

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
ESCUELA DE POSGRADO



EL ANÁLISIS DE CORRESPONDENCIAS CONJUNTO Y
MÚLTIPLE AJUSTADO

Tesis para optar el grado de Magíster en Estadística

AUTOR

Ricardo Elías Saavedra López

ASESOR

Dr. Luis Hilmar Valdivieso Serrano

JURADO

Dr. Sergio Camiz

Dr. Carlos Nilberto Veliz Capuñay

Dr. Luis Hilmar Valdivieso Serrano

LIMA - PERÚ

2012

Resumen

Esta tesis presenta una revisión de los fundamentos teóricos de dos de las más recientes extensiones de la técnica estadística conocida como análisis de correspondencia (AC): el análisis de correspondencia conjunto (ACC) y el análisis de correspondencia múltiple ajustado (ACMA); y muestra una aplicación práctica de éstas a una encuesta de egresados de la Pontificia Universidad Católica del Perú.

El análisis de correspondencia simple (ACS) es el primer alcance del análisis de correspondencias y se presenta cuando cada categoría de una variable se describe en función de la dependencia existente de los valores de otra única variable. Su extensión a más de 2 variables es conocida como el análisis de correspondencia múltiple (ACM).

Si bien se puede encontrar literatura sobre el ACS y el ACM, es importante destacar que el ACC y el ACMA han sido poco difundidos, encontrándose escasa literatura sobre el tema, más aún, en nuestro idioma. Por lo tanto, se hace necesaria una revisión de las dos primeras a modo de contexto y una presentación metodológica y detallada de las dos últimas.

Con la aplicación práctica se pretende obtener una representación de las facultades de los egresados de la PUCP en función del ingreso en su primer empleo relacionado con la formación recibida en la universidad y la percepción del grado de desarrollo de la competencia de comunicación recibida en la universidad. Esta aplicación consistirá en aplicar los 4 métodos descritos, comparándolos mediante nuevas técnicas que permiten reproducir las tablas de contingencia originales a partir de las representaciones obtenidas por los métodos indicados.

Palabras-clave: Análisis de correspondencias simple, análisis de correspondencias múltiple, análisis de correspondencias conjunto, Análisis de correspondencias múltiple modificado, matriz de Burt, inercia, biplot, anidamiento, tabla de contingencia, frecuencia marginal, descomposición en valores singulares de una matriz, valores singulares.

Índice general

1. Aspectos generales	1
1.1. Introducción	1
1.2. Objetivos	2
1.2.1. Objetivo general	2
1.2.2. Objetivos específicos	2
1.3. Organización del Trabajo	3
2. Análisis de correspondencias simple y múltiple	4
2.1. Introducción	4
2.2. Conceptos previos	4
2.2.1. Tabla de contingencia	4
2.2.2. Dependencia e independencia en tablas de contingencia	6
2.3. El análisis de correspondencias simple	7
2.3.1. Objetivos	7
2.3.2. Biplots	9
2.3.3. Inercia total	11
2.3.4. Identificación del número correcto de dimensiones	11
2.3.5. Contribuciones absolutas y relativas	12
2.3.6. Reglas de interpretación	13
2.4. El análisis de correspondencias múltiple	14
2.4.1. Análisis con la matriz binaria	14
2.4.2. Análisis con la matriz de Burt	15
2.4.3. Comparación del ACM de las matrices binaria y de Burt	17
2.4.4. Inercia de la matriz de Burt	17
2.4.5. Identificación del número correcto de dimensiones	17
3. Análisis de correspondencias múltiple ajustado y de correspondencias conjunto	19
3.1. Introducción	19
3.2. Análisis de correspondencias múltiple ajustado	19
3.3. Análisis de correspondencias conjunto	20
3.3.1. Formulación del problema	20
3.3.2. El algoritmo de mínimos cuadrados alternado	21
3.3.3. Soluciones iniciales basadas en el análisis de correspondencia múltiple	23
3.3.4. Inercia total en el ACC	24

4. Aplicaciones	25
4.1. Introducción	25
4.2. Caso de estudio	25
4.3. Alcance y objetivos de la aplicación	29
4.4. Tablas de contingencia	31
4.5. Aplicación del ACS	32
4.6. Aplicación del ACM, ACMA y ACC	38
4.6.1. Aplicación del ACM, ACMA y ACC - 2 variables	38
4.6.2. Aplicación del ACM, ACMA y ACC - 3 variables	47
5. Conclusiones	63
5.1. Conclusiones	63
5.2. Sugerencias para investigaciones futuras	65
A. Resultados teóricos	66
A.1. Álgebra matricial	66
A.2. Traza de una matriz	67
A.3. Autovalores y Autovectores	67
A.4. El teorema de descomposición en valores singulares y el teorema de Eckart y Young	68
B. Encuesta y más aplicaciones	74
B.1. Ficha técnica de la encuesta aplicada en el III Estudio Institucional sobre Egresados PUCP	74
B.1.1. Objetivo general	74
B.1.2. Objetivos específicos	74
B.1.3. Método de estudio	75
B.1.4. Población de estudio y cobertura	75
B.1.5. Ejes y variables de análisis	75
B.1.6. Técnica e instrumento	77
B.1.7. Procedimiento	77
B.2. Mas Aplicaciones	77
B.2.1. Aplicación del ACS	77
B.2.2. Aplicación del ACM	80
B.2.3. Aplicación del ACC	84
B.2.4. Aplicación del ACMA	87
Bibliografía	89

Capítulo 1

Aspectos generales

1.1. Introducción

El análisis de correspondencias simple (ACS) es una técnica estadística fundamentada en la descomposición en valores singulares de una matriz creada a partir de una tabla de contingencia de 2 variables, con el fin de representar los perfiles o frecuencias condicionales de estas variables en un mapa conjunto. Si bien el ACS es una técnica descriptiva, ella tiene sólo sentido bajo la hipótesis de dependencia entre las variables. Bajo esta suposición el ACS introduce una medida de dispersión para la nube de los perfiles conocida como inercia. En el ACS cada valor singular obtenido aporta un porcentaje del total de la inercia indicada. Esto quiere decir que a más dimensiones, mejor estarán representadas las categorías o perfiles de las variables. Aparte, una característica peculiar del ACS es que cuando una dimensión es aumentada al modelo, las anteriores se conservan, conociéndose esta propiedad como anidamiento.

La generalización del ACS a más de dos variables categóricas, y denominada análisis de correspondencias múltiples (ACM), no es un tema obvio ni bien definido, presentándose en la literatura diferentes enfoques para su tratamiento. Entre estos se puede citar los trabajos de Benzecri (1978), Greenacre (1984) y el análisis de homogeneidad de Gifi (1981).

El ACM se basa en la aplicación del análisis de correspondencia a la denominada matriz de Burt, la cual no es sino una matriz cuadrada simétrica formada por las tablas de contingencia del cruce de todos los pares de variables, en donde en la diagonal se hallan los cruces de las variables con ellas mismas. Esta codificación hace que los porcentajes de inercia de las dimensiones de los mapas en el ACM sean artificialmente bajos, lo que conlleva a una subestimación de la calidad de estos mapas.

En este proyecto se intentará profundizar en el estudio del ACM, presentándose dos modificaciones recientemente propuestas por Greenacre: una conocida como análisis de correspondencias conjunto (ACC), Greenacre (1988), y otra denominada análisis de correspondencias ajustado (ACM ajustado), Greenacre (2006). La ventaja de estos enfoques sobre el ACM radica de que ambas son extensiones naturales del ACS, al resultar este un caso particular de ambos (cosa que no ocurre con el ACM clásico) y, por otro lado, que ambas técnicas logran mejor ajustar las inercias explicadas a sus verdaderos valores a diferencia de los pesimistas valores encontrados en el ACM clásico.

El ACC busca disminuir el problema de sub-representatividad de las inercias, planteando en su modelo la eliminación de la influencia de las diagonales en la representación de los

perfiles obteniendo mapas perceptuales de mejor calidad; es decir con mayor inercia. Un problema sin embargo, con esta técnica radica en que a diferencia del ACM ella no es anidada.

Surge finalmente el ACM ajustado, que representa una solución simple, anidada, y que al mismo tiempo resuelve el problema de la baja inercia. Consiste en aplicar algunos ajustes a las inercias principales y a la inercia total del ACM.

En este trabajo se busca fundamentar y discutir todas estas técnicas así como aplicarlas a una data de los resultados de la encuesta a egresados en el 2006 de la PUCP y aplicada en el 2009, tres años después de su egreso; enfatizando las diferencias encontradas para cada uno de los análisis. Todas las técnicas mencionadas se encuentran implementadas en la librería ca de R como consta en Nenadic y Greenacre (2007).

Finalmente se planteará una comparación de los resultados de estas técnicas aplicando métodos recientes que intentan reproducir las tablas de contingencia originales a partir de la técnica aplicada. Estos métodos se citan en el trabajo de Camiz (2010).

1.2. Objetivos

1.2.1. Objetivo general

El objetivo general de la tesis es profundizar en el estudio de las modificaciones al análisis de correspondencias múltiples (ACM): el análisis de correspondencias conjunto (ACC) de Greenacre (1988) y el análisis de correspondencias ajustado (ACMA) de Greenacre (2006), exponer sus características y diferencias aplicando las técnicas a una data real con el apoyo de las implementaciones desarrolladas en R y comparar sus resultados mediante la reproducción de las tablas de contingencias originales en base al planteamiento de Camiz (2010).

1.2.2. Objetivos específicos

De manera específica:

- Revisar la literatura acerca del análisis de correspondencias simple (ACS) y el análisis de correspondencias múltiple (ACM).
- Revisar la literatura acerca del análisis de correspondencias conjunto (ACC) y el análisis de correspondencias múltiple ajustado (ACMA) y exponer sus diferencias con el ACM.
- Revisar la literatura acerca de las técnicas de reproducción de tablas de contingencia a partir de los resultados de un análisis de correspondencia en sus diferentes alcances.
- Aplicar las técnicas ACM, ACC y ACMA a una data real y exponer sus diferencias con especial énfasis en la reproductibilidad de las tablas de contingencias originales a partir de los resultados obtenidos por estas técnicas.

1.3. Organización del Trabajo

En el Capítulo 1 se presenta los aspectos generales de esta investigación y sus objetivos.

En el Capítulo 2 se hace una presentación del análisis de correspondencias simple (ACS) y el análisis de correspondencias múltiple (ACM) en base a una revisión de la literatura correspondiente con el fin de explicar los conceptos necesarios que se utilizarán posteriormente y unificar los distintos enfoques presentados para introducir el ACM.

El Capítulo 3 se centra en el desarrollo teórico del análisis de correspondencia conjunta (ACC) y del análisis de correspondencia múltiple ajustado (ACMA) enfatizando sus diferencias y similitudes entre ellos y con el ACS y el ACM.

El Capítulo 4 incluye una aplicación de las técnicas anteriormente descritas para determinar como los egresados de la PUCP, agrupados según sus facultades, se diferencian en función del ingreso en su primer empleo relacionado con la formación recibida en la universidad y la percepción del grado de desarrollo de la competencia de comunicación recibida en la universidad. La información es obtenida a partir de los resultados de la encuesta a egresados en el 2006 de la PUCP y aplicada en el 2009, tres años después de su egreso. Además se aplicarán los métodos recientes utilizados para reproducir las tablas de contingencias originales a partir de los resultados del análisis de correspondencias con el fin de obtener una medida de ajuste para estos métodos.

Finalmente, en el Capítulo 5 se discute algunas conclusiones obtenidas y se sugiere futuras investigaciones que se puedan derivar de este trabajo.

En el anexo se presenta algunos tópicos de álgebra matricial (Apéndice A) y otras aplicaciones desarrolladas (Apéndice B) sobre la información de la Encuesta de egresados indicada.

La tesis culmina presentando las referencias bibliográficas sobre las que se basa el desarrollo de este trabajo.

Capítulo 2

Análisis de correspondencias simple y múltiple

2.1. Introducción

El análisis de correspondencias simple (ACS) es una técnica multivariada descriptiva propuesta inicialmente por Hirschfeld (1935), y posteriormente desarrollada por Jean-Paul Benzécri (1973). Este es conceptualmente similar al análisis de componentes principales pero aplicado no a un conjunto de variables continuas sino a dos variables categóricas. El análisis busca fundamentalmente representar de una manera gráfica la relación entre estas variables a partir de su tabla de contingencia; es decir, de su distribución de frecuencias conjuntas de sus categorías. En este capítulo se hará una presentación del ACS y del análisis de correspondencias múltiple (Benzécri, 1978) (Greenacre, 1984) (Gifi, 1981), así como de las definiciones que se consideran necesarias para un buen entendimiento de los temas que serán desarrollados en esta tesis. En el apéndice 1 se detallan algunos elementos de álgebra matricial que proveerán una mejor comprensión a las técnicas aquí utilizadas.

2.2. Conceptos previos

2.2.1. Tabla de contingencia

Sea A e B dos variables categóricas con valores o categorías que serán denotadas por simplicidad por $1, \dots, I$ y $1, \dots, J$, respectivamente. Dadas n observaciones conjuntas de ambas variables, se acostumbra representar la distribución de frecuencias conjuntas de esta data mediante la tabla de contingencia del cuadro 2.1. o, equivalentemente, mediante la matriz $N = [n_{ij}]$, cuyas entradas n_{ij} denotan al número o frecuencia de observaciones que toman los valores $A = i$ y $B = j$.

		B					
		1	...	j	...	J	F. marginal de A
A	1	n_{11}	...	n_{1j}	...	n_{1J}	$n_{1.}$
	\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
	i	n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i.}$
	\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
	I	n_{I1}	...	n_{Ij}	...	n_{IJ}	$n_{I.}$
F. marginal de B		$n_{.1}$...	$n_{.j}$...	$n_{.J}$	n

Cuadro 2.1: Tabla de contingencia

Como se aprecia la tabla contiene en su última fila y columna las frecuencias marginales

$n_{i.} = \sum_{j=1}^J n_{ij}$ y $n_{.j} = \sum_{i=1}^I n_{ij}$ de respectivamente el número de observaciones con valores $A = i$ y $B = j$.

Al dividirse la tabla anterior entre n se obtiene la tabla de contingencia de frecuencia relativas o proporciones con matriz asociada

$$P = \frac{1}{n}N.$$

Esta tabla contiene en su última fila y columna los denominados perfiles marginales, los cuales se ilustran en los cuadros 2.2 y 2.3

1	...	i	...	I	Total
r_1	...	r_i	...	r_I	1

Cuadro 2.2: Perfil marginal de A

1	...	j	...	J	Total
c_1	...	c_j	...	c_J	1

Cuadro 2.3: Perfil marginal de B

siendo $r_i = \frac{n_{i.}}{n}$ la proporción de datos que toman como valor de $A = i$ y $c_j = \frac{n_{.j}}{n}$ la proporción de datos que toman como valor de $B = j$.

Otra manera de representar estos perfiles marginales es mediante las matrices diagonales

$$D_r = \begin{bmatrix} r_1 & & 0 \\ & \ddots & \\ 0 & & r_I \end{bmatrix} \quad \text{y} \quad D_c = \begin{bmatrix} c_1 & & 0 \\ & \ddots & \\ 0 & & c_J \end{bmatrix}.$$

Un elemento fundamental en el ACS lo constituyen los perfiles fila y los perfiles columna. Estos están constituidos por las frecuencias condicionales de cada una de las categorías de una de las variables con respecto a la otra. Más concretamente, los perfiles fila (I en total) describen las distribuciones condicionales de las distintas categorías de la variable B dada cada una de las I categorías de la variable A . Ellas se aprecian en el cuadro 2.4:

	B				
	1	...	j	...	J
f_1^A	$\frac{n_{11}}{n_{1.}}$...	$\frac{n_{1j}}{n_{1.}}$...	$\frac{n_{1J}}{n_{1.}}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots
f_i^A	$\frac{n_{i1}}{n_{i.}}$...	$\frac{n_{ij}}{n_{i.}}$...	$\frac{n_{iJ}}{n_{i.}}$
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots
f_I^A	$\frac{n_{I1}}{n_{I.}}$...	$\frac{n_{Ij}}{n_{I.}}$...	$\frac{n_{IJ}}{n_{I.}}$

Cuadro 2.4: Perfiles fila f_i^A

En adelante, la matriz de orden $I \times J$ cuyas filas corresponden a cada uno de los perfiles fila será denotada por F_A ; es decir,

$$F_A = D_r^{-1}P.$$

De otro lado, los perfiles columna (J en total) describen las distribuciones condicionales de las distintas categorías de la variable A dada cada una de las J categorías de la variable B . Ellas se aprecian en el cuadro 2.5.

		f_1^B	\vdots	f_j^B	\vdots	f_J^B
A	1	$\frac{n_{11}}{n_{.1}}$	\dots	$\frac{n_{1j}}{n_{.j}}$	\dots	$\frac{n_{1J}}{n_{.J}}$
	\vdots	\vdots	\ddots	\vdots	\ddots	\vdots
	i	$\frac{n_{i1}}{n_{.1}}$	\dots	$\frac{n_{ij}}{n_{.j}}$	\dots	$\frac{n_{iJ}}{n_{.J}}$
	\vdots	\vdots	\ddots	\vdots	\ddots	\vdots
	I	$\frac{n_{I1}}{n_{.1}}$	\dots	$\frac{n_{Ij}}{n_{.j}}$	\dots	$\frac{n_{IJ}}{n_{.J}}$

Cuadro 2.5: Perfiles columna f_j^B

En adelante, la matriz de orden $I \times J$ cuyas columnas corresponden a cada uno de los perfiles columna será denotada por F_B ; es decir,

$$F_B = PD_c^{-1}.$$

Desde un punto de vista geométrico todo perfil fila, que será representado por el vector columna

$$f_i^A = [f_{i1}^A \ f_{i2}^A \ \dots \ f_{iJ}^A]^\top,$$

siendo $f_{ij}^A = \frac{n_{ij}}{n_{i.}}$, puede verse como un vector de dimensión J . Análogamente, todo perfil columna

$$f_j^B = [f_{1j}^B \ f_{2j}^B \ \dots \ f_{Ij}^B]^\top,$$

siendo $f_{ij}^B = \frac{n_{ij}}{n_{.j}}$, puede verse como un vector de dimensión I . De esta manera la colección de todos los perfiles fila podría representarse como una nube de I puntos en el espacio \mathbb{R}^J (nube de perfiles fila) y la colección de perfiles columna como una nube de J puntos en el espacio \mathbb{R}^I (nube de perfiles columna). El objetivo del ACS es en esencia poder introducir alguna transformación de estos puntos en un espacio de menor dimensión de tal manera que tal representación facilite la interpretación de las relaciones existentes entre las variables en estudio.

2.2.2. Dependencia e independencia en tablas de contingencia

Hay que recordar que, desde un punto de vista inferencial, la existencia o no de algún tipo de relación entre las variables A y B puede analizarse mediante el estadístico ji-cuadrado de Pearson:

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^I \sum_{j=1}^J \frac{n_{i.}(f_{ij}^A - c_j)^2}{c_j} = \sum_{i=1}^I \sum_{j=1}^J \frac{n_{.j}(f_{ij}^B - r_i)^2}{r_i} \quad (2.1)$$

donde $e_{ij} = \frac{n_{i.}n_{.j}}{n} = nr_i c_j$ representa la frecuencia esperada de ser válida la hipótesis de independencia.

Intuitivamente, valores pequeños de χ^2 significan que los valores de n_{ij} y e_{ij} son cercanos

y, por tanto, se puede pensar que A y B son independientes. Por el contrario, valores grandes de χ^2 darían evidencia de que las variables están asociadas.

En este desarrollo se reconstruirá el valor de χ^2 pero no pensado como un estadístico de prueba sino más bien como una medida del grado de asociación, o más certeramente de dispersión de la nube de los perfiles fila y columna. Note que si en (2.1) $\chi^2 = 0$, entonces todos los perfiles fila (columna) resultan ser iguales entre sí e iguales al perfil marginal de B (A). En tal sentido para valores pequeños de χ^2 la nube de perfiles mostrará una configuración poco dispersa y centrada en su correspondiente perfil marginal.

Vale destacar que si bien el ACS es una técnica meramente descriptiva, ella puede también utilizarse en un estudio inferencial para describir la asociación existente de rechazarse la hipótesis de independencia.

2.3. El análisis de correspondencias simple

2.3.1. Objetivos

El análisis de correspondencias simple puede ser formulado como la aproximación ponderada por mínimos cuadrados de la tabla de contingencias $N = [n_{ij}]$ por otra matriz $\tilde{N} = [\tilde{n}_{ij}]$ con menor rango K , tal que

$$\tilde{N} = \arg \min_{H=[h_{ij}]} \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - h_{ij})^2}{e_{ij}} = \arg \min_H n^{-1} \text{tr}(D_r^{-1}(N - H)D_c^{-1}(N - H)^T), \quad (2.2)$$

donde los pesos son los inversos de las frecuencias esperadas $e_{ij} = nr_{i.}c_{.j}$.

Hay muchas maneras de justificar esta particular elección de los pesos a posteriori. Por ejemplo, la matriz resultante de aproximaciones provee una descomposición del estadístico ji-cuadrado. Una justificación a priori es asumir que las frecuencias tienen, bajo la suposición de independencia, una distribución de Poisson. Dado que las varianzas de las frecuencias en cada celda son iguales a sus valores esperados, se podría pensar que el peso sirve para estandarizar la varianza (Greenacre, 1988).

El hecho de que la matriz \tilde{N} tenga un menor rango permitirá además hallar subespacios de dimensiones bajas que mejor se ajusten a la representación de las categorías mostradas en N .

