Tema 6
Análisis de Correspondencias
Curso 2017-18

1 Análisis de datos categóricos

- Ejemplo

Valores observados	Venta de producto			
	A	B	C	
Jóvenes (18-35 años)	20	20	20	60
Adultos (36-55 años)	40	10	40	90
Tercera Edad (56 + años)	20	10	40	70
	80	40	100	220

Valores esperados	Venta de producto			
	A	B	C	
Jóvenes (18-35 años)	21.82	10.91	27.27	60
Adultos (36-55 años)	32.73	16.36	40.91	90
Tercera Edad (56 + años)	25.45	12.73	31.82	70
	80	40	100	220

$$\chi^2_{\text{exp}} = 17.62$$
$$\chi^2_{4,0.95} = 9.49$$

1.1 Independencia de caracteres

- Consideremos dos variables categóricas (A, B)
  - A con n modalidades: $A_1, ..., A_n$
  - B con p modalidades: $B_1, ..., B_p$
- Hipótesis de independencia

$$H_0 : P(A_i \cap B_j) = P(A_i) P(B_j), \forall i, j$$

- Seleccionamos una muestra m de tamaño N y observamos (A, B) en cada elemento

  N_{ij} : núm. de elementos de m que presentan $A_i \cap B_j \quad (\sim Bi(N, P(A_i \cap B_j)))$

- Tabla de contingencia: Matriz de datos experimentales

	B_1	...	B_j	...	B_p	Marginal de A
A_1	N_{11}	...	N_{1j}	...	N_{1p}	$N_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	N_{i1}	...	N_{ij}	...	N_{ip}	$N_{i.} = \sum_{j=1}^{p} N_{ij}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_n	N_{n1}	...	N_{nj}	...	N_{np}	$N_{n.}$
Marginal de B	$N_{.1}$	...	$N_{.j} = \sum_{i=1}^{n} N_{ij}$	...	$N_{.p}$	N

- Estadístico $ji - cuadrado$ para la hipótesis de independencia

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(N_{ij} - e_{ij})^2}{e_{ij}}$$

con

$$e_{ij} = \widehat{E}_{H_0}(N_{ij}) \qquad (N_{ij} \sim_{H_0} Bi(N, P(A_i)P(B_j)))$$

$$e_{ij} = N\widehat{P}_{H_0}(A_i)\,\widehat{P}_{H_0}(B_j) = N\frac{N_{i.}}{N}\frac{N_{.j}}{N} = \frac{N_{i.}N_{.j}}{N}$$

- Distribución asintótica y región crítica para un test de tamaño α

$$\chi^2 \xrightarrow{\mathcal{L}} \chi^2_{(n-1)(p-1)}$$

$$Rc \;:\; \chi^2 \geq \chi^2_{(n-1)(p-1),1-\alpha}$$

1.2 Homogeneidad de poblaciones

- Consideremos una variable categórica B con p modalidades $B_1, ..., B_p$
- Estudio de la homogeneidad de B en n poblaciones $(P_1, ..., P_n)$
- Hipótesis de homogeneidad

$$H_0 : P_1(B_j) = ... = P_n(B_j) (= P(B_j)), \quad j = 1, ..., p$$

- Seleccionamos de forma independiente una muestra m_i de tamaño N_i de cada población.
  - Datos experimentales

N_{ij} : núm. de elementos de m_i que presentan B_j $\quad (\sim Bi(N_i, P_i(B_j)))$

$$\begin{array}{cccccc} & B_1 & ... & B_j & ... & B_p & \\ P_1 & N_{11} & ... & N_{1j} & ... & N_{1p} & N_1 = N_{1.} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ P_i & N_{i1} & ... & N_{ij} & ... & N_{ip} & N_i = N_{i.} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ P_n & N_{n1} & ... & N_{nj} & ... & N_{np} & N_n = N_{n.} \\ & N_{.1} & ... & N_{.j} = \sum_{i=1}^{n} N_{ij} & ... & N_{.p} & N_{..} = N \end{array}$$

- Estadístico $ji-cuadrado$ para la hipótesis de homogeneidad

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(N_{ij}-e_{ij})^2}{e_{ij}}$$

con

$$e_{ij} = \widehat{E}_{H_0}(N_{ij}) \qquad (N_{ij} \sim_{H_0} Bi(N_{i.}, P(B_j))$$

$$e_{ij} = N_{i.}\widehat{P}_{H_0}(B_j) = N_{i.}\frac{N_{.j}}{N} = \frac{N_{i.}N_{.j}}{N}$$

- Distribución asintótica y región crítica para un test de tamaño α

$$\chi^2 \xrightarrow{\mathcal{L}} \chi^2_{(n-1)(p-1)}$$

$$Rc \ : \ \chi^2 \geq \chi^2_{(n-1)(p-1),1-\alpha}$$

Nota 1 *Los problemas de independencia y homogeneidad son distintos. Sin embargo, el estadístico ji-cuadrado tiene la misma expresión y la misma distribución asintótica en ambos problemas.*

1.3 Distancia ji-cuadrado

- La distancia $ji-cuadrado$ es una distancia entre distribuciones de probabilidad
- Partimos de una tabla de contingencia (frecuencias absolutas)

$$\begin{array}{cccccc}
 & B_1 & \dots & B_j & \dots & B_p & \\
A_1 & N_{11} & \dots & N_{1j} & \dots & N_{1p} & N_{1.} \\
\vdots & \vdots & & \vdots & & \vdots & \vdots \\
A_i & N_{i1} & \dots & N_{ij} & \dots & N_{ip} & N_{i.} = \sum_{j=1}^{p} N_{ij} \\
\vdots & \vdots & & \vdots & & \vdots & \vdots \\
A_n & N_{n1} & \dots & N_{nj} & \dots & N_{np} & N_{n.} \\
 & N_{.1} & \dots & N_{.j} = \sum_{i=1}^{n} N_{ij} & \dots & N_{.p} & N
\end{array}$$

- Distribuciones conjuntas y marginales

$$\begin{bmatrix}
f_{11} & \dots & f_{1j} & \dots & f_{1p} \\
\vdots & & \vdots & & \vdots \\
f_{i1} & \dots & f_{ij} & \dots & f_{ip} \\
\vdots & & \vdots & & \vdots \\
f_{n1} & \dots & f_{nj} & \dots & f_{np}
\end{bmatrix} \qquad \begin{bmatrix} f_{1.} \\ \vdots \\ f_{i.} \\ \vdots \\ f_{n.} \end{bmatrix}$$

$$\begin{bmatrix} f_{.1} & \dots & f_{.j} & \dots & f_{.p} \end{bmatrix} \qquad 1$$

1.3.1 Distribuciones condicionadas por filas (perfiles filas)

$$M_r = \begin{bmatrix} r_1' \\ \vdots \\ r_i' \\ \vdots \\ r_n' \end{bmatrix} = \begin{bmatrix} \frac{N_{11}}{N_{1.}} & \dots & \frac{N_{1j}}{N_{1.}} & \dots & \frac{N_{1p}}{N_{1.}} \\ \vdots & & \vdots & & \vdots \\ \frac{N_{i1}}{N_{i.}} & \dots & \frac{N_{ij}}{N_{i.}} & \dots & \frac{N_{ip}}{N_{i.}} \\ \vdots & & \vdots & & \vdots \\ \frac{N_{n1}}{N_{n.}} & \dots & \frac{N_{nj}}{N_{n.}} & \dots & \frac{N_{np}}{N_{n.}} \end{bmatrix} \quad \begin{bmatrix} \frac{N_{1.}}{N} \\ \vdots \\ \frac{N_{i.}}{N} \\ \vdots \\ \frac{N_{n.}}{N} \end{bmatrix}$$

$$\text{Media ponderada} = m_r' = \begin{bmatrix} \frac{N_{.1}}{N} & \dots & \frac{N_{.j}}{N} & \dots & \frac{N_{.p}}{N} \end{bmatrix}$$

$$\begin{bmatrix} r_1' \\ \vdots \\ r_i' \\ \vdots \\ r_n' \end{bmatrix} = \begin{bmatrix} r_{11} & \dots & r_{1j} & \dots & r_{1p} \\ \vdots & & \vdots & & \vdots \\ r_{i1} & \dots & r_{ij} & \dots & r_{ip} \\ \vdots & & \vdots & & \vdots \\ r_{n1} & \dots & r_{nj} & \dots & r_{np} \end{bmatrix} \quad \begin{bmatrix} f_{1.} \\ \vdots \\ f_{i.} \\ \vdots \\ f_{n.} \end{bmatrix}$$

$$m_r' = \begin{bmatrix} f_{.1} & \dots & f_{.j} & \dots & f_{.p} \end{bmatrix}$$

