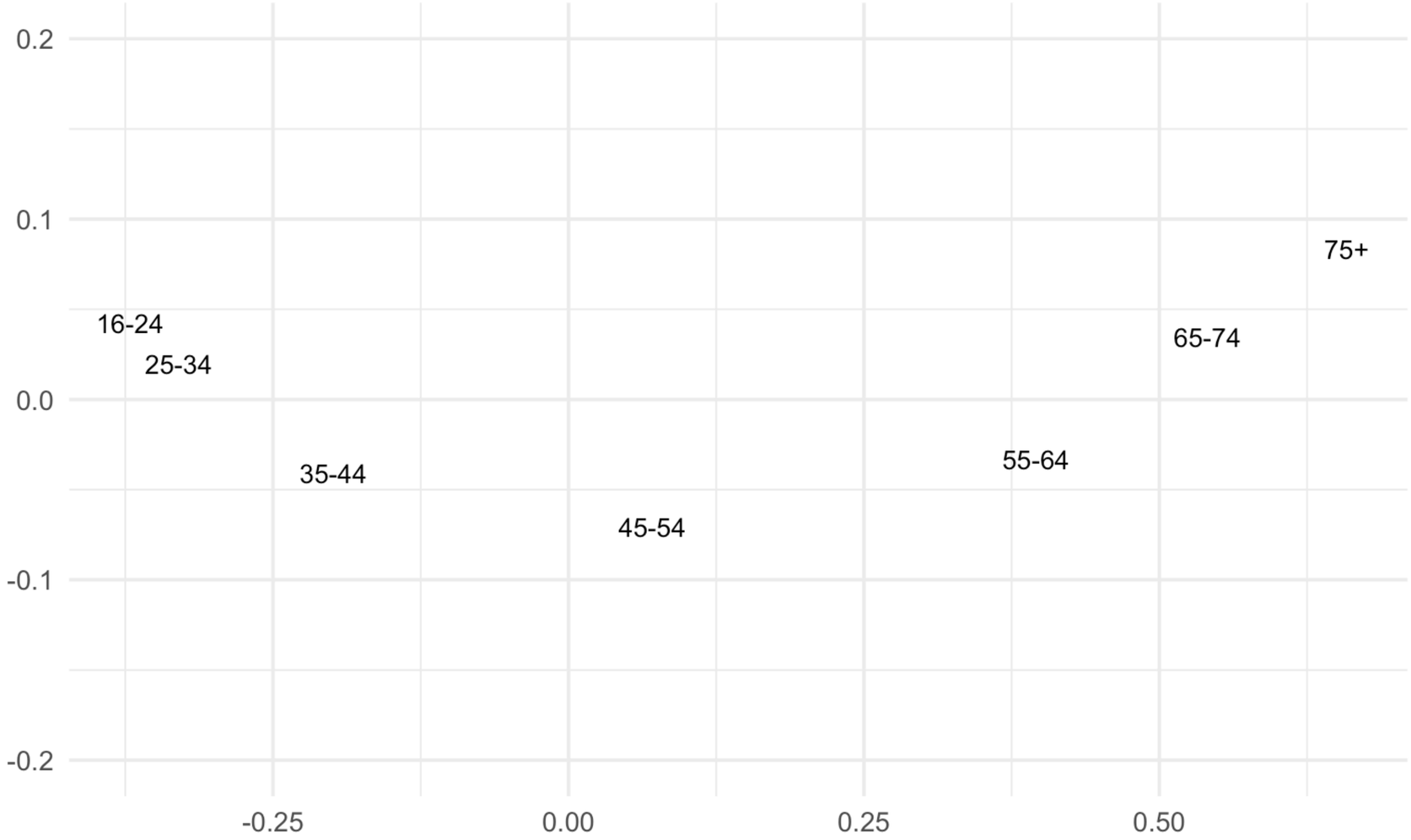
   - $\mathbf{R}$ es la matriz de perfiles por renglón
   - $\mathbf{c} = \mathbf{D}_c\mathbf{1}$ es el centroide