La matriz de aproximaciones \tilde{N} requerida puede obtenerse por una aplicación del teorema de Eckart-Young y más precisamente del corolario (A.4.1) que es probado en el apéndice. Para ello será conveniente escribir la función objetivo a minimizar en (2.2) como:

$$\Upsilon(H) = n \text{tr}(D_r^{-1/2}(Q - R)(Q - R)^T D_r^{1/2}),$$

donde

$$Q = \frac{1}{n} D_r^{-1/2} N D_c^{-1/2}.$$

y

$$R = \frac{1}{n} D_r^{-1/2} H D_c^{-1/2}. \quad (2.3)$$

Por las propiedades de traza de una matriz (véase el apéndice A), se tiene que

$$\Upsilon(H) = ntr((Q - R)(Q - R)^\top) = n\|Q - R\|_F,$$

donde $\|\cdot\|_F$ denota la norma de Frobenius de una matriz. El corolario A.4.1 dice entonces que, vista como función de R , se minimiza de tomarse

$$\tilde{R} = U_0 \Lambda_0^{1/2} V_0^\top,$$

donde

- U_0 es la matriz $I \times K$ conformada por los K primeros vectores singulares a izquierda de Q ,

-

$$\Lambda_0 = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_K \end{bmatrix}$$

es la matriz diagonal de los K primeros autovalores de $Q^\top Q$ y

- V_0 es la matriz $J \times K$ conformada por los K primeros vectores singulares a derecha de Q .

Por tanto la matriz \tilde{N} de rango K que minimiza (2.2) viene dada por la matriz inducida por \tilde{R} a través de (2.3):

$$\tilde{N} = nD_r^{1/2} \tilde{R} D_c^{1/2} = nD_r^{1/2} U_0 \Lambda_0^{1/2} V_0^\top D_c^{1/2}, \quad (2.4)$$

matriz que puede escribirse directamente como:

$$\tilde{N} = n \sum_{k=1}^K \sqrt{\lambda_k} D_r^{1/2} u_k v_k^\top D_c^{1/2}$$

donde u_1, u_2, \dots, u_K y v_1, v_2, \dots, v_K son, respectivamente, los K primeros vectores singulares a izquierda y derecha de Q .

La siguiente proposición, demostrada en el apéndice, indica las características del primer término en esta última descomposición.

Proposición 2.3.1 *El primer autovalor y los primeros vectores singulares de Q en (2.4) vienen dados por*

$$\lambda_1 = 1, \quad u_1 = D_r^{\frac{1}{2}} \mathbf{1}_I \quad \text{y} \quad v_1 = D_c^{\frac{1}{2}} \mathbf{1}_J,$$

donde $\mathbf{1}_I$ y $\mathbf{1}_J$ denotan, respectivamente, a los vectores columna de 1's de ordenes I y J .

De manera similar a lo visto en la demostración del corolario A.4.1 del apéndice es posible particionar las matrices U_0, Λ_0 y V_0 de la fórmula de reconstrucción (2.4) en dos partes, una

correspondiente al primer valor o vector singular y la otra conformada por las componentes restantes. Así se puede expresar \tilde{N} como:

$$\tilde{N} = nD_r^{1/2}\sqrt{\lambda_1}u_1v_1^\top D_c^{1/2} + nD_r^{1/2}U\Lambda^{1/2}V^\top D_c^{1/2},$$

donde u_1 y v_1 son respectivamente los primeros vectores singulares a izquierda y derecha asociados al primer autovalor λ_1 de $Q^\top Q$, U y V son las matrices de vectores singulares a izquierda y derecha descartando estos vectores singulares y Λ es la matriz diagonal de los primeros autovalores de $Q^\top Q$ excluyendo al valor λ_1 . Más aún, haciendo uso de la proposición anterior se tiene de reemplazar el primer término a derecha de esta descomposición que:

$$\tilde{N} = nrc^\top + nD_r^{1/2}U\Lambda^{1/2}\tilde{V}^\top D_c^{1/2},$$

donde nrc^\top es la matriz de frecuencias esperadas bajo independencia.

Puesto que $\tilde{N} = N$, si $K = r(N)$ se sigue que:

$$D_r^{-1/2}(P - rc^\top)D_c^{-1/2} = \frac{1}{n}D_r^{-1/2}(N - nrc^\top)D_c^{-1/2} = U\Lambda^{1/2}V^\top.$$

Esto muestra que la solución de correspondencias simples es equivalente a obtener la descomposición en valores singulares de la matriz de orden $I \times J$:

$$\bar{Q} = D_r^{-1/2}(P - rc^\top)D_c^{-1/2}, \quad (2.5)$$

matriz cuyas entrada (i, j) están caracterizada por:

$$\bar{q}_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{ne_{ij}}}.$$

2.3.2. Biplots

En el AC, los resultados se suelen representar geoméricamente mediante los llamados biplots. Para esto se usan las proyecciones de los perfiles filas y columnas centrados y escalados en un espacio de menor dimensión.

Si se denota por m al rango de la matriz \bar{Q} en (2.5), se podría particionar las matrices U , Λ y V en su descomposición de valores singulares mediante

$$\bar{Q} = \bar{U}\bar{\Lambda}^{1/2}\bar{V}^\top + \tilde{U}\tilde{\Lambda}^{1/2}\tilde{V}^\top = \bar{U}\bar{\Lambda}^{\alpha/2}\bar{\Lambda}^{(1-\alpha)/2}\bar{V}^\top + \tilde{U}\tilde{\Lambda}^{1/2}\tilde{V}^\top, \quad (2.6)$$

donde \bar{U} es la matriz $I \times K$ que contiene a los primeros K vectores singulares a izquierda de \bar{Q} , \tilde{U} los restantes $m - K$ vectores singulares a izquierda, \bar{V} es la matriz $J \times K$ que contiene los primeros K vectores singulares a derecha de \bar{Q} , \tilde{V} los restantes $m - K$ vectores singulares a derecha, $\bar{\Lambda}$ es la matriz diagonal con los primeros K valores singulares de $\bar{Q}^\top \bar{Q}$, $\tilde{\Lambda}$ la matriz diagonal con los restantes $m - K$ valores singulares y $\alpha \in [0, 1]$ un valor arbitrario.

Centrándose en los perfiles fila, note que:

$$D_r^{-1/2}\bar{Q} = D_r^{-1}(P - rc^\top)D_c^{-1/2} = (F_A - \mathbf{1}rc^\top)D_c^{-1/2} = \left[\frac{1}{\sqrt{c_j}}(f_{ij}^A - c_j)\right]$$

representa una matriz de I perfiles centrados y escalados en \mathbb{R}^J . La mejor representación de estos en \mathbb{R}^K puede obtenerse mediante la aproximación

$$D_r^{-1/2}\bar{Q} \simeq D_r^{-1/2}\bar{U}\bar{\Lambda}^{\alpha/2}\bar{\Lambda}^{(1-\alpha)/2}\bar{V}^\top = X\bar{\Lambda}^{(1-\alpha)/2}\bar{V}^\top, \quad (2.7)$$

donde la matriz de orden $I \times K$

$$X = D_r^{-1/2}\bar{U}\bar{\Lambda}^{\alpha/2},$$

contiene las coordenadas de los perfiles proyectados en \mathbb{R}^K .

Similarmente, la mejor representación de los perfiles columna centrados y escalados de \mathbb{R}^I

$$D_c^{-1/2}\bar{Q}^\top = (F_B^\top - \mathbf{1}jr^\top)D_r^{-1/2} = \left[\frac{1}{\sqrt{r_i}}(f_{ji}^B - r_i)\right]$$

en \mathbb{R}^K puede obtenerse mediante la aproximación

$$D_c^{-1/2}\bar{Q}^\top \simeq D_c^{-1/2}\bar{V}\bar{\Lambda}^{(1-\alpha)/2}\bar{\Lambda}^{\alpha/2}\bar{U}^\top = Y\bar{\Lambda}^{\alpha/2}\bar{U}^\top, \quad (2.8)$$

donde la matriz de orden $J \times K$

$$Y = D_c^{-1/2}\bar{V}\bar{\Lambda}^{(1-\alpha)/2},$$

contiene las coordenadas de los perfiles proyectados en \mathbb{R}^K .

Como se aprecia α es un parámetros que determina el tipo de coordenadas. Las elecciones más comunes para este parámetro se ilustran en la tabla siguiente

α	Coordenadas fila proyectados	Coordenadas columna proyectados
$\alpha = 0$	Estandarizadas: $X = D_r^{-1/2}\bar{U}$	Principales: $Y = D_c^{-1/2}\bar{V}\bar{\Lambda}^{1/2}$
$\alpha = 1$	Principales: $X = D_r^{-1/2}\bar{U}\bar{\Lambda}^{1/2}$	Estandarizadas: $Y = D_c^{-1/2}\bar{V}$
$\alpha = \frac{1}{2}$	Simétricas: $X = D_r^{-1/2}\bar{U}\bar{\Lambda}^{1/4}$	Simétricas: $Y = D_c^{-1/2}\bar{V}\bar{\Lambda}^{1/4}$

Los modelos correspondientes a $\alpha = 0$ y $\alpha = 1$ son referidos comúnmente como modelos asimétricos; mientras el modelo con $\alpha = \frac{1}{2}$ como simétrico. Otro modelo común involucra dos modelos asimétricos, uno con $\alpha = 1$ sólo para la filas y $\alpha = 0$ sólo para las columnas. Este modelo es conocido como el modelo simétrico Francés.

Note que las coordenadas principales de filas y columnas están escaladas de tal forma que $X^\top D_r X = Y^\top D_c Y = \bar{\Lambda}$; mientras que las coordenadas estandarizadas de filas y columnas (de allí su nombre) satisfacen $X^\top D_r X = Y^\top D_c Y = I_K$.

2.3.3. Inercia total

El usual estadístico ji-cuadrado χ^2 en la prueba la independencia (2.1) puede descomponerse como:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}} \right)^2 = \sum_{i=1}^I \sum_{j=1}^J n \bar{q}_{ij}^2 \\ &= ntr(\bar{Q}\bar{Q}^\top) = ntr(U\Lambda^{1/2}V^\top V(\Lambda^{1/2})^\top U^\top) \\ &= ntr(U\Lambda U^\top) = ntr(\Lambda) = n \sum_{k=1}^m \lambda_k.\end{aligned}$$

En la terminología del análisis de correspondencias a la cantidad

$$\mathcal{I}(N) = \sum_{k=1}^m \lambda_k = \frac{\chi^2}{n}$$

se le denomina la 'inercia total' de la tabla de contingencia N y a sus componentes λ_k las 'inerCIAS principales' (Greenacre,1984).

2.3.4. Identificación del número correcto de dimensiones

Como se comentó previamente la inercia representan una medida de dispersión de la nube original de perfiles. La inercia total y sus componentes son lo análogo a la varianza total y su descomposición con respecto a los ejes principales en el análisis de componentes principales.

Desde que la inercia total viene dada por la traza de las matrices $\bar{Q}\bar{Q}^\top$ ó $\bar{Q}^\top\bar{Q}$, las inercias de los perfiles centrados y ponderados proyectados en \mathbb{R}^K vienen dadas por estas mismas expresiones para sus aproximaciones (2.7) y (2.8), las cuales resultan ser iguales a:

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

En tal sentido, el cociente

$$\frac{\sum_{k=1}^K \lambda_k}{\sum_{k=1}^m \lambda_k}$$

es usualmente considerado como un criterio de bondad de ajuste en la aproximación obtenida.

La pregunta que surge entonces con el criterio anterior es cuántas dimensiones k considerar. De acuerdo con Camiz y Gomes (2012), esto puede ser resuelto aplicando una prueba clásica de bondad de ajuste (Kendall y Stuart (1961)) o más fácilmente el test de Malinvaud (1987). En el caso clásico, la prueba de Kendall puede ser aplicada secuencialmente al residuo de la reconstrucción parcial para cada dimension α considerada, mediante:

$$Q_\alpha = \sum_{ij} \frac{(n_{ij} - \tilde{n}_{\alpha ij})^2}{\tilde{n}_{\alpha ij}},$$

el cual sigue una distribución ji-cuadrado con $(r - \alpha - 1)(c - \alpha - 1)$ grados de libertad, siendo $\tilde{n}_{\alpha ij}$ el valor estimado de la celda (i, j) para la solución α -dimensional. Malinvaud

(1987) mostró posteriormente que, sustituyendo los valores estimados de la celda por aquellos esperados bajo la hipótesis de independencia, la fórmula se puede aproximar por:

$$\tilde{Q}_\alpha = \sum_{ij} \frac{(n_{ij} - \tilde{n}_{\alpha ij})^2}{nr_i c_j} = \chi^2 - n \sum_{\beta=1}^{\alpha} \lambda_\beta = n \sum_{\gamma=\alpha+1}^m \lambda_\gamma$$

que también se distribuye como Q_α . Entonces, si la prueba ji-cuadrado da valores significativos para la dimensión α , se deberá considerar esa dimensión en la solución del ACS, ya que indicaría que los residuos no se deben al azar.

Siguiendo a Camiz y Gomes (2012), se observa además que la misma propiedad podría ser asociada a la prueba parcial ji-cuadrado para la importancia de cada autovalor, $\chi_\alpha^2 = n\lambda_\alpha$ con $(r-1) + (c-1) - (2\alpha-1)$ grados de libertad. Esta prueba propuesta por Orlóci (1978), permite detectar si existe una relación lineal de ambas filas y columnas que explican el desvío de lo esperado. Si el valor de la prueba es significativo, entonces existe relación lineal y la dimensión debería de considerarse en la solución.

2.3.5. Contribuciones absolutas y relativas

Otro aspecto importante en el ACS es medir las contribuciones absolutas de los perfiles filas y columnas a la inercia de cada dimensión k a considerar y recíprocamente la contribución relativa de cada dimensión en un perfil.

Para la primera basta considerar que de las inercias de los perfiles proyectados en (2.7) y (2.8) se obtiene:

$$\sum_{k=1}^K \lambda_k = \text{tr}(D_r^{1/2} X \bar{\Lambda}^{(1-\alpha)/2} \bar{V}^\top \bar{V} \bar{\Lambda}^{(1-\alpha)/2} \bar{X}^\top D_r^{1/2}) = \sum_{i=1}^I \sum_{k=1}^K X_{ik}^2 \lambda_k^{1-\alpha} r_i.$$

En este sentido se llamará a

$$C_k(i) = \frac{X_{ik}^2 r_i}{\lambda_k^\alpha}$$

la contribución absoluta del perfil fila i en la inercia de la dimensión k . Análogamente se define la contribución absoluta del perfil columna j en la inercia de la dimensión k por

$$C_k(j) = \frac{Y_{jk}^2 c_j}{\lambda_k^\alpha}.$$

Estas contribuciones se utilizan para interpretar el significado de los ejes del biplot utilizando, para cada uno de ellos, las categorías o perfiles con contribuciones más altas.

Por otra parte, las contribuciones relativas son las contribuciones de un eje o dimensión k sobre la inercia del i -ésimo perfil fila y el j -ésimo perfil columna. Para el perfil fila i esta se define como:

$$\frac{X_{ik}^2}{\sum_{k=1}^m X_{ik}^2}$$

y para el perfil columna j como:

$$\frac{Y_{jk}^2}{\sum_{k=1}^m Y_{jk}^2},$$

donde X_{ik} e Y_{jk} representan las coordenadas de los perfiles fila i y de los perfiles columna j en la dimensión k , respectivamente.

Estas se obtienen a partir del cálculo del coseno al cuadrado del ángulo formado entre el perfil de interés y los ejes de las dimensiones proyectadas y miden la importancia de cada dimensión para explicar la posición, en el diagrama cartesiano, de cada una de las categorías de las variables analizadas, representando la parte de la distancia al origen de coordenadas, explicada por dicha dimensión o eje.

Se utilizan para analizar las proximidades entre los puntos o perfiles haciendo más hincapié en aquellas dimensiones cuyas contribuciones sean más elevadas a la hora de explicar dichas proximidades. Aquí interesan sobretudo los puntos que tienen un coseno cuadrado elevado en dicha dimensión. Las proximidades entre estos puntos, observados en el subespacio factorial correspondiente, dan una buena imagen de sus proximidades reales.

2.3.6. Reglas de interpretación

Las representaciones gráficas sobre los ejes o dimensiones $k = 1, 2, \dots, K$ de las I perfiles fila y de los J perfiles columna de la matriz de contingencia N proyectados son provistas por las coordenadas en las filas de las matrices X e Y . Generalmente, un biplot es satisfactorio si el porcentaje acumulado de inercia explicado por los primeros factores ejes es suficientemente grande. La interpretación de los gráficos es la siguiente:

- El punto $(0, 0)$ del gráfico corresponde al centro de gravedad de la nube de puntos; es decir, a los perfiles marginales o medios. De allí que valores alejados del centro del gráfico indican perfiles relativamente atípicos.
- La independencia se refleja en que los perfiles fila o columna son casi idénticos a los perfiles medios. En este caso la inercia es débil y no existen direcciones privilegiadas. Geométricamente ello significa que los puntos se concentran alrededor del centro de gravedad adoptando una forma esférica.
- La suma de las proporciones acumuladas de las primeras inercias principales se utiliza usualmente como criterio para determinar el valor de K ; es decir, el número de dimensiones sobre la cual proyectar la nube de perfiles original. se puede plantear también pruebas sobre este número como las discutidas en 2.3.4.
- La proximidad entre dos filas de las matrices X ó Y indica que la distribución de frecuencias condicionales de los correspondientes perfiles fila o columna son aproximadamente proporcionales. La interpretación opuesta se aplica cuando estos elementos están lejos una de la otra.
- Para encontrar un eventual significado a un eje o dimensión hay que interesarse en primer lugar por los puntos que tengan una mayor contribución absoluta. Estos puntos son los que fijan la posición del eje.
- Cuanto más próximo a 1 esté la contribución relativa de un eje o dimensión k sobre un perfil proyectado, más próxima estará la posición del perfil sobre el eje factorial k y la posición real del punto en el espacio. Por tanto, mayor será su calidad de representación,

así como el porcentaje de inercia del punto que viene explicado por el eje k . Si se han retenido K de los primeros ejes o dimensiones, se obtiene la calidad de la representación de por citar un perfil fila i en este sub-espacio como la suma de las contribuciones de estos ejes sobre este perfil.

2.4. El análisis de correspondencias múltiple

La generalización del AC simple, a más de dos variables categóricas y denominada análisis de correspondencias múltiples (ACM), no es un tema obvio ni bien definido, presentándose en la literatura diferentes enfoques para su tratamiento. Entre estos se puede citar los trabajos de Benzecri (1978), Greenacre (1984) y el análisis de homogeneidad de Gifi (1981).

2.4.1. Análisis con la matriz binaria

Una manera de generalizar el ACS cuando se disponen de Q variables categóricas es aplicándola a una matriz binaria o indicadora Z de orden $n \times J$, donde $J = \sum_{q=1}^Q J_q$, J_q denota al número de categorías de la q -ésima variable y n al número de casos. Esta matriz se forma por la concatenación de Q matrices binarias

$$Z = [Z_1 \ Z_2 \ \dots \ Z_Q],$$

siendo Z_q la matriz binaria de orden $n \times J_q$ que representa a la q -ésima variable y que tiene como entrada (i, j) al valor 1 si el i -ésimo caso toma la categoría j de esta variable o al valor 0 en caso contrario. La proposición siguiente indica la forma del estadístico ji-cuadrado para la matriz binaria Z en caso esta fuera tratada como una tabla de contingencia.

Proposición 2.4.1 *El estadístico ji-cuadrado asociado a la prueba de independencia sobre la matriz binaria Z viene dado por*

$$\chi^2 = n(J - Q).$$

Demostración: Se observa inicialmente que el estadístico ji-cuadrado asociado a $Z = [z_{ij}]$ puede descomponerse como:

$$\begin{aligned} \chi^2(Z) &= \sum_{i=1}^n \sum_{j=1}^J \frac{(z_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^n \frac{(z_{i1} - e_{i1})^2}{e_{i1}} + \sum_{i=1}^n \frac{(z_{i2} - e_{i2})^2}{e_{i2}} + \dots + \sum_{i=1}^n \frac{(z_{iQ} - e_{iQ})^2}{e_{iQ}} \\ &= \chi^2(Z_1) + \chi^2(Z_2) + \dots + \chi^2(Z_Q). \end{aligned}$$

Por tanto, bastará analizar el estadístico $\chi^2(Z_q)$ para un valor arbitrario de $q = 1, 2, \dots, Q$. En la definición de la inercia total se vio que

$$\chi^2(Z_q) = ntr(\bar{Q}^\top \bar{Q}),$$

donde en nuestro caso la matriz \bar{Q} vendrá dada por:

$$\bar{Q} = D_r^{-1/2}(P - rc^\top)D_c^{-1/2},$$

siendo $P = \frac{1}{n}Z_q$, $r = \frac{1}{n}\mathbf{1}_n$, $c = [c_1 \ c_2 \ \dots \ c_{J_q}]^\top$, $D_r = \frac{1}{n}I_n$, $D_c = \begin{bmatrix} c_1 & & 0 \\ & \ddots & \\ 0 & & c_{J_q} \end{bmatrix}$ y c_j la proporción de casos que toman la categoría j de la q -ésima variable.

En otras palabras

$$\bar{Q} = \frac{1}{\sqrt{n}}(Z_q D_c^{-1/2} - \mathbf{1}_n c^\top D_c^{-1/2}).$$

Así

$$\begin{aligned} \chi^2(Z_q) &= \text{tr}((Z_q D_c^{-1/2} - \mathbf{1}_n c^\top D_c^{-1/2})^\top (Z_q D_c^{-1/2} - \mathbf{1}_n c^\top D_c^{-1/2})) \\ &= \text{tr}(D_c^{-1/2} Z_q^\top Z_q D_c^{-1/2}) - 2\text{tr}(D_c^{-1/2} c \mathbf{1}_n^\top Z_q D_c^{-1/2}) + \text{tr}(D_c^{-1/2} c \mathbf{1}_n^\top \mathbf{1}_n c^\top D_c^{-1/2}) \\ &= n\text{tr}(D_c^{-1/2} D_c D_c^{-1/2}) - 2n\text{tr}(D_c^{-1} c c^\top) + n\text{tr}(D_c^{-1} c c^\top) = n(J_q - 1), \end{aligned}$$

y se tiene que $\chi^2(Z) = \sum_{q=1}^Q \chi^2(Z_q) = \sum_{q=1}^Q n(J_q - 1) = n(J - Q)$. \diamond

Se recuerda que la inercia en una tabla de contingencia se calcula como el estadístico ji-cuadrado dividido entre n . Sin embargo, dentro del contexto de la matriz binaria, n tendría que ser reemplazado por nQ . Así se tiene que la inercia de esta matriz viene dada por:

$$\mathcal{I}(Z) = \frac{\chi_Z^2}{nQ} = \frac{n(J - Q)}{nQ} = \frac{J - Q}{Q}.$$

2.4.2. Análisis con la matriz de Burt

Una estructura alternativa de datos para el ACM es tomar, en lugar de la matriz binaria Z , la matriz Burt B definida por:

$$\begin{aligned} B &= Z^\top Z = [Z_1 \ Z_2 \ \dots \ Z_Q]^\top [Z_1 \ Z_2 \ \dots \ Z_Q] \\ &= \begin{bmatrix} Z_1^\top Z_1 & Z_1^\top Z_2 & \dots & Z_1^\top Z_Q \\ Z_2^\top Z_1 & Z_2^\top Z_2 & \dots & Z_2^\top Z_Q \\ \vdots & \vdots & \ddots & \vdots \\ Z_Q^\top Z_1 & Z_Q^\top Z_2 & \dots & Z_Q^\top Z_Q \end{bmatrix} = \begin{bmatrix} D_1 & N_{12} & \dots & N_{1Q} \\ N_{12}^\top & D_2 & \dots & N_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ N_{1Q}^\top & N_{2Q}^\top & \dots & D_Q \end{bmatrix} \end{aligned}$$

y compuesta por todas las tablas de contingencia N_{qs} resultantes de cruzar todas las variables de interés dos a dos para $q \neq s$ y matrices diagonales D_q conteniendo las frecuencias marginales de las categorías de la variable q de cruzar esta variable consigo misma.