- Consideramos cada fila r_i' de M_r como un punto de R^p con peso $f_{i.}$.
- El centro de gravedad m_r (media) de las filas de M_r viene dado por

$$m_r' = \sum_{i=1}^{n} f_{i.} r_i' = \begin{bmatrix} f_{.1} & \dots & f_{.j} & \dots & f_{.p} \end{bmatrix}$$

- Distancia $ji-cuadrado$ entre dos filas de M_r

$$\begin{aligned} d_{\chi^2}(r_i, r_{i'}) &= \sum_{j=1}^{p} \frac{(r_{ij} - r_{i'j})^2}{f_{.j}} = (r_i - r_{i'})' D_p^{-1} (r_i - r_{i'}) \\ D_p &= diag(f_{.1}, ..., f_{.p}) \quad \text{(simétrica, definida positiva)} \end{aligned}$$

- Distancia $ji-cuadrado$ de una fila r_i a m_r

$$d_{\chi^2}(r_i, m_r) = \sum_{j=1}^{p} (r_{ij} - f_{.j})^2 \frac{1}{f_{.j}} = (r_i - m_r)' D_p^{-1} (r_i - m_r)$$

- Expresión alternativa del estadístico $ji-cuadrado$

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{(N_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{\left(N_{ij} - \frac{N_{i.}N_{.j}}{N}\right)^2}{\frac{N_{i.}N_{.j}}{N}} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{N_{i.}^2 \left(\frac{N_{ij}}{N_{i.}} - \frac{N_{.j}}{N}\right)^2}{\frac{N_{i.}N_{.j}}{N}} = N \sum_{i=1}^{n} f_{i.} \sum_{j=1}^{p} (r_{ij} - f_{.j})^2 \frac{1}{f_{.j}} \\ &= N \sum_{i=1}^{n} f_{i.} d_{\chi^2}(r_i, m_r) \end{aligned}$$

- **Inercia total**

$$InT = \frac{\chi^2}{N}$$

es la media ponderada de las distancias ji-cuadrados entre los perfiles filas y su centroide.

- **Inercia de la fila** r_i

$$In\left(r_i\right) = f_{i.} d_{\chi^2}\left(r_i, m_r\right)$$

verificando

$$\sum_{i=1}^{n} In\left(r_i\right) = InT$$

$In\left(r_i\right)$ es una medida de la aportación de la fila r_i al estadístico χ^2. (Valores altos de χ^2 son reflejo de la existencia de asociación)

Nota 2 • *La hipótesis de independencia es equivalente a que todos los perfiles filas y columnas sean iguales (todas las distribuciones condicionadas coinciden). En tal caso todos los perfiles coinciden con la distribución marginal correspondiente.*

- *Un estadístico $ji-cuadrado$ significativo se puede interpretar como una desviación significativa de los perfiles fila y columna respecto de su centroide.*

Nota 3 ***Principio de equivalencia distribucional*** *(Propiedad de la distancia $ji-cuadrado$)*

- Si se agregan dos filas con los mismos perfiles, la distancia entre los perfiles columnas no se altera
- Si se agregan dos columnas con los mismos perfiles, la distancia entre los perfiles filas no se altera
  - Nota: es lógico considerar dos puntos que se solapan como un único punto.

1.3.2 Distribuciones condicionadas por columnas (perfiles columnas)

$$M_c = \begin{bmatrix} c_1' \\ \vdots \\ c_j' \\ \vdots \\ c_p' \end{bmatrix} = \begin{bmatrix} \frac{N_{11}}{N_{.1}} & \dots & \frac{N_{i1}}{N_{.1}} & \dots & \frac{N_{n1}}{N_{.1}} \\ \vdots & & \vdots & & \vdots \\ \frac{N_{1j}}{N_{.j}} & \dots & \frac{N_{ij}}{N_{.j}} & \dots & \frac{N_{nj}}{N_{.j}} \\ \vdots & & \vdots & & \vdots \\ \frac{N_{1p}}{N_{.p.}} & \dots & \frac{N_{ip}}{N_{.p}} & \dots & \frac{N_{np}}{N.p} \end{bmatrix} \qquad \begin{bmatrix} \frac{N_{.1}}{N} \\ \vdots \\ \frac{N_{.j}}{N} \\ \vdots \\ \frac{N_{.p}}{N} \end{bmatrix}$$

$$m_c' = \begin{bmatrix} \frac{N_{1.}}{N} & \dots & \frac{N_{i.}}{N} & \dots & \frac{N_{n.}}{N} \end{bmatrix}$$

$$\begin{bmatrix} c_1' \\ \vdots \\ c_j' \\ \vdots \\ c_p' \end{bmatrix} = \begin{bmatrix} c_{11} & \dots & c_{1i} & \dots & c_{1n} \\ \vdots & & \vdots & & \vdots \\ c_{j1} & \dots & c_{ji} & \dots & c_{jn} \\ \vdots & & \vdots & & \vdots \\ c_{p1} & \dots & c_{pi} & \dots & c_{pn} \end{bmatrix} \qquad \begin{bmatrix} f_{.1} \\ \vdots \\ f_{.j} \\ \vdots \\ f_{.p} \end{bmatrix}$$

$$m_c' = \begin{bmatrix} f_{1.} & \dots & f_{i.} & \dots & f_{n.} \end{bmatrix}$$

- Consideremos cada fila c_j' de M_c como un punto de R^n con peso $f_{.j}$
- El centro de gravedad m_c de las filas de M_c viene dado por

$$m_c' = \sum_{j=1}^{p} f_{.j} c_j' = \begin{bmatrix} f_{1.} & \dots & f_{i.} & \dots & f_{n.} \end{bmatrix}$$

- Distancia $ji-cuadrado$ entre dos columnas (filas de M_c)

$$\begin{aligned} d_{\chi^2}(c_j, c_{j'}) &= \sum_{i=1}^{n} \frac{(c_{ji} - c_{j'i})^2}{f_{i.}} = (c_j - c_{j'})' D_n^{-1} (c_j - c_{j'}) \\ D_n &= diag(f_{1.}, ..., f_{n.}) \quad \text{(simétrica, definida positiva)} \end{aligned}$$

- Distancia $ji-cuadrado$ entre una columna y m_c

$$d_{\chi^2}(c_j, m_c) = \sum_{i=1}^{n} \left(\frac{f_{ij}}{f_{.j}} - f_{i.} \right)^2 \frac{1}{f_{i.}} = (c_j - m_c)' D_n^{-1} (c_j - m_c)$$

- Expresión alternativa del estadístico $ji-cuadrado$

$$\chi^2 = N \sum_{j=1}^{p} f_{.j} d_{\chi^2}(c_j, m_c)$$

- **La inercia total** $\frac{\chi^2}{N}$ es la media ponderada de las distancias ji-cuadrado entre los perfiles columna y su centroide.
- **Inercia de la columna** c_j

$$\begin{aligned} In(c_j) &= f_{.j} d_{\chi^2}(c_j, m_c) \\ \sum_{j=1}^{p} In(c_j) &= InT \end{aligned}$$

2 Análisis de correspondencia (AC)

2.1 Relación entre el AC y el ACP

Análisis de Componentes Principales

- Partimos de n individuos en p dimensiones: $x'_1, ..., x'_n$.
- Matriz de datos $X = \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix}$
- Objetivo: proyectar los puntos sobre un número reducido de direcciones (k) de forma que las distancias entre las proyecciones en el espacio reducido $\left(R^k\right)$ sean similares a las distancias entre los puntos en el espacio inicial R^n.

- Las componentes principales son las direcciones sobre las que las proyecciones tienen máxima varianza y están incorreladas
  - Las proyecciones de $x_1', ..., x_n'$ sobre una dirección e vienen dadas por $e'x_1, ..., e'x_n$
  - Varianza de las proyecciones sobre e

    $$\text{var. proy.} = \frac{1}{n-1}\sum_{i=1}^{n}\left(e'x_i - e'\overline{x}\right)^2 = e'\frac{1}{n-1}\sum_{i=1}^{n}\left(x_n - \overline{x}\right)\left(x_n - \overline{x}\right)' e = e'\widehat{\Sigma}\, e$$

    Alternativamente,

    $$\text{var. proy.} = \frac{1}{n-1}p_e'p_e = p_e'\ diag\left(\frac{1}{n-1}, ..., \frac{1}{n-1}\right)p_e$$

    con

    $$p_e = \begin{pmatrix} x_1' - \overline{x}' \\ \vdots \\ x_n' - \overline{x}' \end{pmatrix} e$$

    el vector de proyecciones centrado.