3. Las primeras dos coordenadas para los renglones se encuentran con $\mathbf{N}_{(2)}\Lambda_{(2)}$

4. Para las columnas (problema dual) se transpose la tabla y se repite el procedimiento (puede haber problemas)

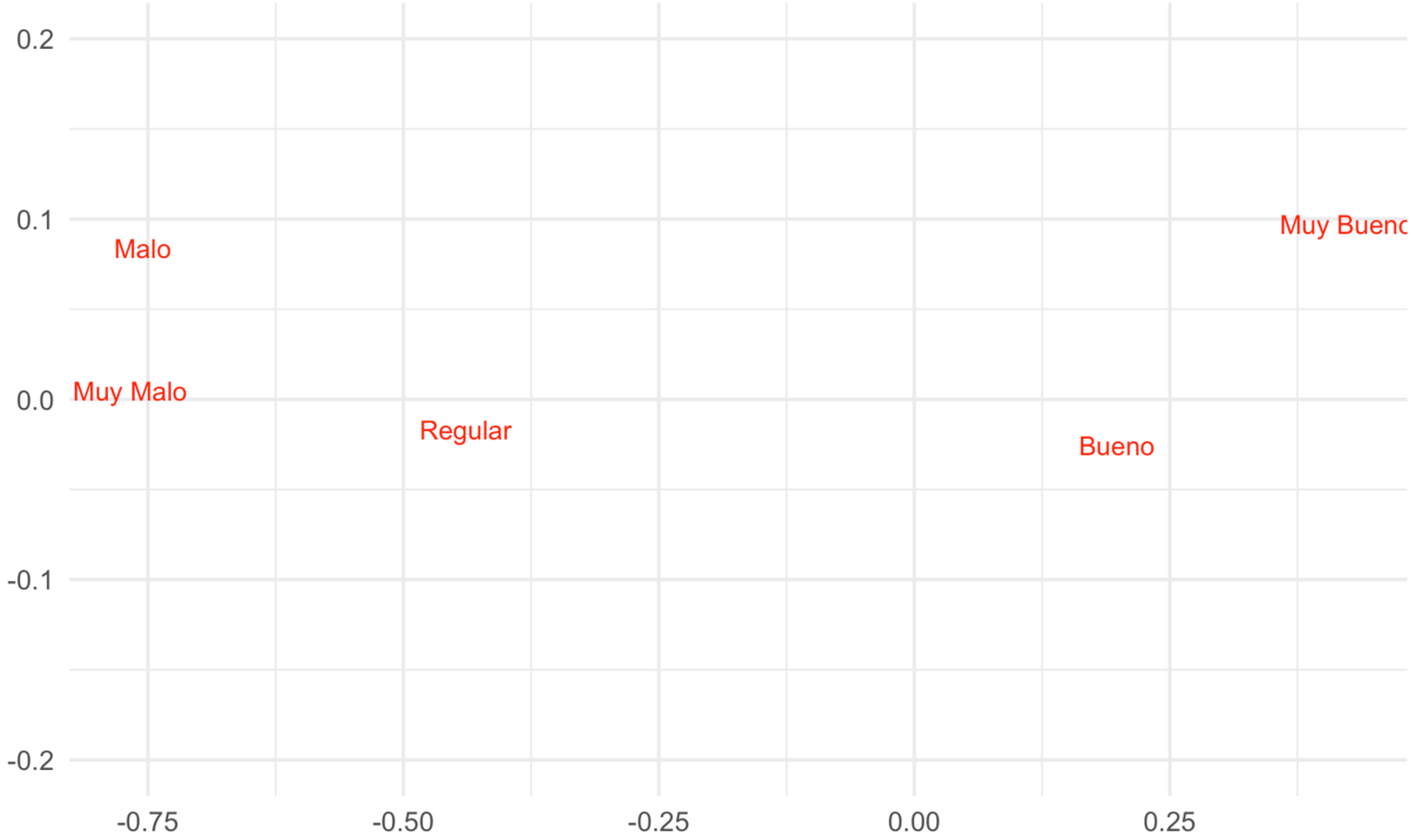
Ejemplo: Estado de salud

- Los grupos de edad



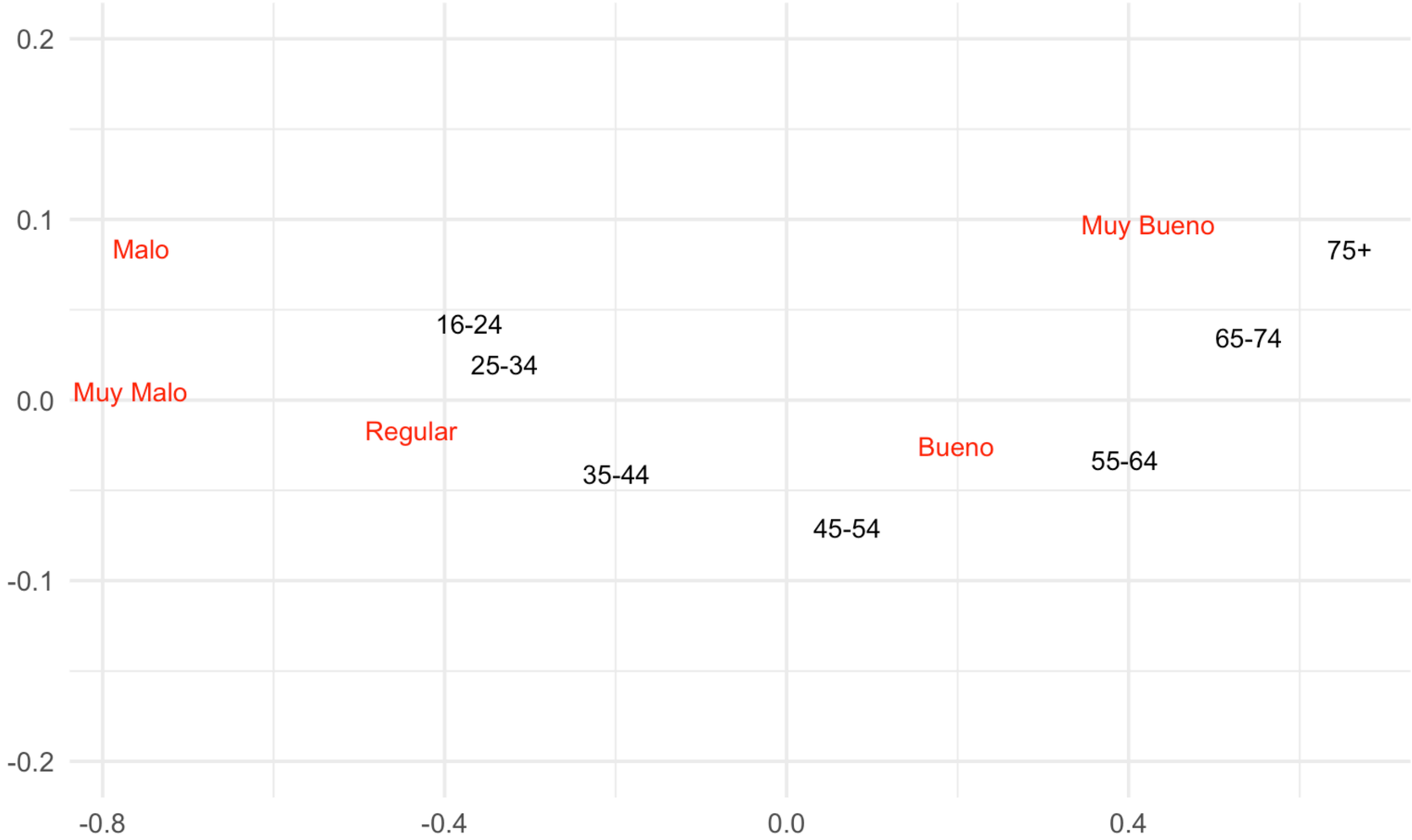
Ejemplo: Estado de salud

- Estado de salud



Ejemplo: Estado de salud

- Ambas variables (problema)



Algoritmo 2