La otra forma 'clásica' de definir el ACM es efectuando el AC a la matriz de Burt. En tal caso un análisis similar al visto en la proposición anterior lleva a descomponer el estadístico ji-cuadrado de independencia para la matriz de Burt de la manera siguiente.

Proposición 2.4.2 *El estadístico ji-cuadrado asociado a la prueba de independencia sobre la matriz de Burt B viene dado por*

$$\chi^2(B) = \sum_{q \neq s} \sum \chi_{qs}^2 + n(J - Q),$$

donde χ_{qs}^2 denota al estadístico ji-cuadrado asociado a las tabla de contingencia N_{qs} .

Demostración: Claramente, por definición, el estadístico ji-cuadrado asociado a B puede descomponerse como la suma de los estadístico ji-cuadrado para las tablas de contingencia N_{qs} , con $q \neq s$ y las matrices D_q de la diagonal. Ello justifica el primer término a derecha en la proposición.

Para la obtención del segundo término basta observar que $D_q = nD_c$ y por tanto uno debería tomar para hallar $\chi^2(D_q) = ntr(\bar{Q}^\top \bar{Q})$, la matriz

$$\bar{Q} = D_r^{-1/2}(P - rc^\top)D_c^{-1/2},$$

con $P = D_c$, $r = c = [c_1 \ c_2 \ \dots \ c_{J_q}]^\top$ y $D_r = D_c = \begin{bmatrix} c_1 & & 0 \\ & \ddots & \\ 0 & & c_{J_q} \end{bmatrix}$ siendo c_j la proporción de casos que toman la categoría j de la q -ésima variable. Esto conduce a

$$\bar{Q} = D_c^{-1/2}(D_c - cc^\top)D_c^{-1/2} = I_{J_q} - (D_c^{-1/2}c)(D_c^{-1/2}c)^\top$$

y, desde que $D_c^{-1}c = \mathbf{1}_{J_q}$, a:

$$\chi^2(D_q) = n(tr(I_{J_q}) - 2tr(c^\top \mathbf{1}_{J_q}) + tr(\mathbf{1}_{J_q} c^\top \mathbf{1}_{J_q} c^\top)) = n(J_q - 1).$$

Luego la proposición se sigue de reemplazar estos términos en la descomposición de $\chi^2(B)$.
◇

Si bien podría parecer inconsistente el definir el ACM tanto como la aplicación del ACS a la matriz binaria Z o a la matriz de Burt B , es necesario aclarar que ambos análisis son en realidad equivalentes al proveer ellos exactamente las mismas coordenadas estándares para los perfiles columna proyectados y generar el análisis de Burt inercias principales que no son sino los cuadrados de estas inercias de trabajarse con la matriz binaria.

En efecto, si se aplica un ACS a la matriz binaria Z , se estaría básicamente obteniendo la descomposición en valores singulares de la matriz

$$\bar{Q}_Z = D_r^{-1/2}(P - rc^\top)D_c^{-1/2},$$

donde en este contexto $D_r = \frac{1}{n}I_n$, $P = \frac{1}{nQ}Z$, $r = \frac{1}{n}\mathbf{1}_n$, $D_c = \frac{1}{Q}D$ y c es el vector columna conteniendo a los perfiles marginales columna de las categorías de todas las Q variables en estudio, siendo D la matriz diagonal $J \times J$ conformada por precisamente estos perfiles marginales.

Un poco de álgebra lleva a escribir

$$\bar{Q}_Z = \frac{1}{\sqrt{nQ}}(Z - \mathbf{1}_n c^\top)D^{-\frac{1}{2}}.$$

Así

$$\begin{aligned} \bar{Q}_Z^\top \bar{Q}_Z &= \frac{1}{nQ^2}(D^{-\frac{1}{2}}(Z^\top Z - Z^\top \mathbf{1}_n c^\top - c \mathbf{1}_n^\top Z + c \mathbf{1}_n^\top \mathbf{1}_n c^\top)D^{-\frac{1}{2}}) \\ &= \frac{1}{nQ^2}(D^{-\frac{1}{2}}(Z^\top Z - 2ncc^\top + ncc^\top)D^{-\frac{1}{2}}) = D^{-\frac{1}{2}}\left(\frac{1}{nQ^2}Z^\top Z - \frac{1}{Q^2}cc^\top\right)D^{-\frac{1}{2}}. \end{aligned}$$

Esta última es precisamente la matriz \bar{Q}_B requerida en la descomposición de valores singulares de aplicar el ACS a la matriz de Burt B , desde que la suma de cualquier fila o columna de B es igual nQ veces el perfil marginal correspondiente esta fila o columna y la suma total de valores en la tabla de Burt es nQ^2 .

En consecuencia, si $\bar{Q}_Z = U\Lambda^{1/2}V^T$ es la descomposición en valores singulares para la matriz binaria, se tiene que

$$\bar{Q}_B = \bar{Q}_Z^T \bar{Q}_Z = V\Lambda V^T.$$

2.4.3. Comparación del ACM de las matrices binaria y de Burt

Las dos formas de definir el ACM se relacionan de la siguiente manera:

- En los dos análisis, las coordenadas estándares son idénticas.
- Las inercias principales del análisis de Burt son los cuadrados de los de la matriz binaria.
- Dado que las inercias principales son menores que 1, sus cuadrados darán valores más pequeños. En consecuencia, los porcentajes de inercia serán siempre mayores en el análisis de Burt.

2.4.4. Inercia de la matriz de Burt

Todas las subtablas que componen la matriz de Burt tienen los mismos valores marginales totales de filas y columnas. La inercia de B será la media de las inercias de las subtablas B_{qs} , que lo componen. Los altos valores de las inercias de las matrices de la diagonal de la matriz de Burt explican porqué su inercia total es tan alta y por lo tanto también explica los bajos porcentajes de la inercia de los ejes. Esto provoca una subestimación de la calidad de los mapas que resulten de este tipo de análisis.

2.4.5. Identificación del número correcto de dimensiones

Nuevamente, la pregunta que surge es cuántas dimensiones considerar. De acuerdo con Camiz y Gomes (2012), esto puede ser resuelto aplicando el argumento sugerido por Benzécri (1979) y Greenacre (1988, 2006) que sostiene que hay que limitar la atención solamente a los autovalores de la matriz binaria mayores que su media. Se prueba que dicha media es $1/Q$.

Siendo los autovalores de la matriz de Burt las raíces de los autovalores de la matriz binaria, entonces solo debería considerarse las dimensiones que cumplan:

$$\sqrt{\lambda_k} > \frac{1}{Q}$$

Por otra parte, Ben Ammou y Saporta (1998, 2003) sugieren estimar la significatividad de los autovalores del ACM (nuevamente considerando la matriz binaria) de acuerdo a su distribución. Para esto muestran que bajo la hipótesis de independencia por pares, la medida de dispersión de los autovalores en el ACM:

$$S_{\lambda}^2 = \frac{1}{k} \sum_{k=1}^K \left(\lambda_k - \frac{1}{Q} \right)^2$$

satisface que

$$\sigma^2 = E[S_{\lambda}^2] = \frac{1}{nQ^2(J-Q)} \sum_{i \neq j} (J_i - 1)(J_j - 1)$$

y consideran que el intervalo $\frac{1}{Q} \pm 2\sigma$ debería de contener aproximadamente alrededor del 95 % de los autovalores.



Capítulo 3

Análisis de correspondencias múltiple ajustado y de correspondencias conjunto

3.1. Introducción

El análisis de correspondencia múltiple, de manera similar al análisis de correspondencia simple, puede verse como una aproximación ponderada por mínimos cuadrados de la matriz de Burt B . El estadístico ji-cuadrado puede ser nuevamente calculado para B como si fuera una tabla de contingencia, lo que se simplifica como se vió en la proposición 2.4.2 en

$$\chi^2(B) = \sum \sum_{q \neq s} \chi_{qs}^2 + n(J - Q),$$

siendo χ_{qs}^2 el estadístico ji-cuadrado para las subtablas de contingencia fuera de la diagonal, $\sum \sum_{q \neq s}$ la doble suma con índices distintos, Q el número de variables en el análisis y J el número de categorías de todas estas variables. Como se aprecia, no hay justificación en esta medida de dispersión de la nube de puntos para ajustar las subtablas en la diagonal de B que contribuyen con el término $n(J - Q)$. De hecho, la presencia de este término infla artificialmente la inercia total al extremo de que los porcentajes alcanzados por los ejes principales pueden ser muy bajos, especialmente cuando $J - Q$ es grande. Una forma más natural de medir la dispersión de la nube de puntos podría ser considerando la suma $\sum \sum_{q \neq s} \chi_{qs}^2$. Esto sugiere una generalización alternativa del análisis de correspondencia que ajusta solamente las tablas fuera de la diagonal, de manera análoga al análisis factorial en donde los valores de la diagonal de la covarianza o matriz de correlación no son de interés directo. Es así que aparecen las alternativas del ACM conocidas como análisis de correspondencia múltiple modificado (ACMA) y el análisis de correspondencia conjunto (ACC).

3.2. Análisis de correspondencias múltiple ajustado

Es posible remediar parcialmente el problema del porcentaje de inercia de un ACM, usando un simple reajuste de escala en la solución ACM. En esta aproximación, que permite obtener una solución anidada, el total de inercia es medido por el promedio de inercia de todos los bloques fuera de la diagonal de B , calculada directamente a partir de las tablas mismas o ajustando la inercia total de B retirando las contribuciones fijas de los bloques de la diagonal tal como se muestra a continuación:

Suma de las inercias de las Q tablas de la diagonal = $J - Q$.

Suma de las inercias de todas las tablas de doble entrada = $Q^2\mathcal{I}(B)$.

La inercia promedio de los bloques fuera de la diagonal (o inercia total ajustada $\mathcal{I}_A(B)$) se obtiene restando la suma de las inercias de las tablas de la diagonal de la suma de las inercias de todas las tablas de doble entrada y dividiéndola entre el total de las tablas fuera de la diagonal ($Q(Q-1)$):

$$\mathcal{I}_A(B) = \frac{Q}{Q-1}(\mathcal{I}(B) - \frac{J-Q}{Q^2}) \quad (3.1)$$

Partes de la inercia (usando la fórmula de Benzécri (1979), citada en Greenacre (1984, p.145)) son entonces calculadas a partir de las inercias principales de la matriz de Burt B :

$$\lambda_k^{aj} = \left(\frac{Q}{Q-1}\right)^2(\sqrt{\lambda_k} - \frac{1}{Q})^2, k = 1, 2, \dots \quad (3.2)$$

Aquí λ_k es la k -ésima inercia principal de la matriz de Burt B . Los ajustes se hacen solamente en las dimensiones para las cuales $\sqrt{\lambda_k} > \frac{1}{Q}$, y por lo tanto no se usan más dimensiones; así, los porcentajes de inercia no suman el 100%. Note que λ_k^{aj} es cercano a λ_k cuando el valor de Q es grande. Estos porcentajes mejoran dramáticamente los resultados de un ACM y son recomendados en su aplicación.

La solución ajustada tiene las mismas coordenadas estándar de la matriz de Burt pero usa las inercias principales ajustadas para calcular las coordenadas principales, consiguiéndose una calidad mejorada en su representación gráfica.

3.3. Análisis de correspondencias conjunto

El análisis de correspondencias conjunto es un algoritmo iterativo que lleva a cabo el análisis de correspondencia de la matriz de Burt, buscando el ajuste solo de las tablas fuera de la diagonal. En este procedimiento, se parte del resultado del ACM para, a continuación, sustituir las tablas de la diagonal por valores estimados a partir de estos resultados.

3.3.1. Formulación del problema

Dada una matriz de Burt B , el análisis de correspondencia múltiple podría ser definido como la aproximación ponderada de mínimos cuadrados de B por otra matriz de menor rango, que minimice:

$$\psi(H) = b^{-1}tr \left(D_B^{-1}(B-H)D_B^{-1}(B-H)^\top \right), \quad (3.3)$$

donde $b = nQ^2$ es el gran total de B y D_B es la matriz diagonal de todos los perfiles marginales de filas y/o columnas (ya que B es simétrica). Este es el análisis de correspondencia simple de B definido en (2.2). El vector de J_q perfiles marginales de la variable q será ahora denotado por r_q , según lo cual $\mathbf{1}_{J_q}^\top r_q = 1$. Este es también el conjunto de perfiles marginales filas de las tablas de contingencia individuales N_{qs} para cualquier s . La matriz diagonal $J_q \times J_q$ formada a partir de los elementos de r_q será denotada ahora por D_q . La expresión (3.3) puede ser

reescrita como:

$$n^{-1} \text{tr} \left(D^{-1} (B - H) D^{-1} (B - H)^{\top} \right) = n^{-1} \sum_{q=1}^Q \sum_{s=1}^Q \text{tr} \left(D_q^{-1} (N_{qs} - H_{qs}) D_s^{-1} (N_{qs} - H_{qs})^{\top} \right), \quad (3.4)$$

siendo $D = QD_B$ una supermatriz diagonal con matrices entrada D_1, D_2, \dots, D_Q y H una supermatriz con submatrices H_{qs} de orden $J_q \times J_s$.

Para simplificar se puede utilizar la notación de norma:

$$\|A - B\|_{qs}^2 = \text{tr} (D_q^{-1} (A - B) D_s^{-1} (A - B)^{\top}),$$

lo cual simplifica (3.4) a

$$n^{-1} \sum_q \sum_s \|N_{qs} - H_{qs}\|_{qs}^2.$$

La función a ser minimizada por el nuevo método consistirá en la suma de los $\frac{1}{2}Q(Q-1)$ términos $n^{-1} \sum_{q < s} \|N_{qs} - H_{qs}\|_{qs}^2$ correspondientes a las subtablas fuera de la diagonal en el triángulo superior de B . Para $Q = 2$ esto consistirá en un solo término y es idéntico a (2.2) dado que $D_1 = D_r$, $D_2 = D_c$ y $N_{12} = N$. En consecuencia el análisis de correspondencia simple es un caso particular de este método, cosa que no ocurre con el ACM.

Se sabe que la matriz H_{qs} de rango 1 es la matriz de frecuencias esperadas nrr^{\top} , donde r denota ahora al super-vector columna de los $r_q (q = 1, \dots, Q)$. Esto corresponde a la misma solución trivial de la matriz de Burt. En cualquier análisis de correspondencia, en particular en el análisis de correspondencias de B, todas las dimensiones están 'anidadas', esto es, la solución de un cierto rango está incluida en soluciones de rango mayor. Si la solución trivial esta anidada o no en soluciones de mayor dimensión es irrelevante ahora, ya que el interés es descomponer los estadísticos ji-cuadrado; es decir, es necesario ajustar las desviaciones de la solución trivial $N_{qs} - nr_q r_s^{\top}$. El problema entonces se puede mejor formular como la minimización de la función de discrepancia:

$$n^{-1} \sum_{q < s} \|N_{qs} - nr_q r_s^{\top} - L_{qs}\|_{qs}^2 \quad (3.5)$$

sobre la clase de todas las super-matrices L , cuyas matrices entrada L_{qs} tienen un rango predefinido K .

3.3.2. El algoritmo de mínimos cuadrados alternado

Puesto que el ACM equivale a una aplicación del ACS a la matriz de Burt B , un refraseo de (2.6) lleva a la siguiente descomposición de esta:

$$B - nrr^{\top} = nD_B^{\frac{1}{2}} \bar{U} \bar{\Lambda} \bar{U}^{\top} D_B^{\frac{1}{2}} + nD_B^{\frac{1}{2}} \tilde{U} \tilde{\Lambda} \tilde{U}^{\top} D_B^{\frac{1}{2}},$$

donde el primer término en el lado derecho de esta ecuación es una matriz de rango K y \bar{U} es una super-matriz cuyas entradas contienen a los primeros vectores singulares (que definen las coordenadas de los perfiles proyectados) asociados a los primeros K valores singulares en $\bar{\Lambda}$ de la matriz simétrica $D_B^{-\frac{1}{2}} (\frac{1}{n} B - rr^{\top}) D_B^{-\frac{1}{2}}$. La solución ACM se obtiene simplemente de

eliminar el segundo término a derecha en esta última ecuación y se lee como:

$$B - nrr^T \approx nD \frac{1}{2} \bar{U} \bar{\Lambda} \bar{U}^T D \frac{1}{2}. \quad (3.6)$$

Con el fin de poder ajustar esta solución y mitigar el efecto de las submatrices en la diagonal, el análisis de correspondencias conjunto (ACC) propone una aproximación del tipo

$$B - nrr^T \approx L + C, \quad (3.7)$$

donde L es una matriz de rango K que, de manera similar a (3.6), toma la forma

$$L = nDX\Lambda_\beta X^T D$$

y C es una matriz diagonal de bloques con submatrices C_{qq} ($q = 1, \dots, Q$) a lo largo de la diagonal y ceros en el resto. En la descomposición de L la matriz de orden $J \times K$, X , determinará básicamente las nuevas coordenadas de los perfiles proyectados, los cuales se asumirán estandarizados en el sentido que $X^T DX = QI$, y la matriz diagonal de orden K , Λ_β , contendrá un parámetro de escala para cada dimensión. Los elementos de cada C_{qq} son parámetros ficticios que permiten el ajuste de las submatrices en la diagonal de $B - nrr^T$ a fin de eliminar su influencia en el modelo de interés.

El presente problema de ajustar las submatrices fuera de la diagonal tiene que ver con el análisis factorial por mínimos cuadrados en dónde los elementos fuera de la diagonal de una matriz de correlación están ajustados. Este método de análisis factorial, comunmente llamado MINRES, ha sido discutido en el contexto de aproximaciones de matrices de rango menor en dónde los pesos cero son asignados a los elementos de la diagonal.

Vale apreciar que si se particiona X en filas de acuerdo a las variables X_1, \dots, X_Q , donde X_q es de orden $J_q \times K$, las submatrices de L vienen dadas por $L_{qs} = nD_q X_q \Lambda_\beta X_s^T D_s$. En términos de estas matrices la minimización de

$$n^{-1} \sum_{q \neq s} \|N_{qs} - nr_q r_s^T - L_{qs}\|_{qs}^2 + n^{-1} \sum_q \|N_{qq} - nr_q r_q^T - L_{qq} - C_{qq}\|_{qq}^2 \quad (3.8)$$

sobre la clase de todas las matrices X , Λ_β y C que satisfacen las condiciones anteriores, es equivalente a minimizar (3.5) desde que el último conjunto de términos en (3.8) puede hacerse siempre cero de tomarse C_{qq} igual a $N_{qq} - nr_q r_q^T - L_{qq}$. Note que (3.8) involucra la suma sobre las sub-tablas fuera de la diagonal en ambos superior e inferior triángulos de B , contando dos veces cada sub-tabla.

La minimización de (3.8) puede ejecutarse iterativamente alternando entre las variables en C y aquellas en X y Λ_β como sigue:

- (1) Comenzar con una solución tentativa para X y Λ_β (a verse en la siguiente sub-sección).
- (2) Para los X y Λ_β dados en (1), minimizar C . Esto se consigue simplemente asignando para cada $q = 1, \dots, Q$:

$$C_{qq} = N_{qq} - nr_q r_q^T - nD_q X_q \Lambda_\beta X_q^T D_q.$$

- (3) Manteniendo fijo el C anterior, minimizar con respecto a X y Λ_β . Esto se consigue ejecutando un ACS a la matrix $B^* = B - C$, esto es, la matriz de Burt con submatrices modificadas en su diagonal.
- (4) Considerar estas últimas soluciones en (1) y repetir el procedimiento hasta que la función de discrepancia entre iteración e iteración sea casi 0.

Para ver por qué (3) se reduce a un análisis de correspondencia de la matriz modificada de Burt B^* , es necesario mostrar que los márgenes de la matriz de Burt B se preservan en B^* . Esto es cierto porque a cualquier nivel la nueva submatriz N_{qq}^* en la diagonal de B^* , que está dada por:

$$N_{qq}^* = nr_q r_q^\top + n D_q X_q \Lambda_\beta X_q^\top D_q \quad (3.9)$$

tiene los mismos márgenes de fila y columna que N_{qq} . Por lo tanto B^* tiene la misma solución trivial que B y la norma $\|\cdot\|_{qs}$ no se afecta por la modificación, dado que las soluciones no triviales en el análisis de B^* proveen el ajuste de mínimos cuadrados requerido para ajustar $B - C - nrr^\top$.