- Las componentes principales vienen determinadas por las soluciones a los problemas
  - Primera componente

$$\max_{e_1} \frac{1}{n-1} p'_{e_1} p_{e_1} \equiv \max_{e_1} \; e'_1 \widehat{\Sigma} \; e_1$$

$$s.a. \;\; e'_1 e_1 = 1$$

  - Segunda componente

$$\max_{e_2} \frac{1}{n-1} p'_{e_2} p_{e_2} \;\equiv \max_{e_2} \; e'_2 \widehat{\Sigma} \; e_2$$

$$s.a. \;\; e'_2 e_2 = 1, \quad e'_2 e_1 = 0$$

  - Tercera componente...

- Las soluciones a los problemas vienen dados por los autovalores-autovectores de $\widehat{\Sigma}$: $(\lambda_1, e_1), ..., (\lambda_p, e_p)$
  - Autovalores $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ de $\widehat{\Sigma}$
  - Verificando

$$\max_{e_1} \frac{1}{n-1} p'_{e_1} p_{e_1} = \max_{e_1} \; e_1' \widehat{\Sigma} \; e_1 = \lambda_1$$
$$s.a. \;\; e_1' e_1 = 1$$

  -

$$\max_{e_2} \frac{1}{n-1} p'_{e_2} p_{e_2} \;\; = \max_{e_2} \; e_2' \widehat{\Sigma} \; e_2 = \lambda_2$$
$$s.a. \;\; e_2' e_2 = 1, \;\; e_2' e_1 = 0$$

  - ...

Nota 4 • *Sea $e_1, ..., e_p$ una base ortonormal de R^p*

- Dados dos puntos $x = (x_1, ..., x_p)'$ e $y = (y_1, ..., y_p)'$, sus proyecciones tienen coordenadas $P_x = \left(e_1'x, ..., e_p'x\right)$ y $P_y = \left(e_1'y, ..., e_p'y\right)$
- La proyección mantiene las distancias

$$d^2(P_x, P_y) = \sum_{i=1}^{p} (e_i'x - e_i'y)^2 = (x-y)' \underbrace{\sum_{i=1}^{p} e_i e_i'}_{I_p} (x-y) = d^2(x, y)$$

- Si $\lambda_i \simeq 0$, las proyecciones de $x_1, ..., x_n$ sobre la dirección e_i tienen poca variabilidad (las proyecciones toman valores parecidos)
- Si se desprecian las direcciones de poca variabilidad correpondientes a $(\lambda_{k+1}, ..., \lambda_p)$, las distancias originales en R^p se mantienen de forma aproximada en el espacio reducido R^k

$$d^2_{R^k}(P_x, P_y) = \sum_{i=1}^{k} (e_i'x - e_i'y)^2 \simeq d^2(x, y)$$

Análisis de correspondencia

- Consideremos dos variables categóricas A, con n modalidades, y B, con p modalidades.
- Consideremos la tabla de contingencia asociada a una muestra de tamaño N
- Sea $F = (f_{ij})_{i=1,...,n,\ j=1,...,p}$ la matriz de frecuencias relativas asociada a una tabla de contingencia $(n \times p)$

$$F = \begin{pmatrix} f_{11} & ... & f_{1j} & ... & f_{1p} \\ \vdots & & \vdots & & \vdots \\ f_{i1} & ... & f_{ij} & ... & f_{ip} \\ \vdots & & \vdots & & \vdots \\ f_{n1} & ... & f_{nj} & ... & f_{np} \\ \hline f_{.1} & ... & f_{.j} & ... & f_{.p} \end{pmatrix} \quad \begin{pmatrix} f_{1.} \\ \vdots \\ f_{i.} \\ \vdots \\ f_{n.} \end{pmatrix}$$

- Sean D_n y D_p las matrices diagonales asociadas a las distribuciones marginales

$$D_n = diag\,(f_{1.}, ..., f_{n.})$$

$$D_p = diag\,(f_{.1}, ..., f_{.p})$$

- Las distribuciones condicionadas por filas y columnas son las filas de las matrices

$$\begin{array}{rcl} \text{Perfiles filas:} & & M_r = D_n^{-1} F \\ \text{Perfiles columnas} & : & M_c = D_p^{-1} F' \end{array}$$

- Objetivo: Representar los perfiles filas y columnas en un número menor de dimensiones (generalmente 2) de forma que los perfiles próximos en la métrica $ji - cuadrado$ tengan representaciones próximas en la distancia euclídea.
  - El objetivo es el mismo que en el ACP con la peculiaridad de que los "elementos" del AC son distribuciones y la distancia que se utiliza es la distancia $ji - cuadrado$

2.2 AC por filas (respecto de la media)

- Consideramos las n distribuciones condicionadas por filas como n puntos de R^p
  - Los puntos son las n filas de la matriz $M_r = D_n^{-1}F = \begin{pmatrix} r_1' \\ \vdots \\ r_n' \end{pmatrix}$
  - El peso de cada punto es $f_{i.}$, $i = 1, ..., n$
  - Media (ponderada) de las filas: $m_r' = \sum_{i=1}^n r_i'\ f_{i.} = \begin{pmatrix} f_{.1} & ... & f_{.p} \end{pmatrix}$
- Sea $u \in R^p$ de forma que $u'D_p^{-1}u = 1$ (norma asociada a la distancia $ji - cuadrado$)
  - El vector de proyecciones (según la distancia $ji - cuadrado$) de las filas sobre un eje $u \in R^p$ viene dado por

$$\begin{pmatrix} r_1'D_p^{-1}u \\ \vdots \\ r_n'D_p^{-1}u \end{pmatrix} = M_rD_p^{-1}u \in R^n$$

  - Media (ponderada) de las proyecciones sobre u

$$\sum_{i=1}^n r_i'D_p^{-1}u\ f_{i.} = m_r'D_p^{-1}u$$

– Vector de proyecciones centrado

$$p_u = \begin{pmatrix} r_1' - m_r' \\ \vdots \\ r_n' - m_r' \end{pmatrix} D_p^{-1} u = \left(M_r - 1_n m_r'\right) D_p^{-1} u$$

– Varianza (ponderada) de las proyecciones sobre u

$$var\,(\text{proyec.}) = \sum_{i=1}^{n} \left(r_i' D_p^{-1} u - m_r' D_p^{-1} u\right)^2 \; f_{i.}$$

∗ Forma alternativa de expresar la varianza de las proyecciones

$$\begin{aligned} var\,(\text{proyec.}) &= p_u' D_n p_u \\ &= u' D_p^{-1} \left(M_r - 1_n m_r'\right)' D_n \left(M_r - 1_n m_r'\right) D_p^{-1} u \end{aligned}$$

2.2.1 Planteamiento del AC por filas

El problema que se plantea en el AC por filas es determinar los ejes $u_1, u_2, .., u_p$, unitarios y ortogonales, que maximicen la varianza (ponderada) de las proyecciones

- – Primer eje

$$\max_{u_1} \; p'_{u_1} D_n p_{u_1} = \max_{u_1} \; u'_1 D_p^{-1} \left(M_r - 1_n m'_r\right)' D_n \left(M_r - 1_n m_r^{'}\right) D_p^{-1} u_1$$

$$s.a. \;\; u'_1 D_p^{-1} u_1 = 1$$

  – Segundo eje

$$\max_{u_2} \; p'_{u_2} D_n p_{u_2}$$

$$s.a. \;\; u'_1 D_p^{-1} u_1 = 1, \;\; u'_1 D_p^{-1} u_2 = 0$$

  – Tercer eje ...