1. Calcular la **matriz de correspondencia** $\mathbf{P} = \frac{\mathbf{N}}{n}$

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

3. Obtener la descomposición SVD de

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

4. Obtener las **coordenadas estándar**

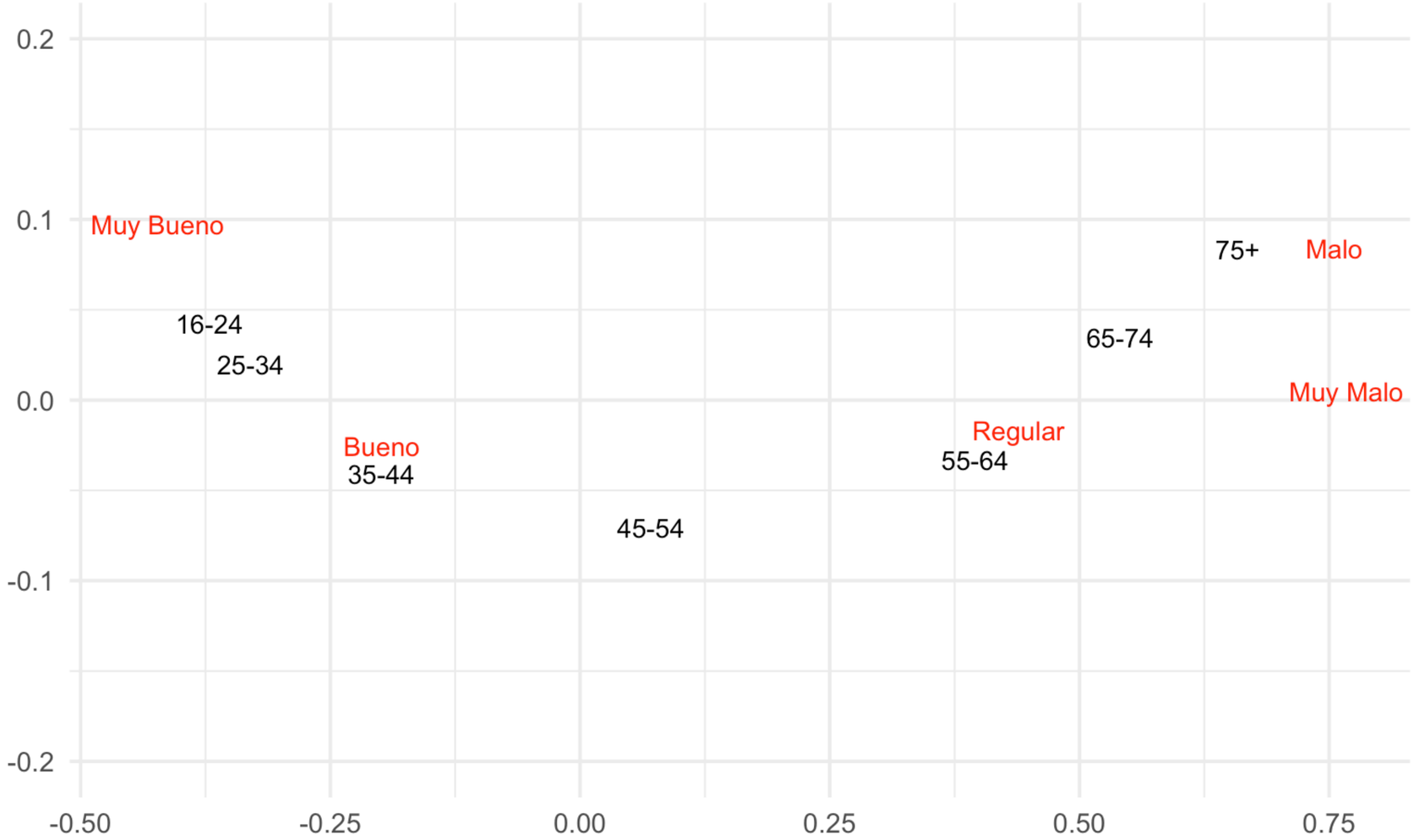
$$\mathbf{X} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U} \qquad \mathbf{Y} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}$$