3.3.3. Soluciones iniciales basadas en el análisis de correspondencia múltiple

En la forma usual de ACM, las soluciones están anidadas y la matriz de Burt completa B se aproxima óptimamente como en (3.6) por

$$B - nrr^\top \approx n D_B^{\frac{1}{2}} \bar{U} \bar{\Lambda} \bar{U}^\top D_B^{\frac{1}{2}}.$$

En la práctica, la matriz solución $X = \bar{U}$ junto con un conjunto nuevo de factores de escala $\Lambda_\beta = \begin{bmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_K \end{bmatrix}$ proveen una solución inicial excelente para el algoritmo del ACC. Si las columnas de X son escritas como $x_{(1)}, \dots, x_{(K)}$ entonces (3.6) puede reescribirse como

$$B - nrr^\top \approx \sum_{k=1}^K n \beta_k D_B^{\frac{1}{2}} x_{(k)} x_{(k)}^\top D_B^{\frac{1}{2}}$$

Si todas las cantidades excepto los parámetros de escala $\beta_k (k = 1, \dots, K)$ son fijos, el problema se reduce a una simple regresión por mínimos cuadrados.

Como $B - nrr^\top$ es simétrico se necesita considerar solamente a las submatrices fuera de la diagonal en un triángulo de la matriz, pero es más simple dar resultados que asuman que todos los elementos en los triángulos inferior y superior estén ajustados. Se introduce la notación $bvec[\cdot]$ para nombrar al vector de elementos de las submatrices fuera de la diagonal encadenados en una secuencia fija. El número de elementos en ese vector es $\sum \sum_{q \neq s} J_q J_s$. Este problema de regresión puede ser expresado así: $y \approx U\beta$, donde $y \approx n^{-1} bvec[B - nrr^\top]$, U tiene como k -ésima columna:

$$bvec[D_B^{\frac{1}{2}} x_{(k)} x_{(k)}^\top D_B^{\frac{1}{2}}]$$

y β es el vector de orden $K \times 1$ de parámetros de regresión. Los pesos de la regresión son los

inversos de los elementos de $bvec[rr^T]$ que están contenidos en una matriz diagonal denotada por D_w . La solución de mínimos cuadrados esta dada entonces por $\hat{\beta} = (U^T D_w U)^{-1} U^T D_w y$.

3.3.4. Inercia total en el ACC

La inercia total ajustada que debe considerarse en el ACC obtenida a partir de la matriz de Burt ajustada no debe incluir a la inercia de las subtablas de su diagonal. Esta inercia total ajustada tambien es igual a:

$$\mathcal{I}(B) - \frac{J - Q}{Q^2}.$$



Capítulo 4

Aplicaciones

4.1. Introducción

En este capítulo se muestra una aplicación de las técnicas AC, ACM, ACMA y ACC, las cuales han sido hechas con datos del III Estudio Institucional sobre Egresados PUCP, recogidas en un censo aplicado en el 2008 a todos los egresados de los semestres 2006-1, 2006-2 y 2007-0 que ya tenían 1 o 2 años de egreso en ese momento. La base de datos objeto de este estudio está conformada por variables categóricas, numéricas y abiertas. Se trabajará con las variables categóricas y se recodificará las numéricas en rangos para convertirlas en categóricas y poder aplicar las técnicas indicadas.

4.2. Caso de estudio

El seguimiento a egresados universitarios se está convirtiendo en la mayoría de las universidades en un sistema consolidado, automatizado y de uso obligatorio para la toma de decisiones en el marco de un sistema universitario de calidad. Esto se produce dado que una de las formas de evaluar los resultados de los procesos formativos de la educación superior es por intermedio del éxito de los egresados en el mundo laboral. En dicho contexto, la Pontificia Universidad Católica del Perú, a través de la Dirección Académica de Planeamiento y Evaluación (DAPE), ha impulsado el desarrollo de estudios institucionales sobre sus egresados con el propósito de generar información útil y actualizada para el análisis y la dirección de las unidades en sus respectivos planes de mejora. Los dos primeros estudios sobre egresados realizados por la PUCP durante los años 2003 y 2004, tuvieron por objetivo generar información integral y estadísticamente representativa sobre nuestros egresados. Los focos de análisis comunes de dichos estudios fueron: la inserción laboral inicial, empleo actual y satisfacción con la formación de pregrado recibida en la PUCP. La información obtenida permitió acompañar procesos de autoevaluación, desarrollo curricular y planeamiento en algunas unidades. Para nuestra aplicación, se tiene la oportunidad de analizar los datos del estudio indicado en la introducción de este capítulo aplicando las técnicas motivo de esta tesis. En los resultados oficiales del estudio se plantearon conclusiones basadas en las observaciones directas a las tablas de contingencias entre las variables, principalmente las facultades de los egresados contra el resto de variables. No se utilizó el análisis de correspondencias. Esta técnica, sin embargo, podría ser usada para llegar a conclusiones más acertadas dada su facilidad de visualización de relaciones entre las variables y los datos disponibles son adecuados para su aplicación. Es importante indicar que actualmente se viene desarrollando IV Estudio Institucional sobre

Egresados PUCP en el que se podrán aplicar las técnicas expuestas en este trabajo, colaborar con las conclusiones del mismo de manera eficaz y más aún comparar resultados obtenidos en el 2008 con los del 2011.

El objetivo del III Estudio Institucional sobre Egresados PUCP es dar a conocer el proceso de inserción y la situación laboral en el año 2008, la valoración de la formación y los estudios post egreso de los egresados de pregrado PUCP 2006-1, 2006-2 Y 2007-0. Es importante aclarar que este estudio fue hecho por una empresa consultora. No fue realizado por la Dirección de Informática de la PUCP.

A partir del estudio, se enfocará la aplicación del análisis de correspondencia en el rubro de inserción laboral analizando las siguientes variables para cada facultad del egresado:

- Ingreso mensual bruto del primer empleo relacionado con su formación (Menor a 1 RMV (1), De 1 hasta 2 RMV (2), Más de 2 RMV hasta 4 RMV (3), Más de 4 RMV (4), No aplica (97), No precisa (99) - RMV: Remuneración mínima vital).
- Tiempo transcurrido entre egresar y obtener su primer empleo relacionado con su formación (Menos de 1 mes(1), De 1 a 3 meses (2), De 4 a 6 meses (3), De 7 a 12 meses (4), Más de 1 año (5), No tuvo empleo relacionado (97), No precisa (99)).
- En qué medida considera el egresado que su formación en la PUCP le permitió desarrollar cada una de las siguientes capacidades (escala del 1 al 5, donde 1 es Nada desarrollada y 5 es Totalmente desarrollada):
 - Autonomía para el aprendizaje.
 - Trabajo en equipo.
 - Liderazgo.
 - Creatividad.
 - Capacidad reflexiva y crítica.
 - Iniciativa y capacidad de hacer propuestas.
 - Capacidad de comunicación.
 - Dominio de idiomas.
 - Dominio de tecnologías.
 - Resolución de problemas.
 - Capacidad de análisis y síntesis.
- Según el egresado, qué tan importante fue para realizar su primer empleo, cada una de las siguientes capacidades (escala del 1 al 5, donde 1 es Nada importante y 5 es Totalmente importante):
 - Autonomía para el aprendizaje.
 - Trabajo en equipo.
 - Liderazgo.

- Creatividad.
- Capacidad reflexiva y crítica.
- Iniciativa y capacidad de hacer propuestas.
- Capacidad de comunicación.
- Dominio de idiomas.
- Dominio de tecnologías.
- Resolución de problemas.
- Capacidad de análisis y síntesis.

Se analizará la asociación de:

- Ingreso mensual bruto del primer empleo relacionado con su formación y la Facultad del egresado.
- Tiempo transcurrido entre egresar y obtener su primer empleo relacionado con su formación y la Facultad del egresado.
- Percepción del grado desarrollo de cada capacidad recibida en la formación PUCP y la Facultad del egresado.
- Percepción del grado de importancia que el egresado asigna a cada capacidad para la obtención de su primer empleo y la Facultad del egresado.

con el fin de escoger con qué variables plantear el alcance del ejercicio a desarrollar.

Al ser la Facultad una variable nominal, se utilizará el coeficiente V de Cramer y el coeficiente de contingencia para determinar la correlaciones entre el Ingreso mensual bruto del primer empleo relacionado con su formación y la Facultad del egresado; así como entre el Tiempo transcurrido entre egresar y obtener su primer empleo relacionado con su formación y la Facultad del egresado se muestra en el cuadro 4.1.

	V de Cramer	Coficiente de contingencia
Ingreso mensual bruto del primer empleo relacionado con su formación	0.175	0.331
Tiempo transcurrido entre egresar y obtener su primer empleo relacionado con su formación	0.069	0.137

Cuadro 4.1: Correlación entre el Ingreso del primer empleo y el Tiempo transcurrido en conseguir el primer empleo; y la Facultad del egresado

Se observa que el coeficiente de asociación es mejor si se usa la variable del Ingreso que la del tiempo. Se trabajará entonces con la primera variable indicada.

La correlación entre la percepción del grado desarrollo de cada capacidad recibida en la formación PUCP y la Facultad del egresado se muestra en el cuadro 4.2.

Se ve que la Percepción del grado de desarrollo del Dominio de tecnologías, de la Capacidad de comunicación y de la Capacidad reflexiva y crítica son los que tienen una correlación más alta con la Facultad del egresado.

En el cuadro 4.3 se ve que la Percepción del grado de importancia que el egresado asigna a las capacidades de Análisis y síntesis, a la capacidad Reflexiva y crítica y al Dominio de

Capacidad	V de Cramer	Coefficiente de contingencia
Dominio de tecnologías	0.185	0.383
Capacidad de comunicación	0.17	0.322
Capacidad reflexiva y crítica	0.169	0.354
Trabajo en equipo	0.141	0.3
Resolución de problemas	0.139	0.296
Dominio de idiomas	0.122	0.263
Autonomía para el aprendizaje	0.115	0.25
Liderazgo	0.115	0.249
Capacidad de análisis y síntesis	0.115	0.248
Creatividad	0.09	0.197
Iniciativa y capacidad de hacer propuestas	0.076	0.167

Cuadro 4.2: Correlación entre la Percepción del grado desarrollo de cada capacidad recibida en la formación PUCP y la Facultad del egresado

Capacidad	V de Cramer	Coefficiente de Contingencia
Capacidad de análisis y síntesis	0.16	0.338
Capacidad reflexiva y crítica	0.145	0.309
Dominio de tecnologías	0.121	0.26
Autonomía para el aprendizaje	0.111	0.241
Trabajo en equipo	0.11	0.24
Capacidad de comunicación	0.105	0.206
Dominio de idiomas	0.102	0.223
Resolución de problemas	0.097	0.212
Liderazgo	0.089	0.196
Iniciativa y capacidad de hacer propuestas	0.088	0.194
Creatividad	0.084	0.184

Cuadro 4.3: Percepción del grado de importancia que el egresado asigna a cada capacidad para la obtención de su primer empleo y la Facultad del egresado

tecnologías para la obtención de su primer empleo son los que tienen una correlación más alta con la Facultad del egresado.

Ahora, para cada capacidad y Facultad de egreso, se verá como se relaciona la percepción del grado de desarrollo recibida en la formación PUCP y la percepción del grado de importancia que el egresado asigna. En este caso se usa el coeficiente Tau de Kendall ya que ambas variables son ordinales. Se obtiene lo mostrado en el cuadro 4.4.

Se ve que las que tienen menor correlación a nivel total son las capacidades de Dominio de idiomas, Capacidad de comunicación y Resolución de problemas. Quiere decir que la percepción del grado de desarrollo recibida en la formación PUCP tiene la relación más débil con la percepción de la importancia que los egresados asignan a estas capacidades, lo que podría estar mostrando una situación que debería tomarse en cuenta.

Entonces para el ejercicio a desarrollar se elegirá la Capacidad de comunicación, ya que puede evidenciar un problema y además la percepción de su grado de desarrollo recibido en la formación PUCP tiene una de las más altas correlaciones con la Facultad del egresado.

En el apéndice se presenta la ficha técnica de esta encuesta.

	TOTAL	CeI	AyC	CCSS	Der	Edu	LlyCCHH	CCyAAC	Arte
Dominio de idiomas	0.201	0.174	0.2	0.137	0.172	0.332	0.303	0.203	0.266
Capacidad de comunicación	0.224	0.163	0.405	0.252	0.217	0.161	0.152	0.318	-0.06
Resolución de problemas	0.272	0.253	0.434	0.159	0.32	0.421	0.282	0.049	-0.068
Iniciativa y capacidad de hacer propuestas	0.275	0.232	0.412	0.325	0.278	0.462	0.335	0.178	0.195
Trabajo en equipo	0.292	0.3	0.312	0.138	0.243	0.368	0.273	0.38	0.329
Liderazgo	0.301	0.27	0.284	0.425	0.315	0.354	0.343	0.321	-0.047
Capacidad de análisis y síntesis	0.312	0.313	0.46	0.392	0.31	0.317	0.11	-0.016	0.663
Autonomía para el aprendizaje	0.318	0.32	0.093	0.191	0.407	0.499	0.219	0.275	0.378
Capacidad reflexiva y crítica	0.319	0.298	0.396	0.405	0.261	0.311	0.022	0.169	0.167
Creatividad	0.324	0.327	0.536	0.169	0.337	0.323	0.213	0.255	0.258
Dominio de tecnologías	0.352	0.297	0.426	0.197	0.362	0.529	0.333	0.249	0.125

Cuadro 4.4: Correlación entre la percepción del grado de desarrollo recibida en la formación PUCP y la percepción del grado de importancia que el egresado asigna para cada Capacidad y Facultad del egresado

4.3. Alcance y objetivos de la aplicación

La aplicación se limitará al análisis de correspondencias de las variables:

- Facultad del egresado
- Ingreso del primer empleo relacionado con su formación
- Grado de desarrollo de la Capacidad de comunicación recibida durante la formación

Las variables que conforman la base de datos analizada en esta tesis son presentadas en el cuadro 4.5.

Eje	Nombre de variable	Etiqueta de variable	Valores
General	Facultad	Facultad del egresado	Ciencias y Artes de las Comunicaciones (CCAACC), Ciencias e Ingeniería (CeI), Ciencias Sociales (CCSS), Administración y Contabilidad (AyC), Arte (Arte), Derecho (Der), Educación (Edu), Letras y Ciencias Humanas (LlyCCHH)
Inserción laboral	Ingreso1erEmp	Ingreso del primer empleo relacionado con su formación	Menor a 1 RMV (1), De 1 hasta 2 RMV (2), Más de 2 RMV hasta 4 RMV (3), Más de 4 RMV (4), No aplica (97), No precisa (99). RMV: Remuneración mínima vital
	FComunica	Percepción del grado de desarrollo de la Capacidad de comunicación recibida durante la formación	del FC1 al FC5, donde FC1 es Nada desarrollada y FC5 es Totalmente desarrollada.
	IComunica	Percepción del grado de importancia que el egresado asigna a la Capacidad de comunicación para la obtención del primer empleo	del IC1 al IC5, donde IC1 es Nada importante y IC5 es Totalmente importante

Cuadro 4.5: Información sobre variables

Las preguntas del cuestionario de la encuesta asociadas a cada variable se muestran en el cuadro 4.6.

Nombre de variable	Pregunta
Ingreso1erEmp	¿Cuál era su ingreso o ganancia total bruta mensual recibida en ese primer empleo?
FComunica	A continuación le daré una lista de capacidades, por favor indíqueme en qué medida considera Ud. que su formación en la PUCP LE PERMITIÓ DESARROLLARLAS, considerando una escala del 1 al 5 donde 1 es Nada desarrollada y 5 es Totalmente desarrollada.
IComunica	A continuación le daré una lista de capacidades, por favor indíqueme QUÉ TAN IMPORTANTE FUERON para realizar este primer empleo, considerando una escala del 1 al 5 donde 1 es Nada importante y 5 es Totalmente importante.

Cuadro 4.6: Preguntas del cuestionario de la encuesta asociadas a las variables

Los objetivos son:

- Desarrollar una aplicación del ACS con la intención de:
 - Aplicar lo explicado en la teoría sobre esto, dado que los siguientes métodos se basan en él.
 - Comparar los resultados con aquellos obtenidos por la función `ca()` de R y aplicar la función `plot()` para obtener los biplots y desarrollar la interpretación correspondiente.
 - Comprobar posteriormente que el ACS es un caso especial del ACC.
 - Comparar posteriormente la reconstrucción de las tablas de cada par de variables a partir de los resultados del ACM, ACMA y ACC con la reconstrucción a partir de los resultados del ACS.
- Desarrollar aplicaciones del ACM (Matriz de Burt), ACMA y ACC a la tabla de contingencia correspondiente a Facultad del egresado contra Ingreso del primer empleo relacionado con su formación, con el objetivo de comparar con los resultados obtenidos con el ACS y comprobar que el ACS es un caso particular del ACC y del ACMA cuando se trabaja con 2 variables.
- Presentar la función `mjca()` de R para análisis de correspondencia múltiple y aplicar la función `plot()` a dicha función para poder obtener los biplots correspondientes, interpretarlos y compararlos.
- Desarrollar aplicaciones del ACM (Matriz de Burt), ACMA y ACC para las tres variables presentadas: Facultad del egresado, Ingreso del primer empleo relacionado con su formación y Percepción del grado de desarrollo de la Capacidad de comunicación recibida durante la formación, enfatizando en la reconstrucción de la tabla de contingencia original y exponiendo las diferencias entre ellas y con respecto al ACS.
- Desarrollar los biplots obtenidos para el ACM, ACMA y ACC y compararlos.
- Desarrollar y comentar un biplot adicional con las variables Facultad del egresado, Percepción del grado de desarrollo de la Capacidad de comunicación recibida durante

la formación y Percepción del grado de importancia asignado por el egresado a la Capacidad de comunicación para la obtención del primer empleo.

4.4. Tablas de contingencia

Las tablas de contingencia de la aplicación son:

- Facultad del egresado contra Ingreso del primer empleo relacionado con su formación.
- Facultad del egresado contra Grado de desarrollo de la Capacidad de comunicación recibida durante la formación.
- Ingreso del primer empleo relacionado con su formación contra Grado de desarrollo de la Capacidad de comunicación recibida durante la formación.

Pueden revisarse en los cuadros 4.7, 4.8 y 4.9.

	Menor 1 RMV	1 a 2 RMV	Mayor 2 a 4 RMV	Mayor 4 RMV	I No precisa
Arte	2	4	6	1	3
AyC	5	38	49	16	2
CCAACC	12	49	23	5	2
CCSS	5	23	25	18	3
CeI	19	197	266	96	21
Der	32	111	76	22	18
Edu	13	11	9	1	1
LLyCCHH	12	44	31	5	12

Cuadro 4.7: Facultad del egresado contra Ingreso del primer empleo relacionado con su formación

	FC1	FC2	FC3	FC4	FC5
Arte	2	1	5	6	2
AyC	2	9	27	48	24
CCAAC		3	20	37	31
CCSS	3	5	25	29	12
CeI	20	88	241	181	69
Der	8	9	58	107	77
Edu		1	7	11	16
LLCCHH		5	25	44	30

Cuadro 4.8: Facultad del egresado contra Grado de desarrollo de la Capacidad de comunicación recibida durante la formación

	FC1	FC2	FC3	FC4	FC5
De 1 a 2 RMV	11	38	149	178	101
Más 4 RMV	4	24	53	54	29
Mayor 2 a 4 RMV	14	47	151	184	89
Menor 1 RMV	1	6	27	37	29
No precisa	5	6	28	10	13

Cuadro 4.9: Ingreso del primer empleo relacionado con su formación contra Grado de desarrollo de la Capacidad de comunicación recibida durante la formación

4.5. Aplicación del ACS

En primer lugar se desarrollará un ACS a la tabla de contingencia Facultad del egresado contra Ingreso del primer empleo relacionado con su formación aplicando la teoría correspondiente. Luego se comparará el resultado con lo hallado por la función CA disponible en R.

Se usará la ecuación 2.5:

$$\bar{Q} = D_r^{-1/2}(P - rc^T)D_c^{-1/2}$$

Además, se sabe que, aplicando el corolario A.5.1., se cumple que:

$$\bar{Q} = \bar{U}\bar{\Lambda}^{1/2}\bar{V}^T$$

Usando la función svd de R, se halla los correspondientes valores de esta descomposición.

Con estos valores se podrá hallar las coordenadas estándares y principales para crear el biplot correspondiente. Se recuerda que las coordenadas estandarizadas son: $X = D_r^{-1/2}\bar{U}$ para las filas y $Y = D_c^{-1/2}\bar{V}$ para las columnas; y, las coordenadas principales son: $X = D_r^{-1/2}\bar{U}\bar{\Lambda}^{1/2}$ para las filas y $Y = D_c^{-1/2}\bar{V}\bar{\Lambda}^{1/2}$ para las columnas.

Se obtiene lo siguiente:

Coordenadas estandarizadas:

Filas:

$$\begin{bmatrix} 0.3801494 & -0.2232067 & 1.1904700 & -0.29775176 & -1 \\ -1.2548974 & 0.7248281 & -0.8096913 & -2.02433082 & -1 \\ -0.6282770 & 0.0970930 & -0.3559082 & 1.06058491 & -1 \\ 2.6520418 & 1.7321832 & -1.3586868 & 0.01363984 & -1 \\ 1.0319695 & -3.7533977 & -2.0416233 & -0.67306177 & -1 \end{bmatrix}$$

Las filas representan a los perfiles fila de la tabla de contingencia (Ingreso) y las columnas a las dimensiones.

Columnas:

$$\begin{bmatrix} 1.0299138 & -3.46268732 & -4.0426072 & 1.0253776 & -4.3882941 \\ -0.6454243 & 0.61596589 & 0.3319949 & 0.9012316 & 1.2228741 \\ 1.1758129 & 0.67631848 & 2.6012669 & -0.2339567 & -0.6828336 \\ -0.6001666 & 0.78519758 & -1.1209336 & -3.5898859 & 0.6082817 \\ -0.7902524 & 0.06927793 & -0.1009586 & 0.3611774 & 0.5487362 \\ 0.9132976 & -0.39360871 & 0.2507044 & -0.4744161 & 0.9014428 \\ 3.1523021 & 3.84052399 & -2.7941672 & 1.5201975 & 0.8629630 \\ 1.1386339 & -1.98051948 & -0.3102529 & 0.2377087 & 1.7777627 \end{bmatrix}$$

Las filas representan a los perfiles columna de la tabla de contingencia (Facultades) y las columnas a las dimensiones.