Nota 5 ***Relación entre la distancia*** $ji-cuadrado$ ***de dos filas y la distancia euclídea de las proyecciones sobre los ejes***

- *Los nuevos ejes* $u_1, ..., u_p$ *verifican*

$$\begin{pmatrix} u_1' \\ \vdots \\ u_p' \end{pmatrix} D_p^{-1} \begin{pmatrix} u_1 & \cdots & u_p \end{pmatrix} = I_p$$

$$\begin{pmatrix} u_1 & \cdots & u_p \end{pmatrix} \begin{pmatrix} u_1' \\ \vdots \\ u_p' \end{pmatrix} D_p^{-1} = \sum_{i=1}^{p} u_\alpha u_\alpha' D_p^{-1} = I_p$$

  - *La coordenadas de* r_i *sobre los nuevos ejes* $u_1, ..., u_p$ *son*

$$P_{r_i} : \left(r_i' D_p^{-1} u_1, ..., r_i' D_p^{-1} u_p \right)$$

  verificándose que

$$r_i = \sum_{\alpha=1}^{p} \left(r_i' D_p^{-1} u_\alpha \right) u_\alpha = \sum_{i=1}^{p} u_\alpha \left(u_\alpha' D_p^{-1} r_i \right) = \underbrace{\sum_{i=1}^{p} u_\alpha u_\alpha' D_p^{-1}}_{I_p} r_i$$

- Sean r_i' y $r_{i'}'$ dos filas de M_r (distribuciones condicionadas por filas)
  - El cuadrado de la distancia euclídea entre las proyecciones de dos filas P_{r_i} y $P_{r_i'}$ coincide con la distancia $ji - cuadrado$ entre las filas r_i y $r_{i'}$

$$\begin{aligned} d^2\left(P_{r_i}, P_{r_{i'}}\right) &= \left(P_{r_i} - P_{r_{i'}}\right)' \left(P_{r_i} - P_{r_{i'}}\right) \\ &= \sum_{\alpha=1}^{p} \left(r_i' D_p^{-1} u_\alpha - r_{i'}' D_p^{-1} u_\alpha\right) \left(r_i' D_p^{-1} u_\alpha - r_{i'}' D_p^{-1} u_\alpha\right) \\ &= (r_i - r_{i'})' D_p^{-1} \underbrace{\sum_{\alpha=1}^{p} u_\alpha u_\alpha' D_p^{-1}}_{I_p} (r_i - r_{i'}) \\ &= (r_i - r_{i'})' D_p^{-1} (r_i - r_{i'}) = d_{\chi^2}(r_i, r_{i'}) \end{aligned}$$

2.2.2 Solución al AC por filas respecto de la media

- El problema que se plantea es determinar los ejes $u_1, u_2, .., u_p$, unitarios y ortogonales, que maximicen la varianza (ponderada) de las proyecciones

  – Primer eje

$$\max_{u_1} \quad p'_{u_1} D_n p_{u_1} = u'_1 D_p^{-1} \left(M_r - 1_n m'_r\right)' D_n \left(M_r - 1_n m_r^{'}\right) D_p^{-1} u$$

$$s.a. \quad u'_1 D_p^{-1} u_1 = 1$$

  ∗ (Operando)

$$\max_{u_1} \quad p'_{u_1} D_n p_{u_1} = u'_1 \; D_p^{-1} \underbrace{\left(F' D_n^{-1} F D_p^{-1} - m_r 1'_p\right)}_{S_r} u_1$$

$$s.a. \quad u'_1 D_p^{-1} u_1 = 1$$

  – Segundo eje

$$\max_{u_2} \quad p'_{u_2} D_n p_{u_2}$$

$$s.a. \quad u'_1 D_p^{-1} u_1 = 1, \quad u'_1 D_p^{-1} u_2 = 0$$

  – Tercer eje ...

- A partir de los resultados recogidos en la Sección 8.1, la solución al AC por filas viene determinado por los autovalores-autovectores de la matriz

$$S_r = F' D_n^{-1} F D_p^{-1} - m_r 1_p' \qquad \left(S_r = M_r' M_c' - m_r 1_p'\right)$$

  - **Los ejes principales** son los autovectores $u_1, u_2, ..., u_p$ normalizados de S_r asociados a los autovalores $\lambda_1 \geq \lambda_2 ... \geq \lambda_p \geq 0$

$$u_\alpha' D_p^{-1} u_\alpha = 1 \quad y \qquad u_\alpha' D_p^{-1} u_{\alpha'} = 0, \quad \alpha \neq \alpha'$$

    * **Factores principales**

$$\varphi_\alpha = D_p^{-1} u_\alpha, \ \alpha = 1, ..., p \Longrightarrow \varphi_\alpha' D_p \varphi_\alpha = 1$$

- Las proyecciones de las filas de $M_r = D_n^{-1} F$ sobre el eje u_α se pueden expresar como

$$\begin{aligned} M_r D_p^{-1} u_\alpha &: \text{ vector de proyecciones de las filas sobre eje } \alpha \\ M_r \varphi_\alpha &: \text{ vector de proyecciones de las filas sobre eje } \alpha \end{aligned}$$

  - La varianza (ponderada) de las proyecciones sobre el eje α es

$$p_{u_\alpha}' D_n p_{u_\alpha} = u_\alpha' \ D_p^{-1} S_r u_\alpha = u_\alpha' \ D_p^{-1} \left(F' D_n^{-1} F D_p^{-1} - m_r 1_p'\right) u_\alpha = \lambda_\alpha$$

    En la sección 2.2.3 se demuestra que $m_r' D_p^{-1} u_\alpha = 0$, (la media ponderada de las proyecciones es 0), de donde

$$u_\alpha' D_p^{-1} F' D_n^{-1} F D_p^{-1} u_\alpha = \varphi_\alpha' F' D_n^{-1} F \varphi_\alpha = \varphi_\alpha' M_r' D_n M_r \varphi_\alpha = \lambda_\alpha \qquad (1)$$

2.2.3 Propiedades de los autovalores - autovectores y de las proyecciones

- El número de autovalores no nulos es $k = \min(n-1, p-1)$
- $(0, m_r)$ es autovalor-vector normalizado de S_r

$$\begin{aligned} S_r m_r &= 0_p \\ m_r' D_p^{-1} m_r &= m_r' 1_p = 1 \end{aligned}$$

- La proyección de cualquier fila sobre el eje m_r vale 1 (la varianza vale 0)

$$M_r D_p^{-1} m_r = D_n^{-1} F \underbrace{D_p^{-1} m_r}_{1_p} = D_n^{-1} \underbrace{F 1_p}_{m_c} = 1_n$$

- Cualquier otro u_α autovector de S_r distinto de m_r verifica que $u_\alpha' 1_p = 0$

$$u_\alpha' \underbrace{D_p^{-1} m_r}_{1_p} = 0 \rightarrow u_\alpha' 1_p = 0$$

- Las proyecciones de m_r
  - Sobre m_r : $m_r' D_p^{-1} m_r = 1$
  - Sobre otro eje: $m_r' D_p^{-1} u_\alpha = 0$
    - La media ponderada de las proyecciones sobre cualquier eje vale 0
    - La varianza ponderada de las proyecciones coincide con la suma de cuadrados ponderada de las proyecciones.

2.2.4 AC por filas respecto del origen

Se verifica la siguiente relación entre los autovalores-autovectores de $S_r = F'D_n^{-1}FD_p^{-1} - m_r 1_p'$ (datos centrados) y $F'D_n^{-1}FD_p^{-1}$ (datos sin centrar)

- $(1, m_r)$ es una autovalor-vector de $F'D_n^{-1}FD_p^{-1}$
- Si $(\lambda_\alpha, u_\alpha)$ es autovalor-vector de S_r, distinto de $(0, m_r)$, entonces $(\lambda_\alpha, u_\alpha)$ es autovalor-vector de $F'D_n^{-1}FD_p^{-1}$
- Para realizar el AC respecto de la media basta con realizar el análisis respecto del origen y desechar los resultados correpondientes a $(1, m_r)$ (Todas las proyecciones de las filas sobre m_r son iguales a 1, y por tanto se pueden desechar).