5. Obtener las **coordenadas principales**

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

Ejemplo: Estado de salud

- Con el algoritmo 2

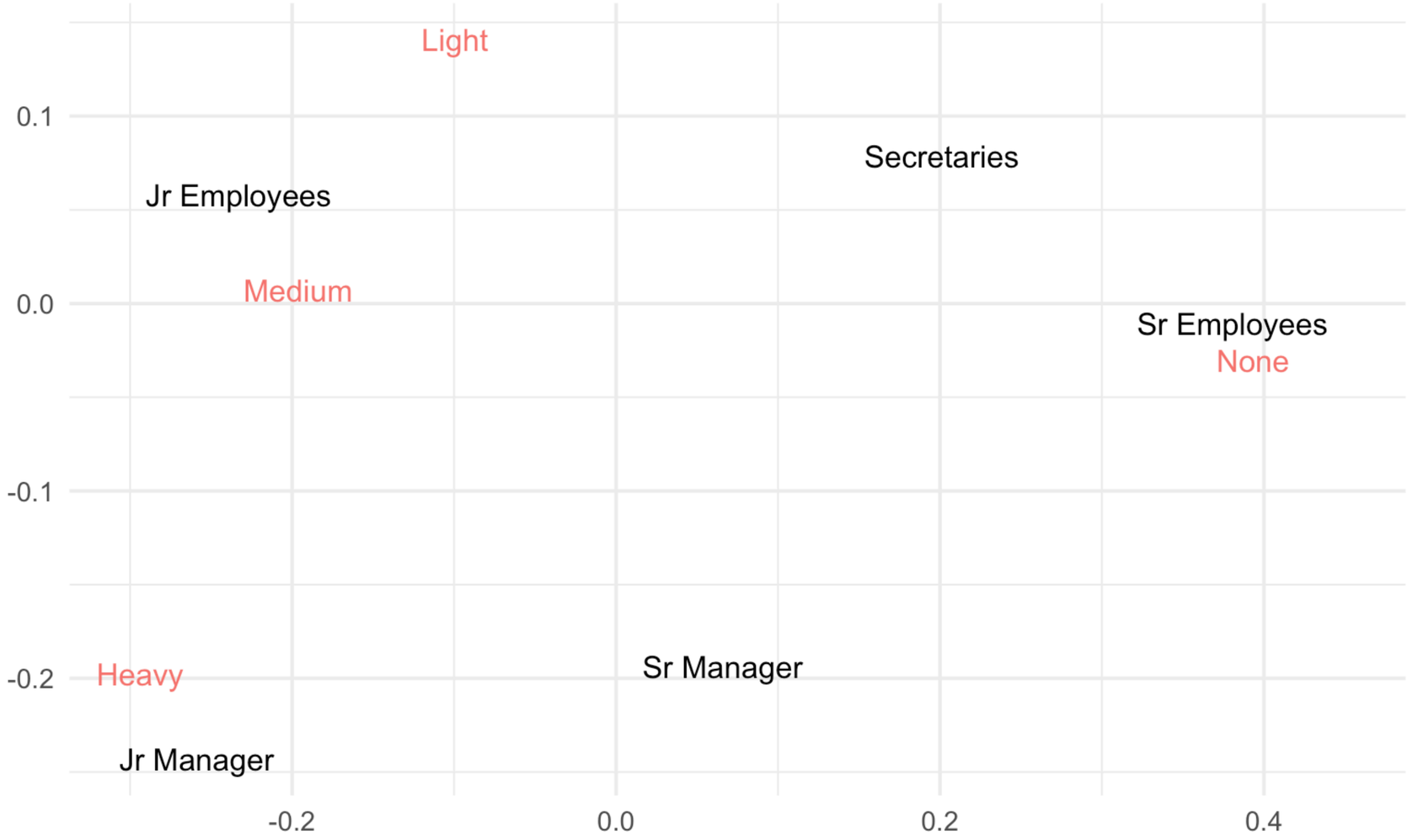


Ejemplo: Trabajadores y hábitos de fumar

- 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: Trabajadores y hábitos de fumar



Inercia

- Como PCA se busca 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: Trabajadores y hábitos de fumar

- 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:

$$87.76\,\% \rightarrow 99.51\,\% \rightarrow 100\,\%$$

Problema dual

- 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: Trabajadores y hábitos de fumar