Coordenadas principales:

Filas:

$$\begin{bmatrix} 0.1129477 & -0.02973753 & 0.12930473 & -0.0209849687 & -3.12456e - 09 \\ -0.3728477 & 0.09656788 & -0.08794586 & -0.1426709244 & -3.12456e - 09 \\ -0.1866699 & 0.01293557 & -0.03865752 & 0.0747479751 & -3.12456e - 09 \\ 0.7879589 & 0.23077645 & -0.14757586 & 0.0009613096 & -3.12456e - 09 \\ 0.3066127 & -0.50006016 & -0.22175407 & -0.0474360928 & -3.12456e - 09 \end{bmatrix}$$

Las filas representan a los perfiles fila de la tabla de contingencia (Ingreso) y las columnas a las dimensiones.

Columnas:

$$\begin{bmatrix} 0.3060019 & -0.461329203 & -0.43909401 & 0.07226663 & -1.371149e - 08 \\ -0.1917646 & 0.082064312 & 0.03606014 & 0.06351706 & 3.820943e - 09 \\ 0.3493506 & 0.090105007 & 0.28254061 & -0.01648882 & -2.133554e - 09 \\ -0.1783179 & 0.104610824 & -0.12175193 & -0.25300822 & 1.900612e - 09 \\ -0.2347951 & 0.009229806 & -0.01096578 & 0.02545508 & 1.714559e - 09 \\ 0.2713536 & -0.052439962 & 0.02723065 & -0.03343593 & 2.816612e - 09 \\ 0.9365933 & 0.511667877 & -0.30349278 & 0.10714058 & 2.696379e - 09 \\ 0.3383042 & -0.263861963 & -0.03369859 & 0.01675325 & 5.554726e - 09 \end{bmatrix}$$

Las filas representan a los perfiles columna de la tabla de contingencia (Facultades) y las columnas a las dimensiones.

También se puede hallar la inercia total, usando los valores singulares en la descomposición:

$$\begin{bmatrix} 0.08827677 & 0.00000000 & 0.00000000 & 0.00000000 & 0.000000e + 00 \\ 0.00000000 & 0.01774988 & 0.00000000 & 0.00000000 & 0.000000e + 00 \\ 0.00000000 & 0.00000000 & 0.01179755 & 0.00000000 & 0.000000e + 00 \\ 0.00000000 & 0.00000000 & 0.00000000 & 0.004967158 & 0.000000e + 00 \\ 0.00000000 & 0.00000000 & 0.00000000 & 0.00000000 & 9.762874e - 18 \end{bmatrix}$$

Entonces:

$$\mathcal{I} = \sum_{k=1}^K \lambda_k = 0.1227914$$

Se observa que con las dos primeras dimensiones se obtiene el 86 % de toda la información.

Las contribuciones totales (2.3.4) son:

Filas

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Ingreso1erEmpDe 1 a 2 RMV	0.053519354	0.018450871	0.52485521	0.032833047	0.37034166
Ingreso1erEmpMas 4 RMV	0.200513902	0.066895662	0.083477024	0.521784244	0.127329208
Ingreso1erEmpMayor 2 a 4 RMV	0.148637376	0.003549781	0.047698204	0.423561781	0.376552841
Ingreso1erEmpMenor 1 RMV	0.546065619	0.232954816	0.143325332	1.44445E-05	0.077639761
Ingreso1erEmpNo precisa	0.051263667	0.678148673	0.200644472	0.021806485	0.048136652

Columnas

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
FacultadArte	0.013176678	0.1489466	0.203014622	0.013060859	0.239219037
FacultadAyC	0.035576831	0.032403362	0.009413255	0.069366475	0.127714514
FacultadCCAAC	0.097679197	0.032316771	0.478074368	0.003867201	0.032942391
FacultadCCSS	0.020694708	0.035422049	0.072189783	0.740418301	0.021258132
FacultadCeI	0.290430742	0.002232033	0.004740217	0.060666916	0.140035499
FacultadDer	0.167729187	0.031153957	0.012638865	0.045258769	0.163403091
FacultadEdu	0.270027404	0.400804939	0.212156862	0.062798923	0.020236545
FacultadLLCCHH	0.104685329	0.316720105	0.007772292	0.00456255	0.255190836

Las contribuciones relativas (2.3.4) son:

Filas

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Ingreso1erEmpDe 1 a 2 RMV	0.414172927	0.028710233	0.542819884	0.014296956	3.1696E-16
Ingreso1erEmpMas 4 RMV	0.787934174	0.052855769	0.043838713	0.115371344	5.53356E-17
Ingreso1erEmpMayor 2 a 4 RMV	0.827792995	0.003975066	0.035501045	0.132730895	2.31927E-16
Ingreso1erEmpMenor 1 RMV	0.892174819	0.076528959	0.031294893	1.32791E-06	1.40288E-17
Ingreso1erEmpNo precisa	0.237704589	0.632268868	0.124337029	0.005689513	2.46851E-17

Columnas

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
FacultadArte	0.185608375	0.421862786	0.382176824	0.010352016	3.72665E-16
FacultadAyC	0.752895791	0.137881709	0.026622744	0.082599757	2.98909E-16
FacultadCCAAC	0.580435922	0.038612609	0.379658434	0.001293036	2.1649E-17
FacultadCCSS	0.261539366	0.090012004	0.121926721	0.526521909	2.97122E-17
FacultadCeI	0.984755874	0.001521723	0.002147977	0.011574426	5.25116E-17
FacultadDer	0.941087807	0.035146631	0.00947709	0.014288472	1.01394E-16
FacultadEdu	0.705945939	0.21069084	0.074125234	0.009237987	5.85102E-18
FacultadLLCCHH	0.61701583	0.375348864	0.006122164	0.001513142	1.66344E-16

En las contribuciones totales se destaca que las modalidades 'más de 4 RMVs' y 'menos de 1 RMV' son las que dan el principal aporte a la dimensión 1; mientras que en el caso de las facultades, Ciencias e Ingeniería y Educación hacen lo propio. En la dimensión 2, destaca nuevamente la modalidad de ingreso 'menos de 1 RMV' y aparece la modalidad 'No precisa'. La facultad de Educación aparece nuevamente en la dimensión 2 acompañada de la facultad de Letras y Ciencias Humanas.

En cuanto a las contribuciones relativas, se puede indicar que sobretodo se debe tomar en cuenta las posiciones de las modalidades 'menos de 1 RMV', 'mayor 2 a 4' y 'más de 4 RMVs' en la dimensión 1 y de las facultades de Ciencias e Ingeniería, y Derecho. En la dimensión 2 en cambio destacar las posiciones de las modalidades 'No precisa' y de las facultades de Letras y Ciencias Humanas, y Arte.

Haciendo uso de las pruebas de Malinvaud y Orloci, se confirma con cuántas dimensiones se debe trabajar.

	p-valor Malinvaud	p-valor Orloci
Dim 1	1.27709E-05	9.54993E-20
Dim 2	0.000625478	0.003546015
Dim 3		0.018787121
Dim 4		0.171364637

Según la prueba de Malinvaud, se puede trabajar con 2 dimensiones y según la prueba de Orloci hasta 3 dimensiones. Se decide trabajar con 2 dimensiones, además por una facilidad de visualización.

En R se cuenta con la función CA del paquete `ca`. Ofrece funciones para los gráficos correspondientes de los resultados en 2 y 3 dimensiones.

El paquete tiene los siguientes componentes:

- ACS:
 - Cálculo: `ca()`
 - Impresión y resumen: los métodos `print()` y `summary()` para objetos `ca`
 - Plotting: `plot.ca()` y `plot3d.ca()`

La idea en AC es reducir la dimensionalidad de una matriz de datos y visualizarla en un subespacio de menor dimensionalidad, comunmente 2 o 3. Para esto, el paquete `ca`, que se encuentra implementado en R (Nenadic, 2007) aplica lo expuesto en los capítulos anteriores.

Para aplicar el método se puede disponer de una tabla de contingencia, por ejemplo en formato Excel, se guarda su contenido en el Clipboard y se ejecuta el comando en R:

```
R > table <- read.table("clipboard")
```

Luego:

```
R > ca(table)
```

Se aplicará lo indicado a la misma tabla de contingencia Facultad del egresado contra Ingreso del primer empleo relacionado con su formación usada anteriormente.

Principal inertias (eigenvalues):

	1	2	3	4
Value	0.088277	0.01775	0.011798	0.004967
Percentage	71.89 %	14.46 %	9.61 %	4.05 %

Rows:

	Ingreso1erEmpDe 1 a 2 RMV	Ingreso1erEmpMas 4 RMV	Ingreso1erEmpMayor 2 a 4 RMV	Ingreso1erEmpMenor 1 RMV	Ingreso1erEmpNo precisa
Mass	0.370342	0.127329	0.376553	0.07764	0.048137
ChiDist	0.175504	0.420036	0.20517	0.834216	0.628885
Inertia	0.011407	0.022465	0.015851	0.054031	0.019038
Dim.1	0.380149	-1.254897	-0.628277	2.652042	1.03197
Dim.2	-0.223207	0.724828	0.097093	1.732183	-3.753398

Columns:

	FacultadArte	FacultadAyC	FacultadCCAAC	FacultadCCSS	FacultadCeI	FacultadDer	FacultadEdu	FacultadLLCCHH
Mass	0.012422	0.085404	0.070652	0.057453	0.465062	0.201087	0.027174	0.080745
ChiDist	0.710273	0.221004	0.458547	0.34868	0.236605	0.279718	1.114719	0.430684
Inertia	0.006267	0.004171	0.014856	0.006985	0.026035	0.015733	0.033766	0.014977
Dim.1	1.029914	-0.645424	1.175813	-0.600167	-0.790252	0.913298	3.152302	1.138634
Dim.2	-3.462687	0.615966	0.676319	0.785198	0.069278	-0.393609	3.840524	-1.980519

La salida contiene los autovalores y porcentajes de la inercia explicada para todas las dimensiones posibles. Adicionalmente, se muestran los valores de las filas y columnas (masas, distancias ji-cuadrado a los puntos de sus promedios, inercias y coordenadas estándar. Sin embargo, estos valores están restringidos a 2 dimensiones para las coordenadas estándares.

Los autovalores mostrados coinciden con los hallados en λ_k . Las masas coinciden con los hallados en r y c^T . Las inercias pueden calcularse a partir de las contribuciones totales y la inercia total. Las coordenadas estándar correspondientes a las dos primeras dimensiones son las mismas halladas anteriormente.

El comando R:

```
R > summary(ca(table))
```

retorna el resumen del CA:

Principal inertias (eigenvalues):

dim	value	%	cum %	scree plot
1	0.088277	71.9	71.9	*****
2	0.01775	14.5	86.3	****
3	0.011798	9.6	96	**
4	0.004967	4	100	
Total:		0.122791	100	

El porcentaje de inercia representado por las dos primeras dimensiones es del 86.3%.

Rows:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	I1ED	370	443	93	113	414	54	-30	29	18
2	I1EM 4	127	841	183	-373	788	201	97	53	67
3	I1EM 2	377	832	129	-187	828	149	13	4	4
4	I1EM 1	78	969	440	788	892	546	231	77	233
5	I1EN	48	870	155	307	238	51	-500	632	678

Columns:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	FcIA	12	607	51	306	186	13	-461	422	149
2	FcAC	85	891	34	-192	753	36	82	138	32
3	FCCA	71	619	121	349	580	98	90	39	32
4	FCCS	57	352	57	-178	262	21	105	90	35
5	FcCI	465	986	212	-235	985	290	9	2	2
6	FcID	201	976	128	271	941	168	-52	35	31
7	FcIE	27	917	275	937	706	270	512	211	401
8	FLLC	81	992	122	338	617	105	-264	375	317

Nuevamente, los autovalores y porcentajes relativos de la inercia explicada son dados para todas las dimensiones. Adicionalmente, se muestran los porcentajes acumulados y un plot correspondiente. Los items dados en Filas y Columnas incluyen las coordenadas principales para las dos primeras dimensiones ($k = 1$ y $k = 2$). Se muestran el cuadrado de las correlaciones (cor) y las contribuciones (ctr) de los puntos. Notar que las cantidades en estas tablas están multiplicadas por 1000 y en consecuencia las cantidades cor y ctr se expresan en permiles. La calidad total (qlt) es dada con respecto a la dimensionalidad de la solución, en este caso es la suma de las correlaciones cuadradas para las dos dimensiones incluidas. Se puede comprobar además que los valores de las coordenadas estándares de las dos primeras dimensiones coinciden con lo hallado anteriormente. También se comprueba lo mismo con las masas y las contribuciones totales.

Para hallar el biplot del ACS se usa el comando R:

$$R > \text{plot}(ca(table))$$

y se obtiene lo mostrado en la figura 4.1.

En la figura 4.1, se observa claramente que los egresados de Ciencias e Ingeniería, Ciencias Sociales y Administración y Contabilidad son los que relativamente perciben ingresos mayores a 2 RMV en su primer empleo. En la otra parte están los egresados de Educación que son quienes relativamente perciben un ingreso menor a 1 RMV en su primer empleo. Por su parte los egresados de Derecho y Ciencias y Artes de la Comunicación perciben relativamente un ingreso de 1 hasta 2 RMV en su primer empleo. Los egresados de Arte no precisaron su respuesta por lo que no se podría dar una conclusión muy certera respecto a ellos. Algo similar sucedió con los egresados de Letras y Ciencias Humanas.

Ahora se presentarán los biplots de las otras 2 tablas de contingencia presentadas usando la función CA de R.

La figura 4.2 muestra evidencia de que los egresados de Ciencias e Ingeniería, así como los de Arte y Ciencias Sociales son quienes enfatizan que la capacidad de comunicación no está desarrollada en alto grado durante su formación. En contraposición se ve que

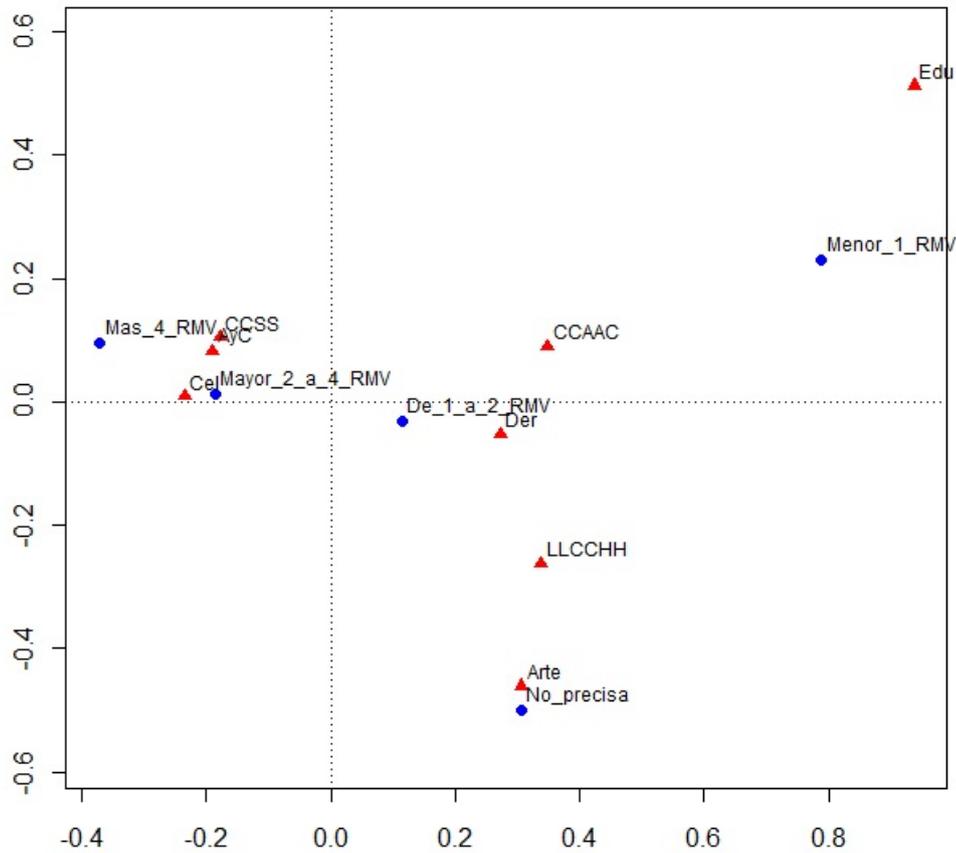


Figura 4.1: Facultad del egresado contra Ingreso del primer empleo relacionado con su formación

los egresados de Ciencias y Artes de la Comunicación y de Educación manifiestan lo contrario. En el medio están los egresados de Derecho, Letras y Ciencias Humanas, y Administración y Contabilidad.

La característica más saltante de figura 4.3 consiste en mostrar que los niveles de ingreso altos en un primer empleo no tienen relación directa con una buena formación en la capacidad de comunicación.

4.6. Aplicación del ACM, ACMA y ACC

4.6.1. Aplicación del ACM, ACMA y ACC - 2 variables

ACM - Facultad del egresado contra Ingreso del primer empleo relacionado con su formación

En este caso la Matriz de Burt que se usará será la siguiente:

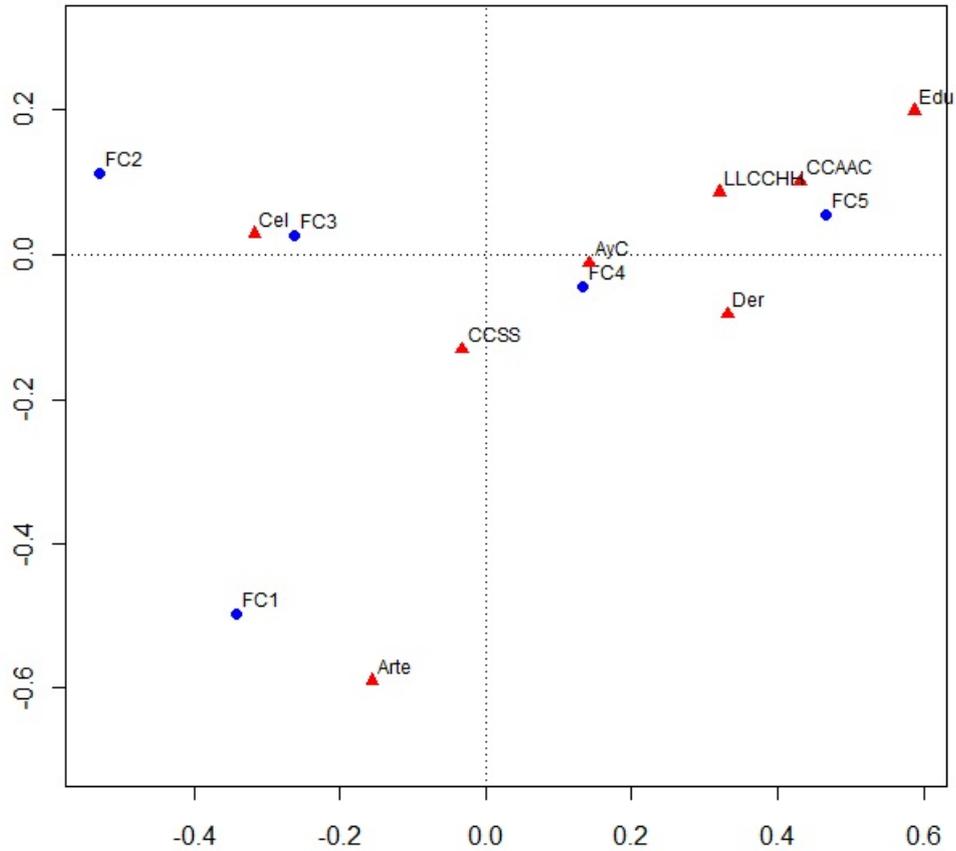


Figura 4.2: Facultad del egresado contra Grado de desarrollo de la Capacidad de comunicación recibida durante la formación

	Arte	AyC	CCAAC	CCSS	CeI	Der	Edu	LLCCHH	De 1 a 2 RMV	Mas 4 RMV	Mayor 2 a 4 RMV	Menor 1 RMV	No precisa
Arte	16	0	0	0	0	0	0	0	4	1	6	2	3
AyC	0	110	0	0	0	0	0	0	38	16	49	5	2
CCAAC	0	0	91	0	0	0	0	0	49	5	23	12	2
CCSS	0	0	0	74	0	0	0	0	23	18	25	5	3
CeI	0	0	0	0	599	0	0	0	197	96	266	19	21
Der	0	0	0	0	0	259	0	0	111	22	76	32	18
Edu	0	0	0	0	0	0	35	0	11	1	9	13	1
LLCCHH	0	0	0	0	0	0	0	104	44	5	31	12	12
De 1 a 2 RMV	4	38	49	23	197	111	11	44	477	0	0	0	0
Mas 4 RMV	1	16	5	18	96	22	1	5	0	164	0	0	0
Mayor 2 a 4 RMV	6	49	23	25	266	76	9	31	0	0	485	0	0
Menor 1 RMV	2	5	12	5	19	32	13	12	0	0	0	100	0
No precisa	3	2	2	3	21	18	1	12	0	0	0	0	62

Y el ACM es un ACS aplicado a esta matriz. Se usará la función `ca` del R para obtener el ACS de esta matriz. Los resultados son los siguientes:

Inercias principales (autovalores) y sus raíces:

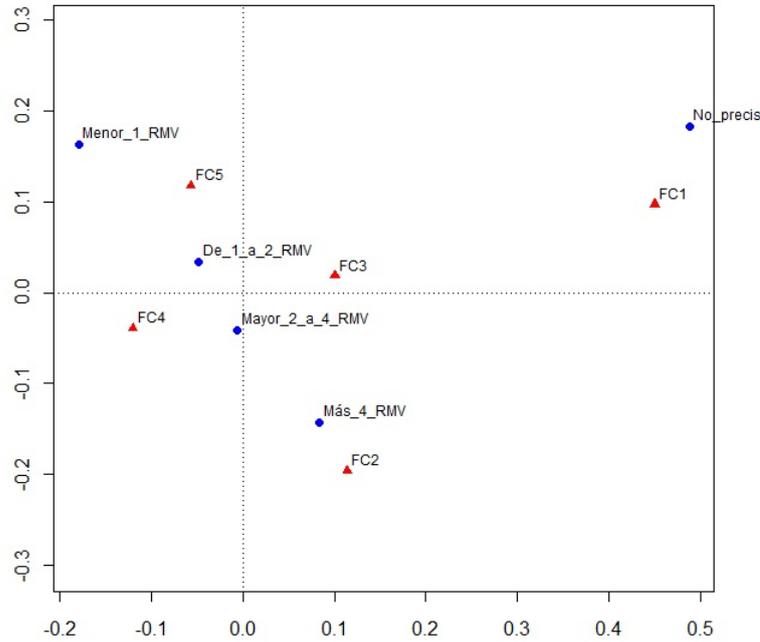


Figura 4.3: ACS - Ingreso del primer empleo relacionado con su formación contra Grado de desarrollo de la Capacidad de comunicación recibida durante la formación

	1	2	3	4	5	6	7	8	9	10	11	12
Autovalores	0.420626	0.321052	0.307258	0.286481	0.25	0.25	0.25	0.216003	0.198641	0.187823	0.123512	0
Porcentaje	14.96 %	11.42 %	10.93 %	10.19 %	8.89 %	8.89 %	8.89 %	7.68 %	7.07 %	6.68 %	4.39 %	0 %
Raíces	0.648556859	0.566614507	0.554308578	0.535239199	0.5	0.5	0.5	0.464761229	0.445691597	0.43338551	0.351442741	0

De acuerdo a Benzécri (1979) y Greenacre (1988), solamente 4 valores singulares son mayores que $1/Q = 0.5$. Entonces se debería trabajar a lo más con las correspondientes 4 dimensiones.