2.3 AC por columnas (respecto de la media)

- Consideramos las p distribuciones condicionadas por filas como p puntos de R^n
  - Los puntos son las p filas de la matriz $M_c = D_p^{-1}F' = \begin{pmatrix} c_1' \\ \vdots \\ c_n' \end{pmatrix}$
  - El peso de cada punto es $f_{.j}$, $j = 1, ..., p$
  - Media (ponderada) de las filas: $m_c' = \sum_{j=1}^{p} c_j' \ f_{.j} = \begin{pmatrix} f_{1.} & ... & f_{n.} \end{pmatrix}$
- Sea $v \in R^n$ de forma que $v' D_n^{-1} v = 1$ (norma asociada a la distancia $ji - cuadrado$)
  - El vector de proyecciones (según la distancia $ji - cuadrado$) de las filas sobre un eje $v \in R^n$ viene dado por

$$\begin{pmatrix} c_1' D_n^{-1} v \\ \vdots \\ c_p' D_n^{-1} v \end{pmatrix} = M_c D_n^{-1} v \in R^p$$

  - Media (ponderada) de las proyecciones sobre v

$$\sum_{j=1}^{p} c_j' D_n^{-1} v \ f_{.j} = m_c' D_n^{-1} v$$

– Vector de proyecciones centrado

$$p_v = \begin{pmatrix} c_1' - m_c' \\ \vdots \\ c_p' - m_c' \end{pmatrix} D_n^{-1} v = \left(M_c - 1_p m_c'\right) D_n^{-1} v$$

– Varianza (ponderada) de las proyecciones sobre u

$$var\,(\text{proyec.}) = \sum_{i=1}^{n} \left(c_j' D_n^{-1} v - m_c' D_n^{-1} v\right)^2 \; f_{.j}$$

Forma alternativa de expresar la varianza de las proyecciones

$$var\,(\text{proyec.}) = p_v' D_p p_v$$

- El problema del AC por columnas es determinar los ejes $v_1, v_2, ..., v_n$, unitarios y ortogonales, que maximicen la varianza ponderada de las proyecciones
  - Primer eje

$$\max_{v_1} \quad p'_{v_1} D_p p_{v_1}$$

$$s.a. \quad v_1' D_n^{-1} v_1 = 1$$

  - Segundo eje

$$\max_{v_2} \quad p'_{v_2} D_p p_{v_2}$$

$$s.a. \quad v_2' D_n^{-1} v_2 = 1, \quad v_2' D_n^{-1} v_1 = 0$$

  - Tercer eje ...

2.3.1 Solución al AC por columnas

- Consideremos las p distribuciones condicionadas por columnas como p puntos de R^n
  - Los puntos son las filas de la matriz $M_c = D_p^{-1}F'$
  - El peso de cada punto es $f_{.j},\ j = 1, ..., p$
- La solución al análisis de la columnas viene determinado por los autovalores-autovectores $(\lambda_\alpha^*, v_\alpha)$ de la matriz

$$S_c = FD_p^{-1}F'D_n^{-1} - m_c 1_n' \qquad (S_c = M_c'M_r' - m_c 1_n')$$

  - **Ejes principales** $v_1, ..., v_n$: autovectores normalizados de S_c $(v_\alpha' D_n^{-1} v_\alpha = 1)$
  - **Factores principales** ψ_α

$$\psi_\alpha = D_n^{-1} v_\alpha \Longrightarrow \psi_\alpha' D_n \psi_\alpha = 1$$

- Las proyecciones de las filas de $D_p^{-1}F'$ (perfiles columna) sobre el eje v_α se pueden expresar como

$$\begin{aligned} M_c D_n^{-1} v_\alpha &: \text{ proyecciones de las columnas sobre eje } \alpha \\ M_c \psi_\alpha &: \text{ proyecciones de las columnas sobre eje } \alpha \end{aligned}$$

- Además, véase Sección 8.1, se verifica que la varianza de las proyecciones

$$\begin{aligned} p_{v_\alpha}' D_p p_{v_\alpha} &= \lambda_\alpha^* \\ v_\alpha' D_n^{-1} F D_p^{-1} F' D_n^{-1} v_\alpha &= \psi_\alpha' F D_p^{-1} F_\alpha' \psi_\alpha = \lambda_\alpha^* \end{aligned}$$

3 Relación entre el AC por filas y el AC por columnas

3.1 Relación entre los autovalores-vectores de S_r y S_c

Proposición 6 *a)* $(0, m_c)$ *es un autovalor-vector normalizado de* S_c

b) Si (λ, u) *es autovalor-vector normalizado de* S_r*, distinto de* $(0, m_r)$*, entonces* $\left(\lambda, \frac{1}{\sqrt{\lambda}} F D_p^{-1} u\right)$ *es un autovalor-vector normalizado de* S_c*.*

Demostración.

a) Análogo a $(0, m_r)$ es un autovalor-vector normalizado de S_r

b) Veamos en primer lugar que $\left(\lambda, F D_p^{-1} u\right)$ es un autovalor-autovector de S_c

$$S_c \left(F D_p^{-1} u\right) = \left(F D_p^{-1} F' D_n^{-1} - m_c 1_n'\right) \left(F D_p^{-1} u\right) = \lambda \left(F D_p^{-1} u\right) ???? \qquad (2)$$

Ya que

$$\begin{aligned} S_r u &= \lambda u \rightarrow \left(F' D_n^{-1} F D_p^{-1} - m_r 1_p'\right) u = \lambda u \\ &\rightarrow F' D_n^{-1} F D_p^{-1} u = \lambda u \rightarrow \\ \left(F D_p^{-1}\right) F' D_n^{-1} F D_p^{-1} u &= \lambda \left(F D_p^{-1}\right) u \rightarrow F D_p^{-1} F' D_n^{-1} \left(F D_p^{-1} u\right) = \lambda \left(F D_p^{-1} u\right) \end{aligned}$$

Además

$$m_c \underbrace{1_n' F}_{m_r'} D_p^{-1} u = m_c \underbrace{m_r' D_p^{-1} u}_{0} = 0$$

Por tanto (2) se verifica.

- Sin embargo $FD_p^{-1}u$ no está normalizado (respecto de la norma considerada en R^n), ya que por (1)
$$\left(u'D_p^{-1}F'\right)D_n^{-1}\left(FD_p^{-1}u\right)=\lambda$$
- Por tanto los autovectores de norma 1 para el problema por columnas son de la forma ($\lambda \neq 0$)
$$v=\frac{1}{\sqrt{\lambda}}FD_p^{-1}u$$

■

Nota 7 *Conclusión*

- Los autovalores de S_r y S_c son los mismos
  - Los ejes tienen la misma representatividad para el conjuntos de puntos, filas y columnas
  - Esto es importante, cuando se representan en un mismo eje las filas y columnas
- Existe una relación entre los autovectores

3.2 Relación entre las coordenadas del AC por filas y columnas

- Relación entre los factores principales asociados a un mismo autovalor no nulo

$$\begin{aligned}\psi_\alpha &= D_n^{-1}v_\alpha = D_n^{-1}\frac{1}{\sqrt{\lambda_\alpha}}FD_p^{-1}u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}D_n^{-1}F\varphi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}M_r\varphi_\alpha \\ \varphi_\alpha &= \frac{1}{\sqrt{\lambda_\alpha}}D_p^{-1}F'\psi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}M_c\psi_\alpha\end{aligned}$$

- Relación entre las coordenadas de las proyecciones para un mismo eje

$$\begin{aligned}\text{coordenadas de las filas sobre el eje } \alpha &: M_r\varphi_\alpha = \sqrt{\lambda_\alpha}\psi_\alpha \\ \text{coordenadas de las columnas sobre el eje } \alpha &: M_c\psi_\alpha = \sqrt{\lambda_\alpha}\varphi_\alpha\end{aligned}$$

- **Ecuaciones de transición**
  - La coordenada de la fila i sobre un eje α, dada por $\sqrt{\lambda_\alpha}\psi_{\alpha_i}$, es una combinación lineal de las coordenadas de c_j sobre φ_α

$$\sqrt{\lambda_\alpha}\psi_{\alpha_i} = \sum_{j=1}^{p}\frac{f_{ij}}{f_{i.}}\varphi_{\alpha_j} = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{j=1}^{p}\frac{f_{ij}}{f_{i.}}\underbrace{\left(\sqrt{\lambda_\alpha}\varphi_{\alpha_j}\right)}_{(1)}$$

$$(1) \quad : \quad \text{Coordenada de } c_j \text{ sobre el eje } \alpha$$

  - Análogamente

$$\sqrt{\lambda_\alpha}\varphi_{\alpha_j} = \sum_{i=1}^{n}\frac{f_{ij}}{f_{.j}}\psi_{\alpha_i} = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{i=1}^{n}\frac{f_{ij}}{f_{.j}}\left(\sqrt{\lambda_\alpha}\psi_{\alpha_i}\right)$$

4 Medidas para la interpretación

4.1 Inercia de los ejes

- A λ_α (varianza de las proyecciones sobre el eje α) **se le denomina inercia del eje** α.

$$In\left(\alpha\right) = \lambda_\alpha$$

- Relación entre la inercia de los ejes y la inercial total
  - Elementos diagonales de S_r

$$s_{jj} = \sum_{i=1}^{n} \frac{f_{ij}^2}{f_{i.}f_{.j}} - f_{.j} = \sum_{i=1}^{n} \frac{\left(f_{ij} - f_{i.}f_{.j}\right)^2}{f_{i.}f_{.j}}$$

$$\text{Traza}\left(S_r\right) = \sum_{j=1}^{p}\sum_{i=1}^{n} \frac{\left(f_{ij} - f_{i.}f.j\right)^2}{f_{i.}f_{.j}} = \frac{1}{N}\chi^2 \quad \textbf{(Inercia total: InT)}$$