- Se quiere comparar contra el promedio nacional de hábitos de fumar

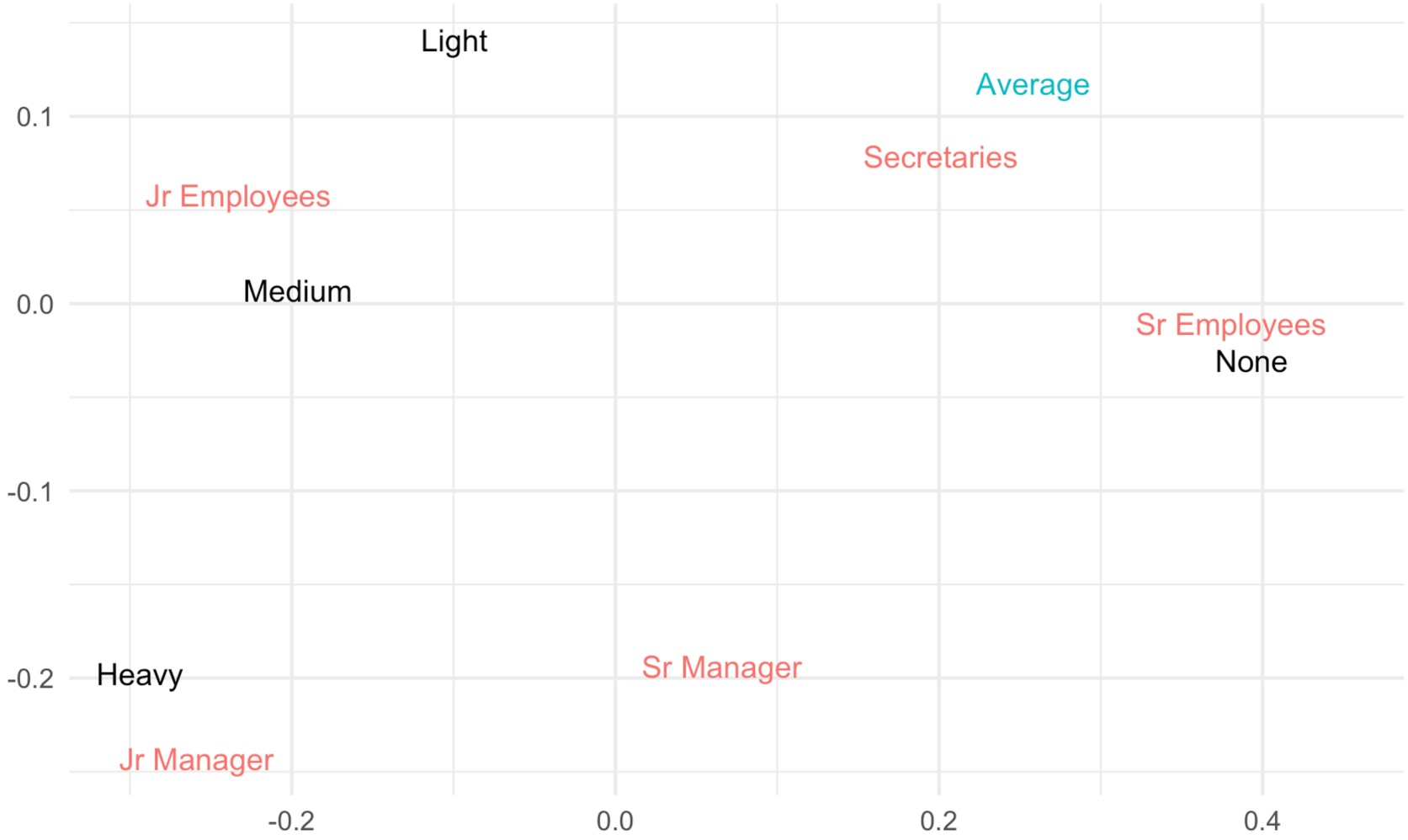
Staff\Nivel	None	Light	Medium	Heavy
Sr Managers	4	2	3	2
Jr Managers	4	3	7	4
Sr Employees	25	10	12	4
Jr Employees	18	24	33	13
Secretaries	10	6	7	2
Promedio	**42%**	**29%**	**20%**	**9%**

- 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: Trabajadores y hábitos de fumar



Ejemplo: Trabajadores y hábitos de fumar

- 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%**		

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

Ejemplo: Trabajadores y hábitos de fumar

Análisis de correspondencias múltiple (MCA)

MCA

- Extensión del análisis de correspondencias simple que permite analizar la relación de K variables categóricas dependientes, cada una con J_k niveles tales que $\sum_k J_k = J$

- Es importante que las variables sean “homogéneas”, e.g. no mezclar variables de opinión con variables demográficas.

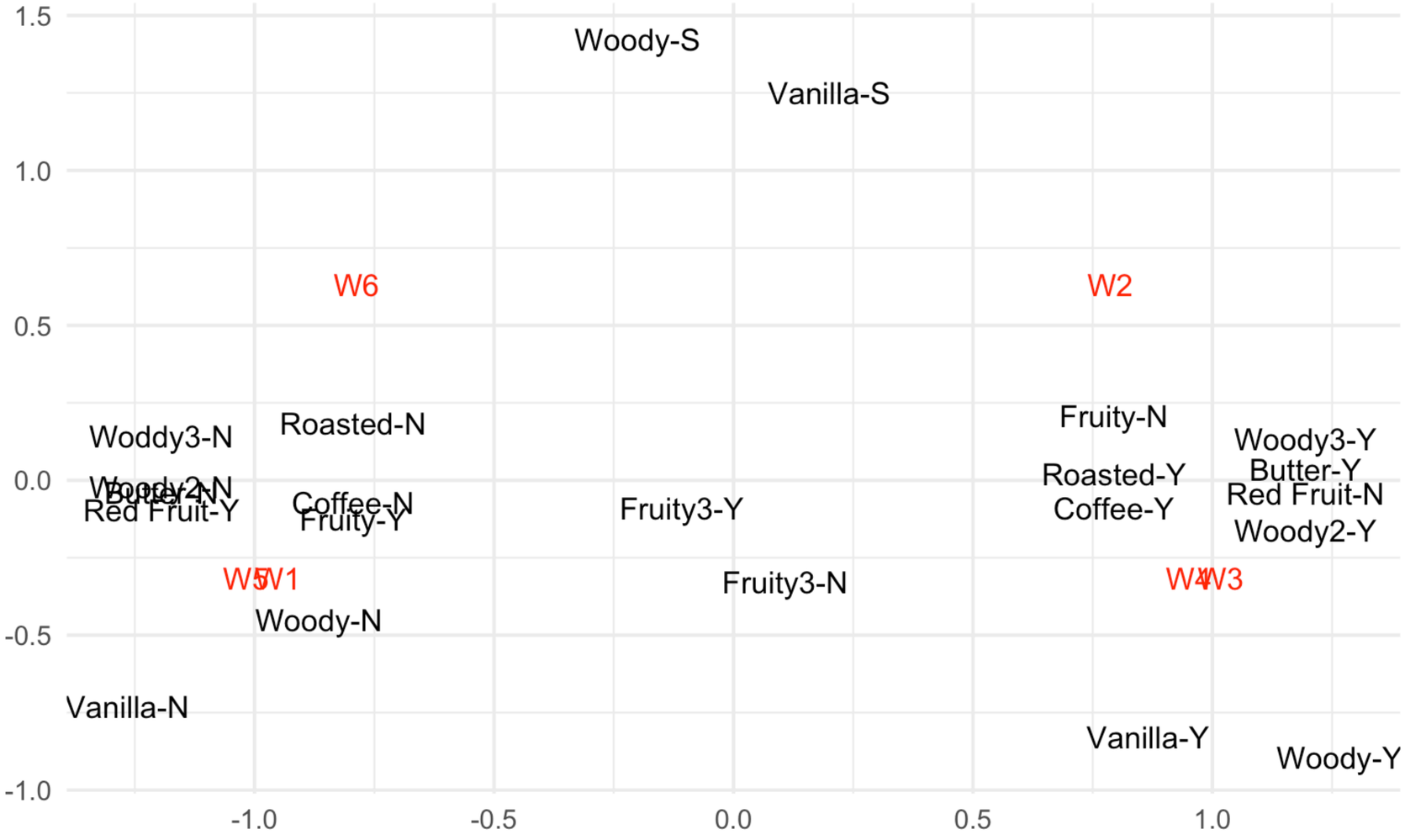
- A grandes rasgos se puede ver como un análisis de correspondencias en una matriz indicadora.