Si se aplica la estimación del promedio de la distribución de autovalores bajo independencia de Ben Ammou y Saporta (1988, 2003), se encuentra que la desviación estándar es $\sigma = 0.031434731$, de modo que el intervalo de confianza al 95 % es $0.437130539 < \lambda < 0.562869461$). En consecuencia, solamente los dos primeros valores singulares están fuera del intervalo de confianza y deberían ser considerados significativos.

De acuerdo a estos resultados, se trabajará con las dos primeras dimensiones.

Continuando con el ejercicio, se tiene:

Filas:

	Arte	AyC	CCAAC	CCSS	CeI	Der	Edu	LLCCHH	Del a 2 RMV	Mas 4 RMV	Mayor 2 a 4 RMV	Menor 1 RMV	No precisa
Mass	0.006211	0.042702	0.035326	0.028727	0.232531	0.100543	0.013587	0.040373	0.185171	0.063665	0.188276	0.03882	0.024068
ChiDist	6.324733	2.31926	2.584967	2.874629	0.776606	1.423238	4.303638	2.405214	0.930326	1.874845	0.921348	2.50758	3.175665
Inertia	0.248461	0.229692	0.236051	0.237383	0.140243	0.203662	0.251648	0.233558	0.160266	0.223784	0.159825	0.244098	0.242725
Dim.1	-1.029914	0.645424	-1.175813	0.600167	0.790252	-0.913298	-3.152302	-1.138634	-0.380149	1.254897	0.628277	-2.652042	-1.03197
Dim.2	3.462687	-0.615966	-0.676319	-0.785198	-0.069278	0.393609	-3.840524	1.980519	0.223207	-0.724828	-0.097093	-1.732183	3.753398

Se halla las coordenadas principales:

name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
FclA	6	107	88	-668	11	7	1962	96	74
FcAC	43	55	82	419	33	18	-349	23	16
FCCA	35	109	84	-763	87	49	-383	22	16
FCCS	29	42	84	389	18	10	-445	24	18
FcCI	233	438	50	513	436	145	-39	3	1
FclD	101	198	72	-592	173	84	223	25	16
FclE	14	481	90	-2044	226	135	-2176	256	200
FLLC	40	312	83	-738	94	52	1122	218	158
I1ED	185	89	57	-247	70	27	126	18	9
I1EM 4	64	236	80	814	188	100	-411	48	33
I1EM 2	188	199	57	407	196	74	-55	4	2
I1EM 1	39	624	87	-1720	470	273	-981	153	116
I1EN	24	493	86	-669	44	26	2127	448	339

El resultado de las Columnas es similar ya que la matriz de Burt es una matriz simétrica. Se observa que los valores de las coordenadas estándares son similares a las obtenidas en el ACS pero los principales varían con respecto al ACS y además el porcentaje de inercia que se consigue con las dos primeras dimensiones alcanza solamente el 26 %.

ACM, ACMA y ACC son ejecutados por la función en R: `mjca()`. La estructura de la función es mantenida similar como su contraparte para el ACS. Las dos diferencias más saltantes están en el formato de ingreso de datos y la restricción del análisis solo a las columnas. La función `mjca()` toma como entrada una matriz de respuestas. En términos de R, esto es una plantilla de datos con las columnas conteniendo a los factores. Dentro de esta función, la matriz es convertida a una matriz de indicadores o a una matriz de Burt, dependiendo del tipo de análisis. La restricción a columnas significa que por defecto, solo los valores de las columnas son dados como salida. Esto es suficiente ya que la matriz de Burt es simétrica y el resultado para las filas sería similar al de las columnas. La opción `lambda` en `mjca()` especifica el tipo de análisis de correspondencia que se ejecutará:

- `lambda = "indicator"`: Análisis basado en un ACS de la matriz de indicadores.
- `lambda = "Burt"`: Análisis basado en la descomposición de autovalores de la matriz de Burt.
- `lambda = "adjusted"`: Análisis basado en la matriz de Burt con el ajuste de inercias (ACMA).
- `lambda = "JCA"`: Análisis de correspondencia conjunto.

Se aplicará la función `mjca()` a los mismos datos y se comparará sus resultados con aquellos obtenidos en la aplicación ACS de la matriz de Burt correspondiente:

```
R > multiple <- read.table("clipboard", header = TRUE)
```

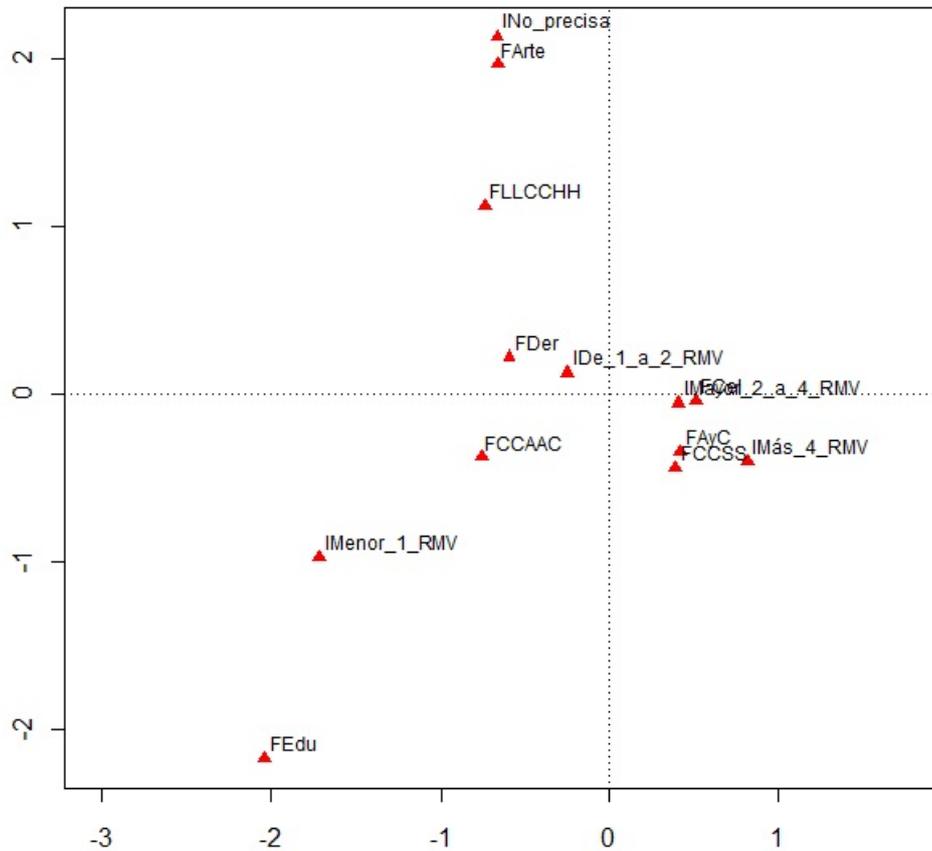


Figura 4.4: ACM - Facultad del egresado contra Ingreso del primer empleo relacionado con su formación

$$R > mjca(multiple[, 1 : 2], lambda = "Burt")$$

Se puede comprobar que los resultados son exactamente los mismos.

Finalmente se mostrará el plot correspondiente a dos dimensiones con coordenadas principales en el cuadro 4.4.

Las conclusiones son similares, a pesar que los puntos se muestran rotados.

El bajo porcentaje de inercia ocurre debido a la inclusión de las tablas diagonales, cuya contribución al ji-cuadrado es muy alta dado que una tabla que se cruza consigo misma obtiene un ji-cuadrado máximo. Por eso se reitera la necesidad de una re-evaluación para conseguir una estimación mejor.

Sin embargo con las pruebas de Benzecri, Greenacre, Ben Ammou y Soporta, sí se ha logrado establecer la capacidad explicativa de los autovalores.

ACMA - Facultad del egresado contra Ingreso del primer empleo relacionado con su formación

El ACMA es similar al ACM, pero partes de la inercia son calculadas a partir de las inercias principales de la matriz de Burt B :

$$\lambda_k^{aj} = \left(\frac{Q}{Q-1}\right)^2 \left(\sqrt{\lambda_k} - \frac{1}{Q}\right)^2, k = 1, 2, \dots \quad (4.1)$$

Recordar que λ_k es la k -ésima inercia principal de la matriz de Burt B . Los ajustes se hacen solamente en las dimensiones para las cuales $\sqrt{\lambda_k} > \frac{1}{Q}$, y por lo tanto no se usan más dimensiones; por tanto, los porcentajes de inercia no suman el 100 %.

Entonces se aplica esta transformación a lo hallado en el ejercicio correspondiente al ACM y se obtiene los siguientes valores ajustados de las inercias principales:

Lambdas ajustados:

	1	2	3	4
Value	0.088276562	0.01774997	0.011797687	0.004967205

Consiguiéndose una representación del 86.3 % de la inercia con las dos primeras dimensiones.

Recordar que en este caso, la inercia total ajustada se calcula de la siguiente forma:

$$\frac{Q}{Q-1} \left(inercia(B) - \frac{J-Q}{Q^2} \right)$$

Reemplazando:

$$\frac{2}{2-1} \left(2.811396 - \frac{13-2}{2^2} \right) = 0.122792$$

Entonces, el porcentaje de inercia correspondiente a las dos primeras dimensiones es igual a $\frac{0.088276562+0.01774997}{0.122792} 100\% = 86.3\%$.

Y se obtienen los siguientes valores de coordenadas estándares y principales.

Coordenadas estándares

	Arte	AyC	CCAAC	CCSS	CeI	Der	Edu	LLCCHH
Mass	0.006211	0.042702	0.035326	0.028727	0.232531	0.100543	0.013587	0.040373
ChiDist	0.636769	0.214885	0.412571	0.286634	0.235792	0.278051	1.091239	0.429861
Inertia	0.002518	0.001972	0.006013	0.00236	0.012928	0.007773	0.016179	0.00746
Dim. 1	-1.029914	0.645424	-1.175813	0.600167	0.790252	-0.913298	-3.152302	-1.138634
Dim. 2	-3.462687	0.615966	0.676319	0.785198	0.069278	-0.393609	3.840524	-1.980519
	De 1 a 2 RMV	Mas 4 RMV	Mayor 2 a 4 RMV	Menor 1 RMV	No precisa			
Mass	0.185171	0.063665	0.188276	0.03882	0.024068			
ChiDist	0.149069	0.402971	0.196351	0.827663	0.608098			
Inertia	0.004115	0.010338	0.007259	0.026593	0.0089			
Dim. 1	-0.380149	1.254897	0.628277	-2.652042	-1.03197			
Dim. 2	-0.223207	0.724828	0.097093	1.732183	-3.753398			

Coordenadas principales

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	FacultadArte	6	304	26	-306	93	2	461	211	10
2	FacultadAyC	43	445	17	192	376	5	-82	69	2
3	FacultadCCAAC	35	310	60	-349	290	15	-90	19	2
4	FacultadCCSS	29	176	28	178	131	3	-105	45	2
5	FacultadCeI	233	493	106	235	492	43	-9	1	0
6	FacultadDer	101	488	64	-271	471	25	52	18	2
7	FacultadEdu	14	458	137	-937	353	40	-512	105	27
8	FacultadLLCCHH	40	496	61	-338	309	16	264	188	21
9	Ingreso1erEmpDe 1 a 2 RMV	185	221	46	-113	207	8	30	14	1
10	Ingreso1erEmpM? 4 RMV	64	420	91	373	394	30	-97	26	4
11	Ingreso1erEmpMayor 2 a 4 RMV	188	416	65	187	414	22	-13	2	0
12	Ingreso1erEmpMenor 1 RMV	39	484	220	-788	446	81	-231	38	16
13	Ingreso1erEmpNo precisa	24	435	78	-307	119	8	500	316	45

Se observa que los valores de las coordenadas estándares y principales son exactamente iguales a las obtenidas en el ACS y además el porcentaje de inercia que se consigue con las dos primeras dimensiones también coincide: 86.3 % pero en este caso el resultado sí es anidado. También se puede comprobar que el caso del ACS entonces es prácticamente un caso particular del ACMA para dos dimensiones.

El plot correspondiente se puede ver en la figura 4.5.

ACC - Facultad del egresado contra Ingreso del primer empleo relacionado con su formación

En el caso del ACC, la función auxiliar `iterate.mjca()` es internamente usada para actualizar la matriz de Burt por iteración de ponderaciones de mínimos cuadrados. La función tiene dos criterios de convergencia, llamadas *epsilon* y *maxit*. La opción *epsilon* establece un criterio de convergencia en términos de diferencia absoluta máxima de la matriz de Burt en un paso de iteración comparado con la matriz de Burt del paso previo. El máximo número de iteraciones está dado por la opción *maxit*. De esta manera, el programa itera hasta que alguna de las condiciones es satisfecha.

Lo que se hará en este caso es ayudarse con el valor *Burt.upd* del resultado de:

$$R > mjca(multiple[, 1 : 2], lambda = "JCA")$$

La matriz de Burt ajustada, obtenida a partir de las iteraciones de esta técnica aplicada a la misma matriz de Burt anterior es la siguiente:

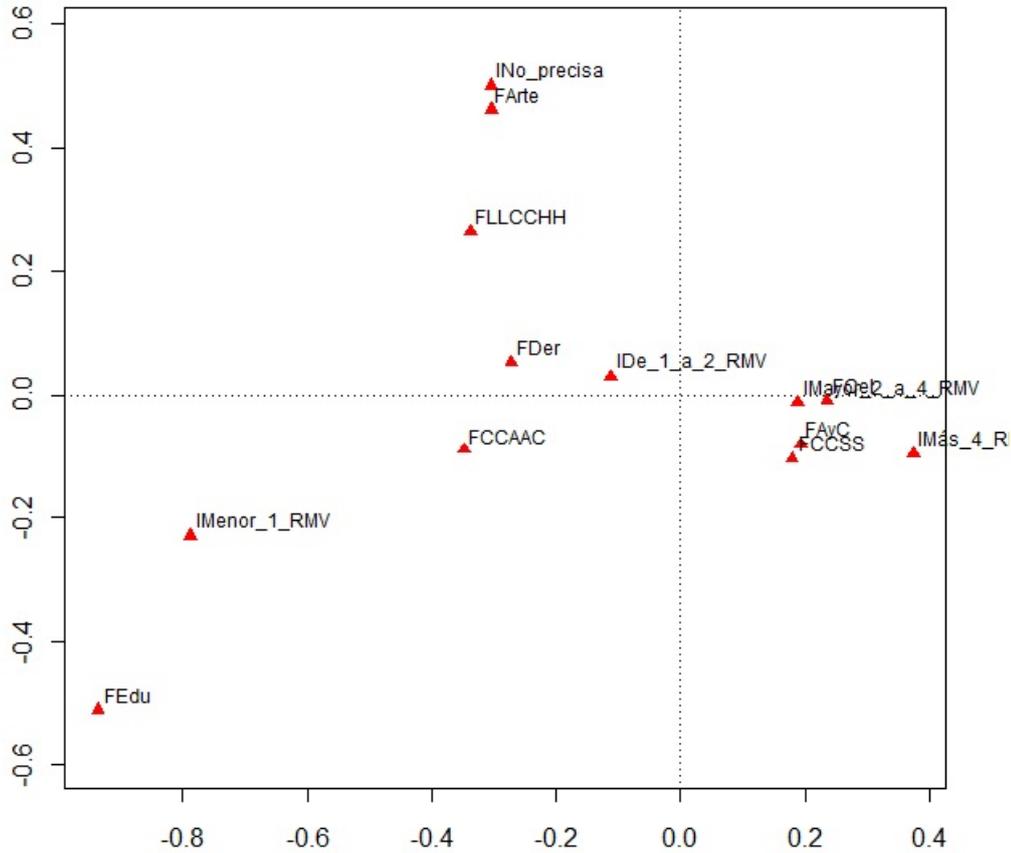


Figura 4.5: ACMA - Facultad del egresado contra Ingreso del primer empleo relacionado con su formación

	Arte	AyC	CCAAC	CCSS	CeI	Der	Edu	LLCCHH	De 1 a 2 RMV
Mas 4 RMV	Mayor 2 a 4 RMV	Menor 1 RMV	No precisa						
Arte	0.57890014	0.7082861	1.184464	0.4174466	5.403811	4.70078	0.08385482	2.922456	4
1	6	2	3						
AyC	0.70828608	11.0320258	6.450718	7.4544651	59.200093	17.5311	2.12428595	5.499024	38
16	49	5	2						
CCAAC	1.18446448	6.4507183	9.462135	4.5019607	30.901167	23.48838	6.05176431	8.959406	49
5	23	12	2						
CCSS	0.41744657	7.4544651	4.501961	5.055778	39.513572	11.84434	1.6884236	3.524015	23
18	25	5	3						
CeI	5.40381139	59.2000925	30.901167	39.5135725	330.438604	94.18427	4.80670086	34.551785	197
96	266	19	21						
Der	4.70078013	17.5311023	23.488385	11.8443384	94.184267	66.06369	11.64085798	29.546578	111
22	76	32	18						
Edu	0.08385482	2.124286	6.051764	1.6884236	4.806701	11.64086	5.62804799	2.976064	11
1	9	13	1						
LLCCHH	2.92245637	5.4990239	8.959406	3.5240151	34.551785	29.54658	2.97606447	16.020671	44
5	31	12	12						
De 1 a 2 RMV	4	38	49	23	197	111	11	44	185.41044
50.818307	166.35113	46.219779	28.200348						
Mas 4 RMV	1	16	5	18	96	22	1	5	50.81831
32.11399	76.79978	2.272402	1.995517						
Mayor 2 a 4 RMV	6	49	23	25	266	76	9	31	166.35113
76.799783	204.27615	19.857547	17.715393						
Menor 1 RMV	2	5	12	5	19	32	13	12	46.21978
2.272402	19.85755	27.091953	4.558319						
No precisa	3	2	2	3	21	18	1	12	28.20035
1.995517	17.71539	4.558319	9.530422						

Se halla ahora las inercias principales, las coordenadas estándares y las coordenadas principales de esta nueva matriz de Burt Ajustada, aplicando un ACS a dicha matriz.

Principal inertias (eigenvalues):

	1	2	3	4	5	6	7	8	9	10	11	12
Value	0.088277	0.01775	0.002949	0.002949	0.001242	0.001242	0	0	0	0	0	0

Total: 0.114409

Diagonal inertia discounted from eigenvalues: 0.0530132 Percentage explained by JCA in 2 dimensions: 86.3(Eigenvalues are not nested) [Iterations in JCA: 19 , epsilon = 6.96e-05]

El valor 0.0530132 representa la inercia de las tablas de la diagonal de la matriz de Burt ajustada y debe ser restada de la inercia total. La nueva inercia total ajustada será igual a 0.0613958. Este valor también debe ser retirado de la suma de las inercias principales de las dos primeras dimensiones que también lo incluyen. Entonces la suma ajustada de las inercias principales de las dos primeras dimensiones es igual a $(0.088277 + 0.01775) - 0.0530132 = 0.0446312$. Entonces el porcentaje de inercia explicado por las dos dimensiones es igual a $(0.0446312/0.0613958) * 100\% = 86.3\%$. La inercia total ajustada también puede obtenerse restando el valor $\frac{J-Q}{Q^2}$ del valor de la inercia de la matriz de Burt original. Se sabe que la inercia de la matriz de Burt original es el promedio de las inercias de cada subtabla producto del cruce de cada variable con la otra. El valor que se resta representa a la suma de las inercias de las tablas de la diagonal (aquellas en que una variable se cruza consigo misma) dividida entre la cantidad total de subtablas. La inercia de la matriz de Burt original obtenida en el ejercicio del ACM es igual a 2.811396. $J = 13$ es la suma de todas las categorías y $Q = 2$ es la cantidad de variables. Entonces se comprueba que efectivamente la inercia total ajustada es igual a $2.811396 - \frac{13-2}{2^2} = 0.0613958$.