  - Suma de los autovalores de S_r verifica

$$\sum_{\alpha} \lambda_\alpha = \frac{1}{N}\chi^2$$

- $In\left(\alpha\right) = \lambda_\alpha$ representa la aportación del eje a la inercia total y refleja la importancia del eje α en la estructura factorial

$$InT = \sum_{\alpha} In\left(\alpha\right)$$

4.2 Contribución relativa de un punto a un eje

- Sea u_α un eje (autovector) del problema por filas correspondiente a λ_α y φ_α el factor correspondiente: $\varphi_\alpha = D_p^{-1} u_\alpha$

$$1 = u'_\alpha D_p^{-1} u_a = \varphi'_\alpha D_p \varphi_\alpha = 1 \rightarrow \sum_{j=1}^{p} f_{.j} \varphi_{\alpha j}^2 = 1$$

- Las proyecciones sobre el eje α de los perfiles columna vienen dadas por

$$M_c \psi_\alpha = \sqrt{\lambda_\alpha} \varphi_\alpha$$

  - La suma (ponderada cada una con peso $f_{.j}$) de los cuadrados de las proyecciones de las columnas coincide con la varianza (ponderada) de las proyecciones de las columnas (la media ponderada de las proyecciones es 0)

$$\begin{aligned} \left(\sqrt{\lambda_\alpha}\varphi_\alpha\right)' D_p \left(\sqrt{\lambda_\alpha}\varphi_\alpha\right) &= \lambda_\alpha \\ \sum_{j=1}^{p} f_{.j} \lambda_\alpha \varphi_{\alpha j}^2 &= \lambda_\alpha = In\,(\alpha) \end{aligned}$$

- Se define la **contribución relativa de la columna** c_j al eje α como

$$CR_\alpha(c_j) = \frac{f_{.j}\lambda_\alpha\varphi_{\alpha j}^2}{\lambda_\alpha} = f_{.j}\varphi_{\alpha j}^2$$

  $CR_\alpha(j)$ **representa la aportación de** c_j **a la inercia el eje** α, verificándose que

$$\sum_{j=1}^{p} CR_\alpha(c_j) = 1$$

- Análogamente se define la **contribución relativa de la fila** i al eje α como

$$CR_\alpha(r_i) = f_{i.}\psi_{\alpha i}^2$$

  verificándose

$$\sum_{i=1}^{n} CR_\alpha(r_i) = 1$$

- La CR es una medida de la importancia de un punto a la hora de explicar la inercia (λ_a) del eje
- Podemos asignar "un nombre al eje" en función de los puntos con mayor contribución.

4.2.1 Relación de las contribuciones a los ejes y la inercia del punto

En la siguiente sección se demuestra que

$$d_{\chi^2}\left(c_j, m_c\right) = d^2\left(P_{c_j}, P_{m_c}\right) = \sum_{\alpha} \lambda_\alpha \varphi_{\alpha_j}^2$$

Por tanto la inercia de la columna c_j

$$In\left(c_j\right) = f_{.j} d_{\chi^2}\left(c_j, m_c\right) = f_{.j} \sum_{\alpha} \lambda_\alpha \varphi_{\alpha_j}^2 = \sum_{\alpha} \lambda_\alpha CR_\alpha\left(j\right)$$

Análogamente

$$In\left(r_i\right) = f_{i.} d_{\chi^2}\left(r_i, m_r\right) = f_{i.} \sum_{\alpha} \lambda_\alpha \psi_{\alpha_i}^2 = \sum_{\alpha} \lambda_\alpha CR_\alpha\left(i\right)$$

4.3 Contribución (relativa) de un eje a un punto (Correlación entre el perfil y el eje)

- Consideremos un producto escalar, y la norma inducida por él.
  - El coeficiente de correlación entre dos vectores, y, z, viene dado por

$$r_{y,z} = \frac{\langle y - \overline{y}, z - \overline{z} \rangle}{\|y - \overline{y}\| \, \|z - \overline{z}\|} \qquad \text{(coseno de ángulo)}$$

- Consideremos en R^p el producto escalar: $\langle y, z \rangle = y' D_p^{-1} z$
  - Consideremos un punto columna centrado

$$c_j - m_c = \left(\frac{f_{1j}}{f_{.j}}, ..., \frac{f_{nj}}{f_{.j}} \right)' - (f_{1.}, ..., f_{n.})' \qquad \text{(media nula)}$$

  - Consideremos un eje del problema por columnas

$$v_a \qquad \text{(media nula)}$$

- **Se define la contribución del eje a a la columna** c_j como (el cuadrado del coeficiente de correlación entre $c_j - m_c$ y v_a (la media de ambos vectores es 0))

$$cr_\alpha(j) = \frac{\left((c_j - m_c)' D_p^{-1} v_a \right)^2}{(c_j - m_c)' D_p^{-1} (c_j - m_c) \quad \times v_a' D_p^{-1} v_\alpha}$$

– Teniendo en cuenta que las proyecciones de las columnas sobre el eje α vienen dadas por $\sqrt{\lambda_\alpha}\varphi_\alpha$ y que $m_c' D_p^{-1} v_a = 0$, se tiene

$$(c_j - m_c)' D_p^{-1} v_a = \sqrt{\lambda_\alpha}\varphi_{\alpha j} \tag{3}$$

– Por las condiciones sobre los ejes

$$v_a' D_p^{-1} v_\alpha = 1 \tag{4}$$

– Por otro lado, $(c_j - m_c)' D_p^{-1} (c_j - m_c)$ es la distancia ji-cuadrado de la columna c_j a m_c

$$d_{\chi^2}(c_j, m_c) = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - f_{i.} \right)^2 = (c_j - m_c)' D_p^{-1} (c_j - m_c)$$

∗ Por la relación entre la distancia $ji - cuadrado$ y la distancia euclídea de las proyecciones: $d_{\chi^2}(c_j, m_c) = d^2\left(P_{c_j}, P_{m_c}\right)$

Coordenadas de la proyección de c_j sobre los distintos ejes

$$P_{c_j} = \left(\sqrt{\lambda_1}\varphi_{1_j}, \sqrt{\lambda_2}\varphi_{2_j}, ..., \sqrt{\lambda_\alpha}\varphi_{\alpha_j}, ..., 1\right)$$

Coordenadas de la proyección de m_c sobre los distintos ejes

$$P_{m_c} = (0, 0, ..., 1)$$

· Por tanto

$$d_{\chi^2}(c_j, m_c) = d^2\left(P_{c_j}, P_{m_c}\right) = \sum_\alpha \lambda_\alpha \varphi_{\alpha_j}^2 \tag{5}$$

- De (3) , (4) y (5) , se tiene que

$$cr_{\alpha}\left(c_j\right) = \frac{\lambda_{\alpha}\varphi_{\alpha_j}^2}{d_{\chi^2}\left(c_j, m_c\right)} = \frac{\lambda_{\alpha}\varphi_{\alpha_j}^2}{d^2\left(P_{c_j}, P_{m_c}\right)}$$

  y representa la proporción debida al eje α en la posición de la columna c_j.

  Se verifica que

$$\sum_{\alpha} cr_{\alpha}\left(c_j\right) = 1$$

- Esta medida indica qué ejes representan bien a un punto
- **Contribución del hiperplano principal** (ejes considerados en el análisis, generalmente un plano) **a la columna** c_j

$$QLT\left(c_j\right) = \sum_{\alpha} cr_{\alpha}\left(c_j\right)$$

  $QLT\left(c_j\right)$ mide la calidad de la representación de c_j en el hiperplano principal. Puntos con bajo $QLT\left(c_j\right)$ están mal representados

- **Análogamente se define la contribución del eje** α a la fila r_i

$$cr_\alpha(r_i) = \frac{\lambda_\alpha \psi_{\alpha_i}^2}{d_{\chi^2}(r_i, m_r)} = \frac{\lambda_\alpha \psi_{\alpha_i}^2}{d^2(P_{r_i}, P_{m_r})}$$

- **Contribución del hiperplano principal** (ejes considerados en el análisis, generalmente un plano) **a la fila** r_i

$$QLT(r_i) = \sum_\alpha cr_\alpha(r_i)$$

  $QLT(r_i)$ mide la calidad de la representación de r_i en el hiperplano principal. Puntos con bajo $QLT(r_i)$ están mal representados

5 Representación de los resultados del AC

- Al igual en ACP, los resultados del AC se suelen representar sobre planos considerando los ejes principales 2 a 2
- El porcentaje de representatividad de cada eje $\lambda_\alpha / \sum \lambda$ da una idea conservativa de la proporción de información representada por el eje u_α. Es conservativa porque sólo es una forma parcial de medir la información. Puede ocurrir que aún con pequeños porcentajes pueda representarse toda la información contenida en el conjunto total (capítulo VII de Lebart)
- Ya que los autovalores del problema por filas y columnas son los mismos, los ejes de los dos problemas tienen la misma representatividad
- Dada la simetría del problema, conviene representar conjuntamente filas y columnas.
- Debido a la relación entre la distancia euclídea y la distancia ji-cuadrado,
  - Es válido interpretar la distancia euclídea entre las filas y entre las columnas.
  - Es válido interpretar la distancia euclídea de las filas o columnas al origen (distancia entre los perfiles y el centro de gravedad)

- Es válido interpretar la posición relativa de una fila (columna) respecto a todos los puntos columnas (fila).
- No debe interpretarse (salvo en casos especiales) la próximidad de un punto fila a un punto columna.