- Las variables pueden ser cuantitativas también, siempre que se agrupen.

Ejemplo: Características de vinos

		Expert 1			Expert 2				Expert 3		
Wine	Oak	Fruity	Woody	Coffee	Red	Roasted	Vanillin	Woody	Fruity	Butter	Woody
W1	1	10	001	01	10	01	001	01	01	01	01
W2	2	01	010	10	01	10	010	10	01	10	10
W3	2	01	100	10	01	10	100	10	01	10	10
W4	2	01	100	10	01	10	100	10	10	10	10
W5	1	10	001	01	10	01	001	01	10	01	01
W6	1	10	010	01	10	01	010	01	10	01	01
W?	?	01	010	.5 .5	10	10	010	.5 .5	10	.5 .5	01

Ejemplo: Características de vinos



Ejemplo: Características de vinos

- Inercias principales

$$\lambda_1^2 = 0.8532; \qquad \lambda_2^2 = 0.2; \qquad \lambda_3^2 = .1151; \qquad \lambda_4^2 = 0.0317$$

- Inercia total de 1.2

- Contribución a la inercia total

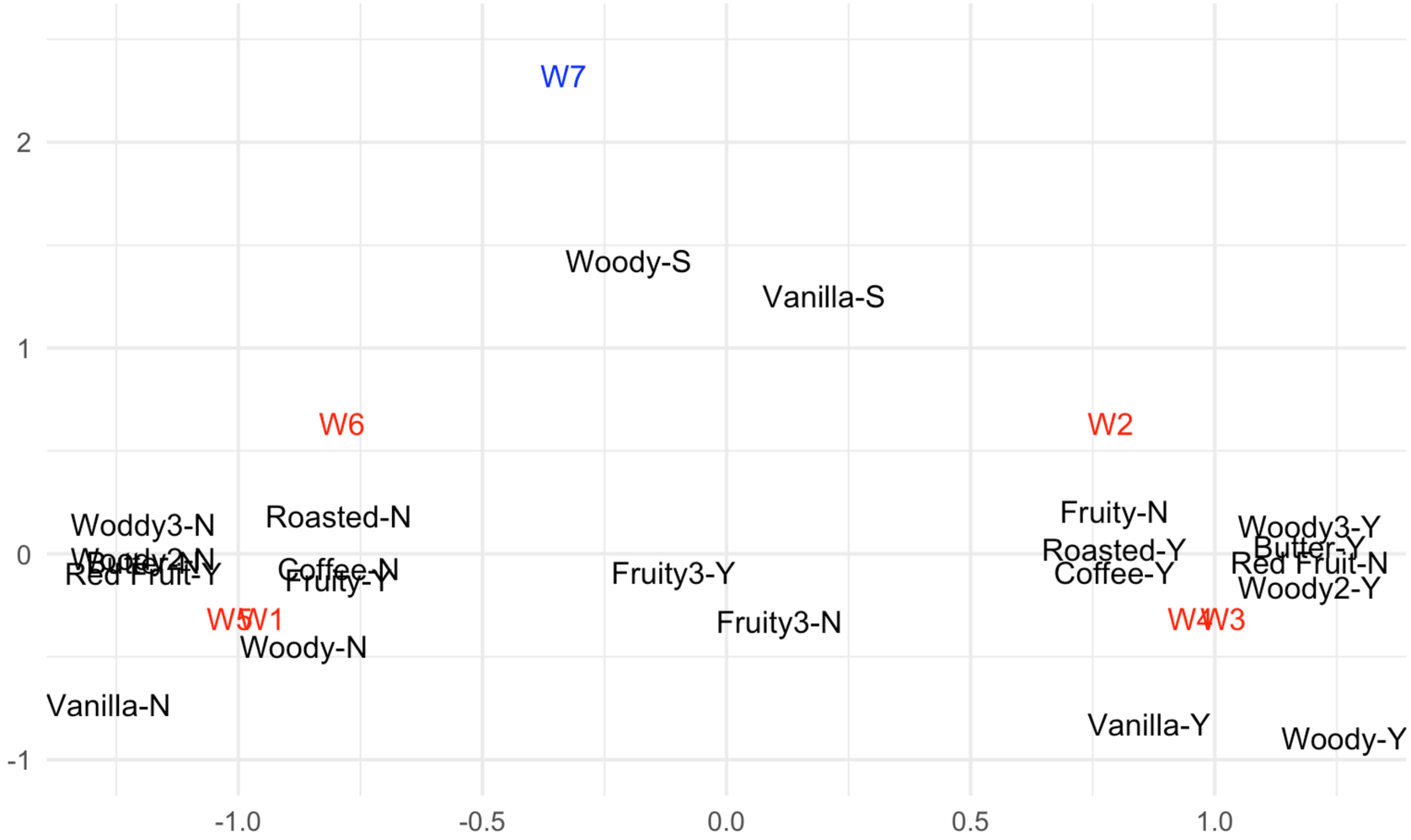
$$0.7110 \qquad 0.1666 \qquad 0.0959 \qquad 0.0263$$