Rows:

	Arte	AyC	CCAAC	CCSS	CeI	Der	Edu	LLCCHH
Mass	0.006211	0.042702	0.035326	0.028727	0.232531	0.100543	0.013587	0.040373
ChiDist	0.636769	0.214885	0.412571	0.286634	0.235792	0.278051	1.091239	0.429861
Inertia	0.002518	0.001972	0.006013	0.00236	0.012928	0.007773	0.016179	0.00746
Dim. 1	-1.029914	0.645424	-1.175813	0.600167	0.790252	-0.913298	-3.152302	-1.138634
Dim. 2	-3.462687	0.615966	0.676319	0.785198	0.069278	-0.393609	3.840524	-1.980519

	De 1 a 2 RMV	Mas 4 RMV	Mayor 2 a 4 RMV	Menor 1 RMV	No precisa
Mass	0.185171	0.063665	0.188276	0.03882	0.024068
ChiDist	0.149069	0.402971	0.196351	0.827663	0.608098
Inertia	0.004115	0.010338	0.007259	0.026593	0.0089
Dim. 1	-0.380149	1.254897	0.628277	-2.652042	-1.03197
Dim. 2	-0.223207	0.724828	0.097093	1.732183	-3.753398

Rows:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	FclA	6	756	22	-306	231	7	-461	525	74
2	FcAC	43	942	17	192	796	18	82	146	16
3	FCCA	35	765	53	-349	717	49	90	48	16
4	FCCS	29	520	21	178	387	10	105	133	18
5	FcCI	233	993	113	235	992	145	9	2	1
6	FclD	101	988	68	-271	952	84	-52	36	16
7	FclE	14	957	141	-937	737	135	512	220	200
8	FLLC	40	996	65	-338	619	52	-264	377	158
9	IIED	185	614	36	-113	574	27	-30	40	9
10	I1EM 4	64	914	90	373	856	100	97	57	33
11	I1EM 2	188	908	63	187	904	74	13	4	2
12	I1EM 1	39	984	232	-788	906	273	231	78	116
13	I1EN	24	930	78	-307	254	26	-500	676	339

Se observa que los valores de las coordenadas estándares y principales son exactamente iguales a las obtenidas en el ACS y además el porcentaje de inercia que se consigue con las dos primeras dimensiones también coincide: 86.3% aunque los resultados no estén anidados (el resultado de dos dimensiones no contiene exactamente la mejor solución unidimensional como eje principal). Sin embargo, se puede comprobar que el caso del ACS entonces es un caso particular del ACC para dos dimensiones.

El plot correspondiente se puede ver en el cuadro 4.6.

4.6.2. Aplicación del ACM, ACMA y ACC - 3 variables

Como se indicó anteriormente, en esta subsección se desarrollará la aplicación del ACM (Matriz de Burt), ACMA y ACC para las tres variables presentadas: Facultad del egresado, Ingreso del primer empleo relacionado con su formación y Grado de desarrollo de la Capacidad de comunicación recibida durante la formación, determinando la cantidad de dimensiones con las que se debe trabajar y enfatizando en la reconstrucción de la tabla de contingencia original y exponiendo las diferencias entre ellas y con respecto al ACS.

Para reconstruir una tabla de contingencia original se usa la siguiente expresión:

$$\tilde{N} = nrc^T + nD_r^{1/2}\bar{U}\bar{\Lambda}^{1/2}\bar{V}^T D_c^{1/2},$$

El ejercicio que se hará considerará 3 casos: valores esperados, valores considerando solamente hasta la 1era dimensión y valores considerando hasta la segunda dimensión para cada técnica: ACM, ACMA y ACC.

Para evaluar la calidad de la reconstrucción, se hallará el valor absoluto de la suma de las diferencias celda por celda entre los valores reales y los valores reconstruidos.

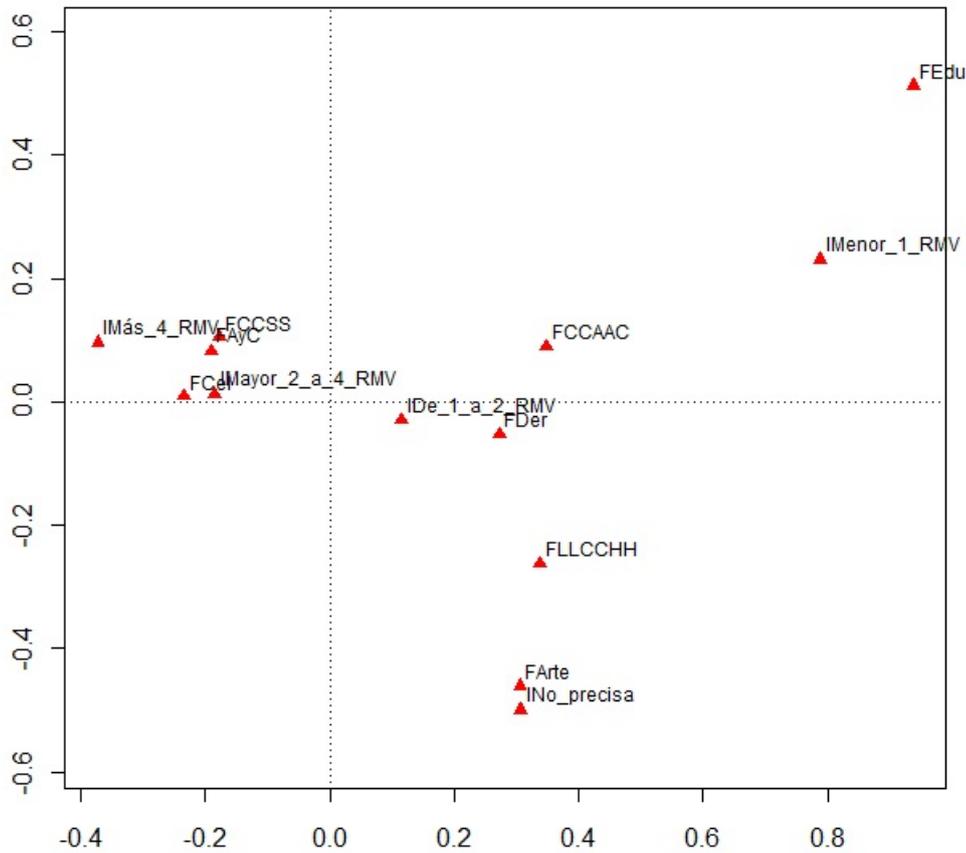


Figura 4.6: ACC - Facultad del egresado contra Ingreso del primer empleo relacionado con su formación

Mientras más pequeña sea la suma, mejor será la calidad de la reconstrucción (Camiz, 2010). Recordar que en este caso los valores originales son aquellos correspondientes a la matriz de Burt.

Finalmente, se hará el mismo ejercicio de evaluación de la calidad de la reconstrucción, utilizando solamente las submatrices de la matriz de Burt correspondientes a cada par de variables y las compararemos con la reconstrucción aplicada al ACS.

Para las tres técnicas entonces, se halla la siguiente matriz de Burt:

	Arte	AyC	CCAAC	CCSS	CeI	Der	Edu	LLCCHH	FC1	FC2
FC3	FC4	FC5	De 1 a 2 RMV	Mas 4 RMV	Mayor 2 a 4 RMV	Menor 1 RMV	No precisa			
Arte	16	0	0	0	0	0	0	0	2	1
5	6	2	4	1	6	2	3			
AyC	0	110	0	0	0	0	0	0	2	9
27	48	24	38	16	49	5	2			
CCAAC	0	0	91	0	0	0	0	0	0	3
20	37	31	49	5	23	12	2			
CCSS	0	0	0	74	0	0	0	0	3	5
25	29	12	23	18	25	5	3			
CeI	0	0	0	0	599	0	0	0	20	88
241	181	69	197	96	266	19	21			
Der	0	0	0	0	0	259	0	0	8	9
58	107	77	111	22	76	32	18			
Edu	0	0	0	0	0	0	35	0	0	1
7	11	16	11	1	9	13	1			
LLCCHH	0	0	0	0	0	0	0	104	0	5
25	44	30	44	5	31	12	12			
FC1	2	2	0	3	20	8	0	0	35	0
0	0	0	11	4	14	1	5			
FC2	1	9	3	5	88	9	1	5	0	121
0	0	0	38	24	47	6	6			
FC3	5	27	20	25	241	58	7	25	0	0
408	0	0	149	53	151	27	28			
FC4	6	48	37	29	181	107	11	44	0	0
0	463	0	178	54	184	37	10			
FC5	2	24	31	12	69	77	16	30	0	0
0	0	261	101	29	89	29	13			
De 1 a 2 RMV	4	38	49	23	197	111	11	44	11	38
149	178	101	477	0	0	0	0			
Mas 4 RMV	1	16	5	18	96	22	1	5	4	24
53	54	29	0	164	0	0	0			
Mayor 2 a 4 RMV	6	49	23	25	266	76	9	31	14	47
151	184	89	0	0	485	0	0			
Menor 1 RMV	2	5	12	5	19	32	13	12	1	6
27	37	29	0	0	0	100	0			
No precisa	3	2	2	3	21	18	1	12	5	6
28	10	13	0	0	0	0	62			

Recordar que las coordenadas estandarizadas se hallan:

$$X = D_r^{-1/2} \bar{U}$$

$$Y = D_c^{-1/2} \bar{V}$$

y por lo tanto:

$$\bar{U} = D_r^{1/2} X$$

$$\bar{V} = D_c^{1/2} Y$$

Pero, como la matriz de Burt es simétrica, se cumple que:

$$c = r$$

$$X = Y$$

$$D_c = D_r$$

$$\bar{V} = \bar{U}$$

Entonces la expresión de la reconstrucción se reescribe como:

$$\tilde{N} = nrr^T + nD_r^{1/2}\bar{U}\bar{\Lambda}^{1/2}\bar{U}^T D_r^{1/2},$$

Los valores de r y de D_r se obtendrán a partir de la matriz de Burt directamente. Los valores de X , en el caso de ACM y ACMA, se obtendrán a partir de los resultados del ACS aplicado a la matriz de Burt. En el caso de ACC, se obtendrán a partir del ACS aplicado a la matriz de Burt ajustada. Los valores de $\bar{\Lambda}$ se obtendrán dependiendo de la técnica. Si es el ACM, a partir de los resultados del ACS aplicado a la matriz de Burt. Si es ACC, a partir de los resultados del ACS aplicado a la matriz de Burt ajustada. Si es ACMA, a partir de los valores actualizados según el ajuste correspondiente.

ACM - Reconstrucción

Los resultados del ACS aplicado a la matriz de Burt, son los siguientes:

Inercias principales (autovalores) y raíces de éstas:

	Autovalores	Porcentaje	Raices
1	0.239403	13.88 %	0.489288259
2	0.158463	9.18 %	0.398074114
3	0.143555	8.32 %	0.378886527
4	0.132221	7.66 %	0.363622057
5	0.127776	7.41 %	0.35745769
6	0.119429	6.92 %	0.345585011
7	0.115584	6.70 %	0.33997647
8	0.11065	6.41 %	0.332640948
9	0.106256	6.16 %	0.325969324
10	0.098061	5.68 %	0.31314693
11	0.092772	5.38 %	0.304584964
12	0.084448	4.89 %	0.290599381
13	0.081823	4.74 %	0.286047199
14	0.074066	4.29 %	0.272150694
15	0.040793	2.36 %	0.20197277
16	0	0 %	0
17	0	0 %	0

De acuerdo a Benzécri (1979) y Greenacre (1988), solamente 8 valores singulares son mayores que $1/Q = 0.33$. Entonces se debería trabajar a lo más con las correspondientes 8 dimensiones.

Si se aplica la estimación del promedio de la distribución de autovalores bajo independencia de Ben Ammou y Saporta (1988, 2003), se encuentra que la desviación estándar es $\sigma = 0.028777723$, de modo que el intervalo de confianza al 95 % es $(0.275777887 < \lambda < 0.39088878)$. En consecuencia, solamente los dos primeros valores

singulares están fuera del intervalo de confianza y deberían ser considerados significativos.

De acuerdo a estos resultados, se trabajará con las dos primeras dimensiones.

Continuando con el ejercicio, se tiene:

Filas:

	Arte	AyC	CCAAC	CCSS	Cel	Der	Edu	LLCCHH	FC1	FC2	FC3	FC4	FC5	De 1 a 2 RMV	Mas 4 RMV	Mayor 2 a 4 RMV	Menor 1 RMV
No precisa																	
Mass 0.016046	0.004141	0.028468	0.023551	0.019151	0.155021	0.067029	0.009058	0.026915	0.009058	0.031315	0.10559	0.119824	0.067547	0.123447	0.042443	0.125518	0.02588
ChiDist 2.610357	5.176796	1.89703	2.126088	2.348998	0.660178	1.178764	3.534762	1.973634	3.48423	1.825791	0.863773	0.779746	1.180501	0.760438	1.534661	0.752987	2.052958
Inertia 0.109334	0.11097	0.102448	0.106455	0.105672	0.067563	0.093136	0.113175	0.104841	0.109963	0.104388	0.078781	0.072854	0.094132	0.071385	0.099961	0.071167	0.109074
Dim.1 -0.172699	-0.011517	0.055793	-1.605111	0.425466	1.060713	-1.14939	-3.017268	-1.186948	1.073769	1.652137	0.673059	-0.298699	-1.432187	-0.420166	1.173716	0.524949	-2.359633
Dim.2	6.064245	-1.544076	-1.357406	-0.129334	-0.046971	0.507367	-0.324192	1.096055	4.20683	-0.425944	0.543313	-0.707286	0.038705	-0.24962	-0.553111	-0.237606	0.155878 4.990806

Recordar que Dim.1 y Dim.2 representan los valores de las coordenadas estándares X en dos dimensiones.

Ahora, se aplica la fórmula de la reconstrucción para tres casos: el primero solamente con valores esperados (el primer término de la suma); el segundo agregando la primera dimensión (considerando solamente la primera columna de U y el primer valor de Λ); el tercero agregando la segunda dimensión (incluyendo la segunda columna de U y el segundo valor de Λ). Se obtiene 1 tabla reconstruida para cada caso. Luego se halla los residuos y la suma de los residuos absolutos comparando cada tabla reconstruida con la tabla original de la matriz de Burt. Se obtiene lo siguiente:

1 - Residuos - Valores esperados

	Arte	AyC	CCAAC	CCSS	Cel	Der	Edu	LLCCHH	FC1	FC2	FC3	FC4	FC5	De 1 a 2 RMV	Mas 4 RMV	Mayor 2 a 4 RMV	Menor 1 RMV
No precisa																	
Arte 2	16	1	1	1	7	3	0	1	2	1	0	0	1	2	1	0	1
AyC 3	1	101	8	6	51	22	3	9	1	1	8	8	2	3	2	8	4
CCAAC 2	1	8	85	5	42	18	2	7	2	6	9	4	13	15	7	11	5
CCSS 1	1	6	5	70	34	15	2	6	1	2	2	2	3	4	9	3	1
Cel 8	7	51	42	34	320	120	16	48	4	32	51	34	52	25	20	40	28
Der 6	3	22	18	15	120	207	7	21	1	15	24	14	25	15	11	22	12
Edu 1	0	3	2	2	16	7	34	3	1	2	4	2	9	2	3	4	10
LLCCHH 7	1	9	7	6	48	21	3	96	3	5	8	7	9	5	8	8	4
FC1 3	2	1	2	1	4	1	1	3	34	3	11	13	7	2	0	1	2
FC2 0	1	1	6	2	32	15	2	5	3	110	38	43	25	7	9	1	3
FC3 8	0	8	9	2	51	24	4	8	11	38	279	147	83	2	1	3	5
FC4 12	0	8	4	2	34	14	2	7	13	43	147	297	94	7	5	10	1
FC5 0	1	2	13	3	52	25	9	9	7	25	83	94	208	4	4	9	9
De 1 a 2 RMV 23	2	3	15	4	25	15	2	5	2	7	2	7	4	300	61	180	37
Mas 4 RMV 8	1	2	7	9	20	11	3	8	0	9	1	5	4	61	143	62	13
Mayor 2 a 4 RMV 23	0	8	11	3	40	22	4	8	1	1	3	10	9	180	62	302	38
Menor 1 RMV 5	1	4	5	1	28	12	10	4	2	3	5	1	9	37	13	38	92
No precisa	2	3	2	1	8	6	1	7	3	0	8	12	0	23	8	23	5

Suma de residuos absolutos 0 total: 7102

2 - Residuos - Valores 1ra. Dimensión

	Arte	AyC	CCAAC	CCSS	CeI	Der	Edu	LLCCHH	FC1	FC2	FC3	FC4	FC5	De 1 a 2 RMV	Mas 4 RMV	Mayor 2 a 4 RMV	Menor 1 RMV
No precisa																	
Arte 2	16	1	1	1	7	3	0	1	2	0	0	0	1	2	1	0	1
AyC 3	1	101	7	6	53	21	3	9	1	2	8	9	3	2	2	7	3
CCAAC 3	1	7	76	3	7	35	8	14	0	6	6	3	8	4	4	3	8
CCSS 0	1	6	3	69	42	11	1	4	1	4	2	4	1	2	6	6	2
CeI 5	7	53	7	42	167	49	9	19	5	17	15	1	38	24	27	21	29
Der 4	3	21	35	11	49	173	19	35	5	7	7	2	18	8	11	7	15
Edu 1	0	3	8	1	9	19	30	8	1	6	7	7	6	10	4	6	1
LLCCHH 6	1	9	14	4	19	35	8	90	1	5	5	0	9	4	1	4	7
FC1 3	2	1	0	1	5	5	1	1	34	6	15	11	2	1	3	3	2
FC2 1	0	2	6	4	17	7	6	5	6	94	59	33	4	8	6	18	15
FC3 9	0	8	6	2	15	7	7	5	15	59	250	132	44	19	19	29	20
FC4 13	0	9	3	4	1	2	7	0	11	33	132	289	113	4	5	23	11
FC5 1	1	3	8	1	38	18	6	9	2	4	44	113	155	24	23	27	25
De 1 a 2 RMV 24	2	2	4	2	24	8	10	4	1	8	19	4	24	285	46	160	55
Mas 4 RMV 7	1	2	4	6	27	11	4	1	3	6	19	5	23	46	129	80	5
Mayor 2 a 4 RMV 22	0	7	3	6	21	7	6	4	3	18	29	23	27	160	80	278	15
Menor 1 RMV 6	1	3	8	2	29	15	1	7	2	15	20	11	25	55	5	15	71
No precisa 59	2	3	3	0	5	4	1	6	3	1	9	13	1	24	7	22	6

Suma de residuos absolutos 1 total: 6389

3 - Residuos - Valores 2da. Dimensión

	AyC	CCAAC	CCSS	CeI	Der	Edu	LLCCHH	FC1	FC2	FC3	FC4	FC5	De 1 a 2 RMV	Mas 4 RMV	Mayor 2 a 4 RMV	Menor 1 RMV	No precisa Arte
Arte 13	4	3	1	7	7	0	5	3	1	7	10	2	2	2	3	0	7
AyC 13	4	92	14	7	54	15	3	3	7	5	3	8	3	9	3	1	2
CCAAC 9	3	14	72	4	8	30	9	10	5	4	15	16	8	0	1	2	8
CCSS 0	1	7	4	69	42	11	1	4	1	4	1	3	2	2	6	6	2
CeI 3	7	54	8	42	167	47	9	18	4	17	13	4	38	23	28	22	30
Der 8	7	15	30	11	47	168	19	40	1	9	2	12	18	3	14	12	15
Edu 0	0	3	9	1	9	19	30	7	1	6	8	8	6	10	4	6	1
LLCCHH 4	5	3	10	4	18	40	7	86	6	6	3	12	9	0	4	8	8
FC1 11	3	7	5	1	4	1	1	6	27	4	25	4	2	6	1	2	1
FC2 6	1	5	4	4	17	9	6	6	4	94	56	38	4	7	7	20	15
FC3 12	7	3	15	1	13	2	8	3	25	56	235	110	44	27	13	21	19
FC4 18	10	8	16	3	4	12	8	12	4	38	110	256	112	16	4	11	10
FC5 2	2	3	8	2	38	18	6	9	2	4	44	112	155	24	23	27	25
De 1 a 2 RMV 12	2	9	0	2	23	3	10	0	6	7	27	16	24	281	49	164	54
Mas 4 RMV 2	2	3	1	6	28	14	4	4	1	7	13	4	23	49	126	84	5
Mayor 2 a 4 RMV 11	3	1	2	6	22	12	6	8	2	20	21	11	27	164	84	274	14
Menor 1 RMV 7	0	2	8	2	30	15	1	8	1	15	19	10	25	54	5	14	71
No precisa 29	7	13	9	0	3	8	0	4	11	6	12	18	2	12	2	11	7

Suma de residuos absolutos 2 total: 6403

Se puede repetir el ejercicio, pero esta vez se considerará solamente los valores fuera de las tablas diagonales. Se obtiene lo siguiente:

Suma de residuos absolutos 0 sin tablas diagonales: 1598

Suma de residuos absolutos 1 sin tablas diagonales: 1585

Suma de residuos absolutos 2 sin tablas diagonales: 1795

Se ve como la suma de residuos absolutos, tanto los totales como aquellos obtenidos sin considerar las tablas diagonales, disminuye al considerar los valores de la 1era. dimensión pero sin embargo, aumenta un poco al incluir los valores de la 2da. dimensión. A simple vista se puede concluir que el ACM no reconstruye muy bien la tabla original. Nuevamente, esto es de esperarse dada la inclusión de las tablas diagonales en el cálculo de la inercia.

ACMA - Reconstrucción

El ACMA trabaja con la matriz de Burt original. Entonces los valores de r , X y \bar{U} son los mismos. Lo que hace es ajustar las inercias principales según lo ya visto.