6 Complementos al AC

6.1 Elementos suplementarios

- Podemos representar sobre los ejes otros puntos (distribuciones) que no se han utilizado en la construcción y ver su posición relativa con respecto a las proyecciones de los perfiles filas o columnas.
- Si $r'_+ \in R^p$ es un punto fila suplementario sus proyecciones sobre los ejes son

$$r'_+\varphi_1, ..., r'_+\varphi_\alpha, ..., r'_+\varphi_k$$

- Si $c'_+ \in R^n$ es un punto columna suplementario sus proyecciones sobre son

$$c'_+\psi_1, ..., c'_+\psi_\alpha, ..., c'_+\psi_k$$

- En particular son de interés los puntos suplementarios que representan las distribuciones extremas (vértices) del análisis por filas y columnas
  - Vértices para el problema por filas: $vr_1 = (1, 0, ..., 0)'_{1\times p}, ..., vr_p = (0, ..., 0, 1)'_{1\times p}$
    - Coordenadas de las proyecciones de vr_j, $j = 1, ..., p$

$$\left(\varphi_{1j}, ..., \varphi_{\alpha j}, ..., \varphi_{kj}\right)$$

    - Nótese la relación de las coordenadas de la proyección de vr_j con las de la c_j, $j = 1, ..., p$

$$\begin{array}{rcl} & & \text{Coordenadas de la proyección de } c_j \text{ sobre los distintos ejes} \\ P_{c_j} & = & \left(\sqrt{\lambda_1}\varphi_{1_j}, \sqrt{\lambda_2}\varphi_{2_j}, ..., \sqrt{\lambda_\alpha}\varphi_{\alpha_j}, ..., 1\right) \end{array}$$

- Vértices para el problema por columnas $vc_1 = (1, 0, .., 0)'_{1\times n}, ..., vc_n = (0, ..., 0, 1)'_{1\times n}$
  - Coordenadas de las proyecciones de vc_i, $i = 1, ..., n$

$$\left(\psi_{1i}, ...\psi_{\alpha i}, ..., \psi_{ki)}\right)$$

  - Nótese la relación de las coordenadas de la proyección de vc_i con las de la r_i, $i = 1, ..., n$

$$P_{r_i} = \underbrace{\left(\sqrt{\lambda_1}\psi_{1_i}, \sqrt{\lambda_2}\psi_{2_i}, ..., \sqrt{\lambda_\alpha}\psi_{\alpha_i}, ..., 1\right)}_{\text{Coordenadas de la proyección de } r_i \text{ sobre los distintos ejes}}$$

- La representación de los vértices es útil para la interpretación del gráfico.
  - En general, a menor distancia de r_i a vr_j mayor f_{ij}
  - En SPSS, la normalización por filas representa las coordenadas de los perfiles de las filas y los vértices correspondientes.La normalización por columnas representa las coordenadas de los perfiles columnas y los vértices asociados.

6.2 Fórmula de reconstitución

- Autovalores-vectores de $S_r = F'D_n^{-1}FD_p^{-1} - m_r 1_p' : (0, m_r)$ y $(\lambda_\alpha, u_\alpha)$.
- Autovalores-vectores de $S_c = FD_p^{-1}F'D_n^{-1} - m_c 1_n' : (0, m_c)$ y $(\lambda_\alpha, v_\alpha) = \left(\lambda_\alpha, \frac{1}{\sqrt{\lambda_\alpha}} FD_p^{-1}u_\alpha\right)$
- Consideremos las matrices de datos sin centrar $S_r^* = F'D_n^{-1}FD_p^{-1}$ y $S_c^* = FD_p^{-1}F'D_n^{-1}$.
  - Autovalores-vectores de $S_r^* : (1, m_r)$ y $(\lambda_\alpha, u_\alpha)$
  - Autovalores-vectores de $S_c^* : (1, m_c) = \left(1, FD_p^{-1}m_r\right)$ y $(\lambda_\alpha, v_\alpha) = \left(\lambda_\alpha, \frac{1}{\sqrt{\lambda_\alpha}} FD_p^{-1}u_\alpha\right)$
  - Para todos los autovalores-vectores de S_r^* y S_c^* se verifica la relación

$$FD_p^{-1}u_\alpha = \sqrt{\lambda_\alpha}v_\alpha$$

Por tanto

$$\begin{aligned} \underbrace{FD_p^{-1}\sum_\alpha u_\alpha u_a'}_{I_p} &= \sum_\alpha \sqrt{\lambda_\alpha}v_\alpha u_\alpha' \\ F &= \sum_\alpha \sqrt{\lambda_\alpha}v_\alpha u_\alpha' = m_c m_r' + \sum_{\alpha(resto)} \sqrt{\lambda_\alpha}v_\alpha u_\alpha' \end{aligned}$$

De donde

$$\begin{aligned} D_n^{-1} F\ D_p^{-1} &= D_n^{-1}\left(m_c m_r' + \sum_{\alpha(resto)} \sqrt{\lambda_\alpha} v_\alpha u_\alpha' \right) D_p^{-1} = \\ &= 1_{n\times p} + \sum_{\alpha(resto)} \sqrt{\lambda_\alpha}\left(D_n^{-1} v_\alpha\right)\left(u_\alpha'\ D_p^{-1}\right) \\ &= 1_{n\times p} + \sum_{\alpha(resto)} \underbrace{\sqrt{\lambda_\alpha}\,(\psi_\alpha)}_{(1)}\underbrace{(\varphi_\alpha')}_{(2)} \end{aligned}$$

Igualando componente a componente

$$\frac{f_{ij}}{f_{i.} f_{.j}} - 1 = \sum_{\alpha(resto)} \underbrace{\sqrt{\lambda_\alpha}\,(\psi_{ai})}_{(1)}\underbrace{\left(\varphi'_{\alpha_j}\right)}_{(2)}$$

Coordenadas de la proyección de r_i : $\left(\sqrt{\lambda_1}\,(\psi_{1i}), ..., \sqrt{\lambda_k}\,(\psi_{ki})\right)$

Coordenadas de la proyección de vr_j : $\left(\varphi_{1_j}, ..., \varphi_{k_j}\right)$

$$\frac{f_{ij}}{f_{i.} f_{.j}} - 1 = \text{Producto escalar de las proyecciones del perfil } r_i \text{ y vértice } vr_j$$

6.3 Expresiones alternativas de las coordenadas

- Consideremos la matriz

$$S = D_n^{-1/2}\left(F - m_f m_c'\right) D_p^{-1/2} = \left(\frac{f_{ij} - f_{i.}f_{.j}}{\sqrt{f_{i.}f_{.j}}}\right) = \left(\frac{N_{ij} - e_{ij}}{\sqrt{Ne_{ij}}}\right)$$

- Según el teorema de descomposición espectral

$$S = U^* D_\lambda V^{*'}$$

donde

$$\begin{aligned}
U^* &= (u_1^*, ..., u_k^*) \text{ son los autovectores de } SS' \text{y } U^{*\prime}U^* = I \\
V^* &= (v_1^*, ..., v_k^*) \text{ son los autovectores de } S'S \text{ y } V^{*\prime}V^* = I \\
D_{\sqrt{\lambda}} &= diag\left(\sqrt{\lambda_1}, ..., \sqrt{\lambda_k}\right) \text{ con } \lambda_i \text{ los autovalores de } SS' \text{ y } S'S \\
k &= \min\left(n-1, p-1\right)
\end{aligned}$$