- Inercia acumulada

$$71.10\,\% \rightarrow 87.76\,\% \rightarrow 97.36\,\% \rightarrow 100\,\%$$

Ejemplo: Características de vinos

- Añadiendo el vino del que no se conoce el tipo de barrica



Ejemplo: Encuesta de familia y cambios de rol

- Encuesta de 1994 del Programa Internacional de Investigación sobre la familia y los cambios de rol de género realiza a 33,590 individuos de 24 países

- Una de las preguntas de interés fue:

  ‘Una mujer con un niño en edad escolar en casa, ¿debe trabajar a tiempo completo, a tiempo parcial, o debe permanecer en casa?

- Se puede responder: ‘no está seguro/no sabe’.

- Se consideran cuatro preguntas y las respuestas de los 3418 individuos de Alemania

Ejemplo: Encuesta de familia y cambios de rol

- Se utiliza one-hot-encoding para pasar de las originales a matriz indicadora

Preguntas				**Pregunta 1**				**Pregunta 2**				**Pregunta 3**				**Pregunta 4**			
1	2	3	4	W	w	H	?	W	w	H	?	W	w	H	?	W	w	H	?
1	3	2	2	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0
2	3	3	2	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0
4	3	3	2	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0
4	4	4	4	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
4	4	4	4	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
1	3	2	1	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0

Ejemplo: Encuesta de familia y cambios de rol

- Se utiliza one-hot-encoding para pasar de las originales a matriz indicadora

Pregunta 1				**Pregunta 2**				**Pregunta 3**				**Pregunta 4**			
W	w	H	?	W	w	H	?	W	w	H	?	W	w	H	?
1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0
0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0
0	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0
0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0

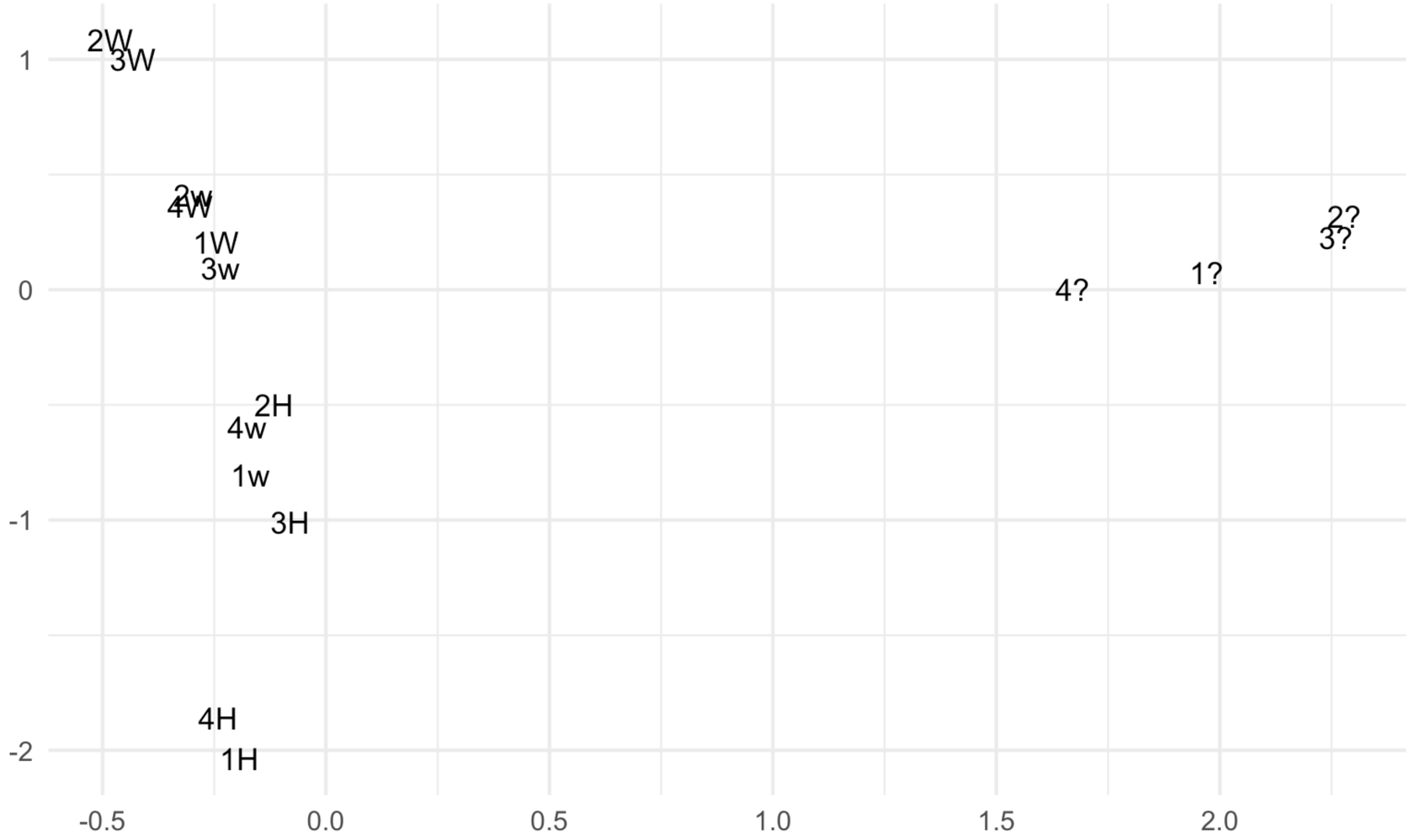
- Se genera una matriz de 3,418 renglones y 16 columnas

Ejemplo: Encuesta de familia y cambios de rol

- Se puede usar la matriz de Burt $\mathbf{X}^T\mathbf{X}$ para el análisis

	1W	***1w***	***1H***	***1?***	***2W***	***2w***	***2H***	***2?***	***3W***	***3w***	***3H***	***3?***	***4W***	***4w***	***4H***	***4?***
1W	2501	0	0	0	172	1107	1131	91	355	1710	345	91	1766	538	40	157
1w	0	476	0	0	7	129	335	5	16	261	181	18	128	293	17	38
1H	0	0	79	0	1	6	72	0	1	17	61	0	14	21	38	6
1?	0	0	0	362	1	57	108	196	7	96	55	204	51	45	2	264
2W	172	7	1	1	181	0	0	0	127	48	4	2	165	15	0	1
2w	1107	129	6	57	0	1299	0	0	219	997	61	22	972	239	13	75
2H	1131	335	72	108	0	0	1646	0	24	989	573	60	760	616	84	186
2?	91	5	0	196	0	0	0	292	9	50	4	229	62	27	0	203
3W	355	16	1	7	127	219	24	9	379	0	0	0	360	14	1	4
3w	1710	261	17	96	48	997	989	50	0	2084	0	0	1348	567	23	146
3H	345	181	61	55	4	61	573	4	0	0	642	0	202	286	73	81
3?	91	18	0	204	2	22	60	229	0	0	0	313	49	30	0	234
4W	1766	128	14	51	165	972	760	62	360	1348	202	49	1959	0	0	0
4w	538	293	21	45	15	239	616	27	14	567	286	30	0	897	0	0
4H	40	17	38	2	0	13	84	0	1	23	73	0	0	0	97	0
4?	157	38	6	264	1	75	186	203	4	146	81	234	0	0	0	465

Ejemplo: Encuesta de familia y cambios de rol



Ejemplo: Encuesta de familia y cambios de rol

- Inercia total de 1.145222

- Dos primeras inercias principales (cuadrado de las inercias de la matriz indicadora)

$$\lambda_1^2 = 0.4807494; \qquad \lambda_2^2 = 0.2633772$$

- Contribución a la inercia total

$$0.419787 \qquad 0.2299791$$

- Inercia acumulada

$$41.97\,\% \rightarrow 64.97\,\%$$

Ejemplo: Encuesta de familia y cambios de rol

- La inercia total es el promedio de las inercias de cada subtabla

Preguntas	Pregunta 1	Pregunta 2	Pregunta 3	Pregunta 4	Promedio
Pregunta 1	3	0.3657	0.4262	0.6457	1.1094
Pregunta 2	0.3657	3	0.8942	0.3477	1.1519
Pregunta 3	0.4262	0.8942	3	0.4823	1.2007
Pregunta 4	0.6457	0.3477	0.4823	3	1.1189
Promedio	1.1094	1.1519	1.2007	1.1189	1.1452

Mejoras

- MCA subestima la importancia de las inercias principales:
  - Con la matriz indicadora debido al aumento de la dimensionalidad
  - Con la matriz de Burt al inflar la inercia total con las tablas en la diagonal que no son de interés
- JCA es un algoritmo iterativo que hace análisis de correspondencias en la matriz de Burt tomando como valores faltantes los elementos de la diagonal (requiere algoritmo de imputación)
- Alternativamente se puede hacer un ajuste de la inercia y se re-escala la solución (mediante una regresión lineal)

Inercia ajustada

- Para las inercias de MCA con la matriz de Burt, Greenacre (1993) propuso:

$$\lambda_i^c = \left[\left(\frac{K}{K-1} \right) \left(\sqrt{\lambda}_i - \frac{1}{K} \right) \right]^2$$

- La inercia total se puede calcular como la suma de los valores singulares re-escalados o alternativamente se utiliza la inercia promedio de las matrices fuera de la diagonal (Greenacre, 1993) dada por:

$$\bar{\mathcal{J}} = \frac{K}{K-1} \left(\sum_i \lambda_i^2 - \frac{J-K}{K^2} \right)$$

Ejemplo: Encuesta de familia y cambios de rol

- Inercias corregidas

$$\lambda_1^c = 0.3494558; \qquad \lambda_2^c = 0.1231569$$

- Contribución a $\bar{\mathcal{J}} = 0.5269629$

$$0.6631 \qquad 0.2337$$

- Inercia acumulada

$$66.31\,\% \to 89.68\,\%$$