- Relación entre los autovalores y autovectores de SS', $S'S$ y S_r y S_c

$$\begin{array}{ccc||cc}
S'S & S_r = F'D_n^{-1}FD_p^{-1} - m_r 1_p' & \| & SS' & S_c = FD_p^{-1}F'D_n^{-1} - m_c 1_n' \\
\lambda & \lambda & \| & \lambda & \lambda \\
v^* & D_p^{1/2}v^* & \| & u^* & D_n^{1/2}u^*
\end{array}$$

- Expresión alternativa de las proyecciones

Proyecciones de las filas sobre eje α

Opción 1	
$\sqrt{\lambda_\alpha} D_n^{-1} v_\alpha$	v_α autov. de S_c
$\sqrt{\lambda_\alpha} D_n^{-1/2} \left(D_n^{-1/2} v_\alpha \right)$	$D_n^{-1/2} v_\alpha$ autov. de SS'

Opción 2	
$\sqrt{\lambda_\alpha} D_n^{-1/2} u_\alpha^*$	u_α^* autov. de SS'
$\sqrt{\lambda_\alpha} D_n^{-1} \left(D_n^{1/2} u_\alpha^* \right)$	$D_n^{1/2} u_\alpha^*$ autov. de S_c

Las proyecciones de las filas sobre los distintos ejes

$$R = D_n^{-1/2} U^* D_\lambda = \left(\sqrt{\lambda_1} D_n^{-1/2} u_1^*, \sqrt{\lambda_2} D_n^{-1/2} u_2^*, ..., \sqrt{\lambda_k} D_n^{-1/2} u_k^* \right)$$

- Análogamente

Proyecciones de las columnas sobre eje α

Opción 1

$\sqrt{\lambda_\alpha} D_p^{-1} u_\alpha$	u_α autov. de S_r
$\sqrt{\lambda_\alpha} D_p^{-1/2} \left(D_p^{-1/2} u_\alpha \right)$	$D_p^{-1/2} u_\alpha$ autov. de $S'S$

Opción 2

$\sqrt{\lambda_\alpha} D_p^{-1/2} v_\alpha^*$	v_α^* autov. de $S'S$
$\sqrt{\lambda_\alpha} D_p^{-1} \left(D_p^{1/2} v_\alpha^* \right)$	$D_p^{1/2} v_\alpha^*$ autov. de S_f

Las proyecciones de las columnas sobre los distintos ejes

$$C = D_p^{1/2} V^* D_\lambda = \left(\sqrt{\lambda_1} D_p^{-1/2} v_1^*, \sqrt{\lambda_2} D_p^{-1/2} v_2^*, ..., \sqrt{\lambda_k} D_p^{-1/2} v_k^* \right)$$

- Se verifica

$$R' D_n R = C' D_p C = D_{\sqrt{\lambda}}^2$$

7 Análisis de Correspondencia Múltiple

8 Apéndice

8.1 Maximización de una forma cuadrática bajo una restricción cuadrática

Proposición 8 *Sea A una matriz simétrica y M una matriz simétrica definida positiva*

- – La solución a
$$\max_{u_1} \quad u_1' A u_1$$
$$s.a. \quad u_1' M u_1 = 1$$
se alcanza en el autovector u_1 asociado al mayor autovalor de λ_1 de $M^{-1}A$ y además $u_1' A u_1 = \lambda_1$.
- – La solución a
$$\max_{u_2} \quad u_2' A u_2$$
$$s.a. \quad u_2' M u_2 = 1, \quad u_1' M u_2 = 0$$
se alcanza en el autovector u_2 asociado al segundo mayor autovalor de λ_2 de $M^{-1}A$ y además $u_2' A u_2 = \lambda_2$.
- – La solución a
$$\max_{u_3} \quad u_3' A u_3$$
$$s.a. \quad u_3' M u_3 = 1, \quad u_3' M u_2 = 0, \qquad u_3' M u_1 = 0$$
se alcanza en ...

Demostración

- Determinación del eje 1

$$\max_{u_1} u_1' A u_1$$
$$s.a. \;\; u_1' M u_1 = 1$$

  Multiplicadores de Lagrange

$$\begin{aligned} L &= u_1' A u_1 - \lambda (u_1' M u_1 - 1) \\ \frac{\partial L}{\partial u_1} &= 2Au_1 - 2\lambda M u_1 = 0 \Rightarrow Au_1 = \lambda M u_1 \Rightarrow u_1' A u_1 = \lambda \\ \frac{\partial L}{\partial \lambda} &= u_1' M u_1 - 1 = 0 \end{aligned}$$

  - Además, si M es definida positiva

$$M^{-1} A u_1 = \lambda u_1$$

    Por tanto u_1 es un autovector de $M^{-1}A$ correspondiente al autovalor λ.
  - Como además $u_1' A u_1 = \lambda$, se tiene u_1 debe ser el autovector asociado al mayor autovalor λ_1 de $M^{-1}A$.

- Determinación del eje 2

$$\max_{u_2} u_2' A u_2$$
$$s.a. \;\; u_2' M u_2 = 1, \;\; u_1' M u_2 = 0$$

– Multiplicadores de Lagrange

$$L = u_2' A u_2 - \lambda \left(u_2' M u_2 - 1 \right) - \mu_2 u_1' M u_2$$

–

$$\frac{\partial L}{\partial u_2} = 2Au_2 - 2\lambda M u_2 - \mu_2 M u_1 = 0$$

– De la determinación del eje 1 se tiene que

$$Au_1 = \lambda_1 M u_1 \rightarrow u_2' A u_1 = \lambda_1 u_2' M u_1 = 0$$

– Multiplicando $u_1' \frac{\partial L}{\partial u_2}$ y considerando la ecuación anterior

$$2u_1' A u_2 - 2\lambda u_1' M u_2 - \mu_2 u_1' M u_1 = 0 \rightarrow \mu_2 = 0$$

– Por tanto

$$Au_2 - \lambda M u_2 = 0 \rightarrow Au_2 = \lambda M u_2$$

– M definida positiva

$$M^{-1} A u_2 = \lambda u_2$$

– Multiplicando $u_2' \frac{\partial L}{\partial u_2}$

$$u_2' A u_2 = \lambda$$

– Por tanto la solución al problema viene dada por (λ_2, u_2) segundo mayor autovalor - vector de $M^{-1}A$, verificándose que

$$u_2' A u_2 = \lambda_2$$

8.2 Análisis general con cualquier distancia y cualquier criterio

- Sea X una matriz (n, p) de datos
- Sea M una matriz simétrica $(p \times p)$, definida positiva (define una distancia en R^p)
- Sea N una matriz $(n \times n)$ diagonal, cuyos elementos diagonales representan los pesos de los n elementos fila de X
- Sea $u \in R^p$ un vector de norma 1 :

$$u'Mu = 1$$

- Las n proyecciones de las filas de X sobre el eje u son los n filas del vector

$$p_u = XMu \in \mathbb{R}^n$$

- La suma ponderada de los cuadrados de las proyecciones sobre u viene dada por

$$p'_u N p_u = u'MX'NXMu$$

- Consideremos el problema de maximizar la suma de cuadrados ponderada de las proyecciones

$$\max_{u_1} \quad p'_{u_1} N p_{u_1} = \max_{u_1} \; u_1' M X' N X M u_1$$

$$s.a. \; u_1' M u_1 = 1$$

  - La solución es (λ_1, u_1) : mayor autovalor-vector de $X'NXM$.

- Consideremos el problema

$$\max_{u_2} \; u_2' M X' N X M u_2$$

$$s.a. \;\; u_2' M u_2 = 1, \quad u_1' M u_2 = 0$$

  - La solución es (λ_2, u_2) : segundo mayor autovalor-vector de $X'NXM$.

8.3 Nota

Sean $u_1, ..., u_p \in R^p$ y M una matriz simétrica definida positiva de forma que $u_i' M u_j = \delta_{ij}$.

Entonces $u_1, ..., u_p$ forman una base de R^p y

$$\forall x \in R^p \rightarrow x = \sum_{i=1}^{p} (x' M u_i)\, u_i = \underbrace{\sum_{i=1}^{p} u_i u_i' M x}_{I_p}$$

$$\begin{pmatrix} u_1' \\ \vdots \\ u_p' \end{pmatrix} M \,(u_1 ... u_p) = I_p = (u_1 ... u_p) \begin{pmatrix} u_1' \\ \vdots \\ u_p' \end{pmatrix} M = \sum_{i=1}^{p} u_i u_i' M$$

Las coordenadas de x en la nueva base son $x' M u_i$, $i = 1, ..., p$