Si se cuenta con las inercias principales (autovalores) y raíces de éstas resultado del cálculo aplicado a la matriz de Burt obtenidas en el ACM recordamos que las 8 primeras cumplen con la condición ($\sqrt{\lambda} > 1/Q$). Como se determinó que se debe considerar las dos primeras dimensiones, entonces se hallará los valores ajustados de las inercias principales para las dos primeras. Se obtiene:

$$\Lambda = \begin{bmatrix} 0.054724362 & 0.009430579 \end{bmatrix}$$

Ahora, de manera similar al ejercicio con el ACM se aplica la fórmula de la reconstrucción para los mismos tres casos. Se muestra directamente los resultados de los residuos absolutos para cada caso:

	Totales	Sin tablas diagonales
Valores esperados	7102	1598
1ra. dimensión	6762	1344
2da. dimensión	6748	1336

ACC - Reconstrucción

En este caso se trabajará con la matriz de Burt ajustada que minimiza el efecto de las tablas de la diagonal en la inercia total. Es la siguiente:

	Arte	AyC	CCAAC	CCSS	Cel	Der	Edu	LLCCHH	FC1	FC2	FC3	FC4	FC5	De 1 a 2 RMV	M? 4 RMV	Mayor 2 a 4 RMV	Menor 1 RMV
No precisa																	
Arte	0.5230503	0.7187354	0.6747485	0.8289163	7.49407	3.535799	0.3773166	1.847364	2	1	5	6	2	4	1	6	2
3																	
AyC	0.7187354	10.6904626	8.5374405	6.5304862	51.69033	21.185244	3.0016319	7.64567	2	9	27	48	24	38	16	49	5
2																	
CCAAC	0.6747485	8.5374405	16.0515023	3.4791774	2.484859	36.401382	8.9126651	14.458226	0	3	20	37	31	49	5	23	12
2																	
CCSS	0.8289163	6.5304862	3.4791774	4.6685611	42.7046814	10.917246	0.6986635	4.172268	3	5	25	29	12	23	18	25	5
3																	
Cel	7.49407	51.6903299	2.4848585	42.7046814	454.59766	38.385709	-11.8829418	13.525632	20	88	241	181	69	197	96	266	19
21																	
Der	3.535799	21.1852436	36.4013821	10.9172462	38.385709	90.704605	20.1148332	37.755182	8	9	58	107	77	111	22	76	32
18																	
Edu	0.3773166	3.0016319	8.9126651	0.6986635	-11.882942	20.114833	5.463469	8.314362	0	1	7	11	16	11	1	9	13
1																	
LLCCHH	1.847364	7.6456699	14.4582255	4.1722679	13.525632	37.755182	8.3143624	16.281296	0	5	25	44	30	44	5	31	12
12																	
FC1	2	2	0	3	20	8	0	0	1.905856	4.331769	14.49334	9.405534	4.863497	11	4	14	1
5																	
FC2	1	9	3	5	88	9	1	5	4.331769	15.543831	45.21216	39.493759	16.418479	38	24	47	6
6																	
FC3	5	27	20	25	241	58	7	25	14.493344	45.212162	144.68323	134.781374	68.82989	149	53	151	27
28																	
FC4	6	48	37	29	181	107	11	44	9.405534	39.493759	134.78137	177.093009	102.226325	178	54	184	37
10																	
FC5	2	24	31	12	69	77	16	30	4.863497	16.418479	68.82989	102.226325	68.661809	101	29	89	29
13																	
De 1 a 2 RMV	4	38	49	23	197	111	11	44	11	38	149	178	101	181.45796	58.5090761	176.65967	40.95617
19.4171217																	
M? 4 RMV	1	16	5	18	96	22	1	5	4	24	53	54	29	58.50908	26.7582936	70.50572	7.25617
0.9707437																	
Mayor 2 a 4 RMV	6	49	23	25	266	76	9	31	14	47	151	184	89	176.65967	70.5057161	195.69306	29.66724
12.4743068																	
Menor 1 RMV	2	5	12	5	19	32	13	12	1	6	27	37	29	40.95617	7.2561704	29.66724	13.72909
8.3913304																	
No precisa	3	2	2	3	21	18	1	12	5	6	28	10	13	19.41712	0.9707437	12.47431	8.39133
20.7464974																	

Se obtiene los siguientes valores:

Inercias principales (autovalores):

dim	value
1	0.106936
2	0.019131
3	0.002410
4	0.001164
5	0.000638
6	0.000478
7	0.000104

Total: 0.135940

Inercia de la diagonal descontada de los autovalores: 0.0773076 Porcentaje explicado por el ACC en 2 dimensiones: 83.2(Los autovalores no están anidados) [Iteraciones en ACC: 50, epsilon = 0.610006]

Rows:

	Arte	AyC	CCAAC	CCSS	Cel	Der	Edu	LLCCHH	FC1	FC2	FC3	FC4	FC5	De 1 a 2 RMV	M? 4 RMV	Mayor 2 a 4 RMV	Menor 1 RMV
No precisa																	
Mass	0.004141	0.028468	0.023551	0.019151	0.155021	0.067029	0.009058	0.026915	0.009058	0.031315	0.10559	0.119824	0.067547	0.123447	0.042443	0.125518	0.02588
0.016046																	
ChiDist	0.571358	0.177218	0.693584	0.268835	0.456603	0.497199	1.315272	0.554321	0.493036	0.3682	0.182218	0.128921	0.31017	0.111282	0.306283	0.145846	0.541906
0.931669																	
Inertia	0.001352	0.000894	0.011329	0.001384	0.03232	0.01657	0.01567	0.00827	0.002202	0.004245	0.003506	0.001992	0.006498	0.001529	0.003982	0.00267	0.0076
0.013928																	
Dim.1	-0.101968	0.05726	-2.043126	0.537353	1.390467	-1.506968	-3.799205	-1.616533	0.704636	1.048824	0.470719	-0.214959	-0.935239	-0.245198	0.771896	0.375843	-1.501146
-0.674201																	
Dim.2	-3.425201	0.99175	0.993599	0.168533	-0.113398	-0.101498	0.553408	-0.791676	-2.460116	-0.217536	-0.579938	0.592919	0.285514	0.233274	0.789307	0.427096	-0.452861
-6.493119																	

Ahora, de manera similar al ejercicio con el ACM se aplica la fórmula de la reconstrucción para los mismos tres casos. Se muestra directamente los resultados de los residuos absolutos para cada caso:

	Totales	Sin tablas diagonales
Valores esperados	7102	1598
1ra. dimensión	5607	633
2da. dimensión	5404	520

Son mucho mejores resultados que los obtenidos en el ACM. El ACC reconstruye mejor. Véase un resumen de las tres técnicas en cuanto a los resultados de los residuos absolutos totales para cada uno de los tres casos:

	ACM	ACC	ACMA
Valores esperados	7102	7102	7102
1ra. dimensión	6389	5607	6762
2da. dimensión	6403	5404	6748

y un resumen de las tres técnicas en cuanto a los resultados de los residuos absolutos sin considerar las tablas de la diagonal para cada uno de los tres casos:

	ACM	ACC	ACMA
Valores esperados	1598	1598	1598
1ra. dimensión	1585	633	1344
2da. dimensión	1795	520	1336

Se puede observar que en lo que se refiere a reconstrucción de la matriz de Burt, los resultados del ACC son mejores que los resultados obtenidos por el ACM y el ACMA. Se puede comprobar que el ACMA reduce lentamente la cantidad de residuos de la tabla completa y de las subtablas fuera de la diagonal. El ACM, no solo no reduce lentamente la cantidad de residuos al pasar a la primera dimensión sino que aumenta esta cantidad al pasar a la segunda dimensión. Es decir, con la segunda dimensión la cantidad de residuos es aún mayor que la desviación total obtenida cuando se consideran los valores esperados por independencia. Otra cosa interesante observada además es que, para la tabla total, la reducción de residuos es más lenta en el ACMA que en el ACM.

Por el contrario, el ACC muestra una reducción mucho más rápida de la cantidad de residuos. Con la segunda dimensión incluida, la cantidad de residuos en el ACC correspondientes a las tablas fuera de la diagonal es menos de la mitad de su par en el ACMA y menos de la tercera parte de su par en el ACM.

Ahora se evaluará de manera similar pero solamente considerando las submatrices de cada par de variables a partir de la reconstrucción de la matriz de Burt para cada uno de los tres casos (valores esperados, 1ra. dimensión, segunda dimensión).

Finalmente se hará la reconstrucción considerando los resultados del ACS aplicado a cada par de variables.

ACM - Evaluación de la reconstrucción de las subtablas

Se extrae la subtabla correspondiente a Facultad del egresado contra Ingreso del primer empleo relacionado con su formación para cada caso, y se obtiene los siguientes residuos:

Residuos totales - valores esperados:

	Arte	AyC	CCAAC	CCSS	CeI	Der	Edu	LLCCHH
De 1 a 2 RMV	2	3	15	4	25	15	2	5
Mas 4 RMV	1	2	7	9	20	11	3	8
Mayor 2 a 4 RMV	0	8	11	3	40	22	4	8
Menor 1 RMV	1	4	5	1	28	12	10	4
No precisa	2	3	2	1	8	6	1	7

Suma de residuos absolutos total: 321

Residuos totales - 1ra. dimensión:

	Arte	AyC	CCAAC	CCSS	CeI	Der	Edu	LLCCHH
De 1 a 2 RMV	2	2	4	2	24	8	10	4
Mas 4 RMV	1	2	4	6	27	11	4	1
Mayor 2 a 4 RMV	0	7	3	6	21	7	6	4
Menor 1 RMV	1	3	8	2	29	15	1	7
No precisa	2	3	3	0	5	4	1	6

Suma de residuos absolutos total: 257

Residuos totales - 2da dimensión:

	Arte	AyC	CCAAC	CCSS	CeI	Der	Edu	LLCCHH
De 1 a 2 RMV	2	9	0	2	23	3	10	0
Mas 4 RMV	2	3	1	6	28	14	4	4
Mayor 2 a 4 RMV	3	1	2	6	22	12	6	8
Menor 1 RMV	0	2	8	2	30	15	1	8
No precisa	7	13	9	0	3	8	0	4

Suma de residuos absolutos total: 280

Resumen

Haciendo lo mismo para las reconstrucciones halladas en el ACMA y el ACC, y agregando la aplicación del mismo ejercicio a los resultados del ACS de las tablas de contingencia correspondientes, se obtiene los resultados que se muestran en este resumen.

Se muestra la siguiente tabla con los residuos totales hallados para cada par de variables, cada técnica, y cada nivel de reconstrucción para la comparación correspondiente:

Facultad del egresado contra Ingreso del primer empleo relacionado con su formación para cada caso

	ACS	ACM	ACC	ACMA
Valores esperados	321	321	321	321
1ra. dimensión	114	257	114	273
2da. Dimensión	93	280	103	271

Facultad del egresado contra Grado de desarrollo de la Capacidad de comunicación recibida durante la formación

	ACS	ACM	ACC	ACMA
Valores esperados	369	369	369	369
1ra. dimensión	74	223	100	314
2da. Dimensión	48	289	100	314

Ingreso del primer empleo relacionado con su formación para cada caso contra Grado de desarrollo de la Capacidad de comunicación recibida durante la formación

	ACS	ACM	ACC	ACMA
Valores esperados	109	109	109	109
1ra. dimensión	73	313	102	85
2da. Dimensión	42	328	57	83

Se observa que:

- En todos los casos, el ACC reconstruye mejor la tabla original cuando se consideran 2 dimensiones y es la más aproximada a los resultados de un ACS. Se observa que la cantidad de residuos en un caso es menor que la tercera parte de su similar en el ACMA y en otro caso es menor que la sexta parte de su similar en el ACM.
- En todos los casos, la reconstrucción brindada por el ACM empeora al pasar de considerar 1 dimensión a 2 dimensiones. Peor aún, en un caso casi triplica su valor al pasar a la primera dimensión.
- En todos los casos, el ACC reconstruye mejor las tablas originales que el ACMA cuando se consideran 2 dimensiones. Es la que más se aproxima al resultado obtenido por el ACS. Incluso en dos de los casos reconstruyó mucho más rápido que el ACMA con la primera dimensión.

Se muestra el biplot obtenido a partir de la aplicación del ACC a las 3 variables.

Lo que se observa es que existen principalmente 2 grupos: el primero formado por la facultad de Ciencias e Ingeniería, la facultad de Ciencias Sociales y la facultad de Ciencias Administrativas, que suelen tener mejores sueldos en su primer empleo a pesar de que juzgan que el grado de desarrollo en su capacidad de comunicación en su formación en la PUCP no fue muy alto. El segundo grupo formado por la facultad de Derecho,

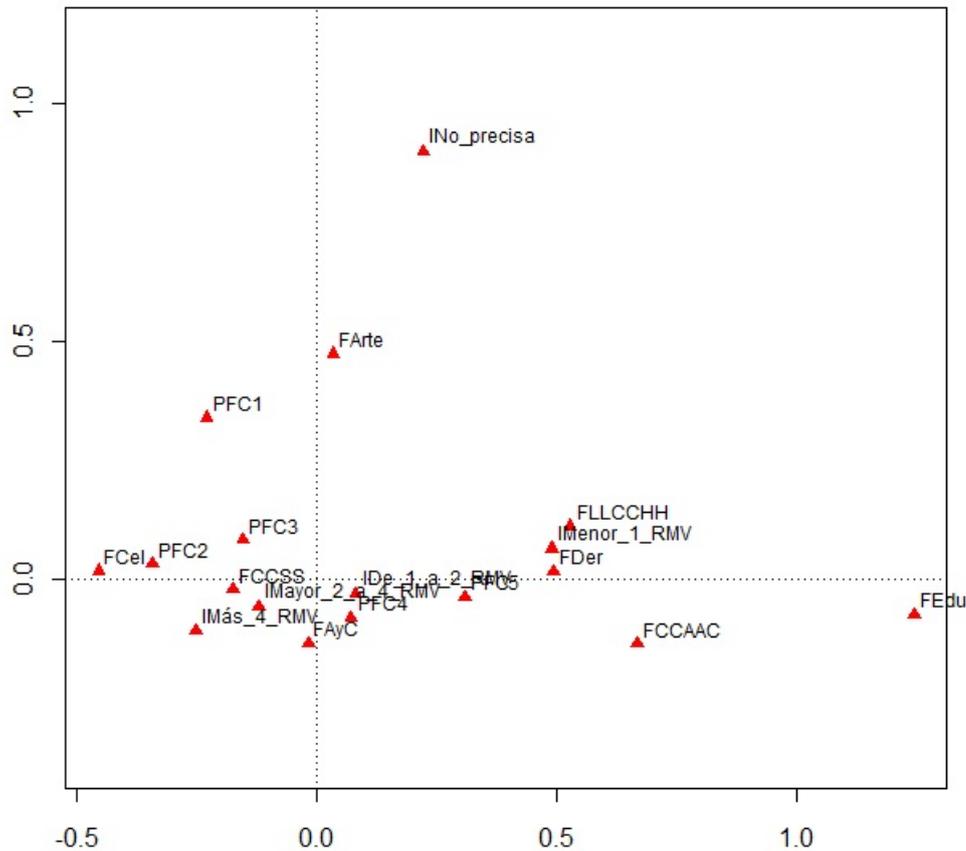


Figura 4.7: ACC - Facultad del egresado contra Ingreso del primer empleo relacionado con su formación contra Grado de desarrollo de la Capacidad de comunicación recibida durante la formación

la facultad de Ciencias y Artes de la Comunicación y la facultad de Letras y Ciencias Humanas que indican tener sueldos más bajos en su primer empleo a pesar de que juzgan que el grado de desarrollo en su capacidad de comunicación en su formación en la PUCP fue muy alto. La facultad de Arte está muy alejada y cercana a los que no precisan el sueldo en su primer empleo, por lo que no se podría concluir mucho en cuanto a esta facultad. La facultad de Educación podría situarse también en el segundo grupo, a pesar de aparecer bastante alejada del resto. Sin embargo, se debe resaltar que se está comparando las facultades de la PUCP. No puede concluirse que la capacidad de comunicación no es necesaria para obtener un mejor ingreso en el primer empleo para el caso de algunas facultades de egreso PUCP. Lo más probable es que debe haber otras características que influyen en el nivel de ingreso obtenido. Para completar una conclusión respecto a este punto, habría que hacer un estudio similar comparando facultades compatibles de diferentes universidades y ver como influye esta capacidad en el nivel de ingreso del primer empleo.

El biplot obtenido a partir de la aplicación del ACM a las 3 variables se puede apreciar en la figura 4.8; y el obtenido a partir de la aplicación del ACMA a las 3 variables puede apreciarse en la figura 4.9.

El biplot ACM se diferencia del ACMA prácticamente solo en la escala. Después, las posiciones relativas entre los puntos son similares.

El biplot ACC muestra más evidente algunos detalles esperados. Se observa por ejemplo más claramente la proximidad de la facultad de ciencias e ingeniería al valor 2 de la percepción del grado de desarrollo de la capacidad de comunicación. En el caso del ACM y ACMA se ve justo en medio del valor 2 y el valor 3. Otro detalle interesante es que la facultad de Educación aparece bastante alejada del resto de puntos en el ACC evidenciando probablemente la baja representatividad de su población, algo que no se aprecia en los biplots del ACM o el ACMA.

Finalmente, en la figura 4.10, se muestra el biplot de la aplicación del ACC a las mismas variables anteriores, pero cambiando la variable de ingreso del primer empleo por la de la percepción del grado de importancia que el egresado asigna a la capacidad de comunicación.

Lo que se aprecia en la figura 4.10, enfatizando lo representado en el eje horizontal, es que en todos los casos, a excepción de la facultad de educación y de ciencias sociales, siempre la percepción del grado de desarrollo de la capacidad está a un punto de distancia por debajo de la percepción del grado de importancia que el egresado asigna a esta capacidad. En el caso de la facultad de ingeniería la percepción del grado de desarrollo está cerca del valor 3 pero la percepción del grado de importancia está más cerca del valor 4. El resto de facultades, a excepción de las indicadas, están más cerca del valor 4 de percepción del grado de desarrollo y más cerca del valor 5 de percepción del grado de importancia. El caso de ciencias sociales muestra estar en ambas percepciones a una distancia similar del valor 4. Y el caso de educación, muestra una percepción del grado de desarrollo más cercano al valor 5 que lo que dista al valor 5 de la percepción del grado de importancia.

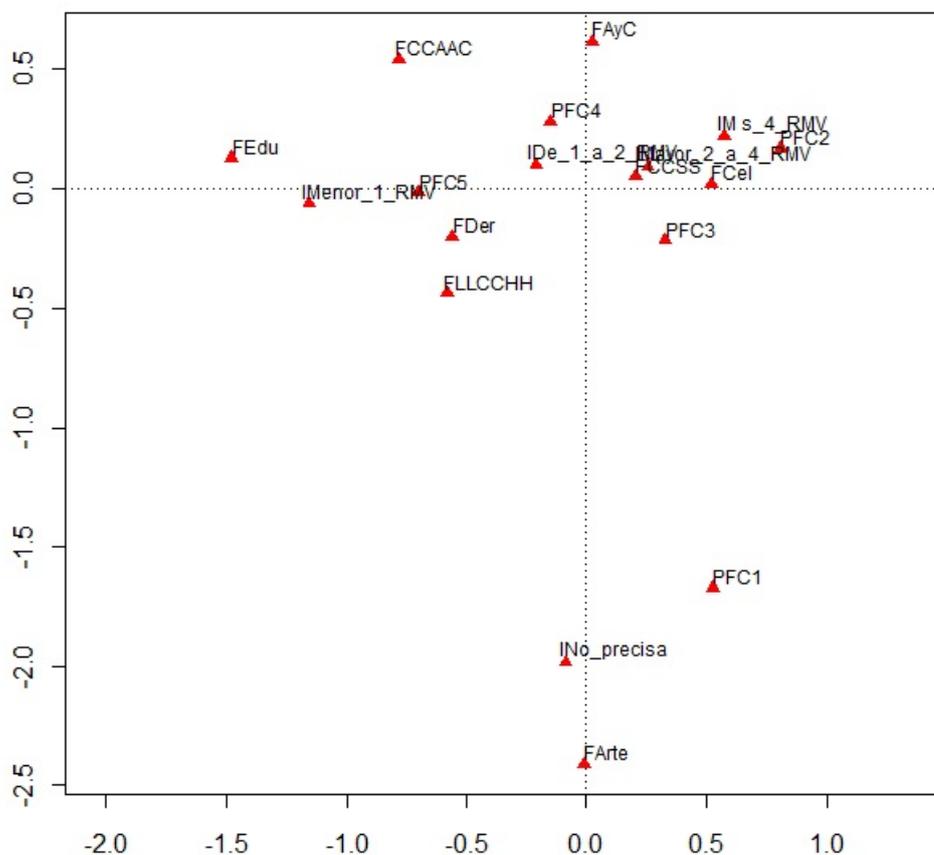


Figura 4.8: ACM - Facultad del egresado contra Ingreso del primer empleo relacionado con su formación contra Grado de desarrollo de la Capacidad de comunicación recibida durante la formación

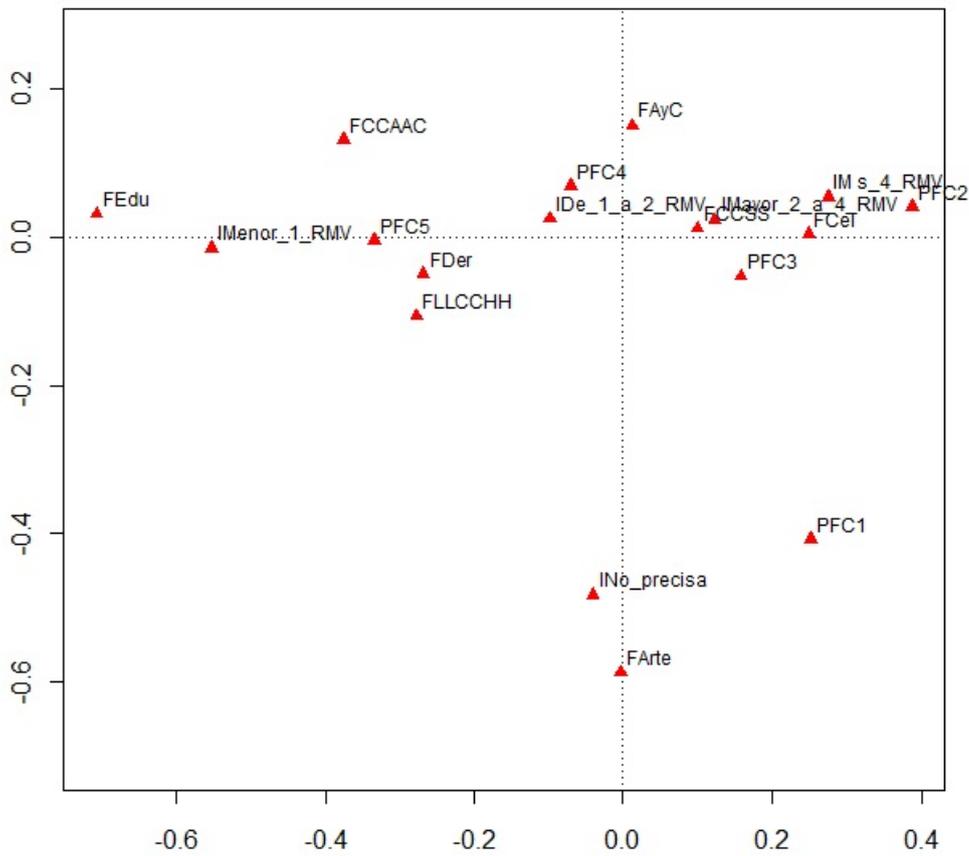


Figura 4.9: ACMA - Facultad del egresado contra Ingreso del primer empleo relacionado con su formación contra Grado de desarrollo de la Capacidad de comunicación recibida durante la formación

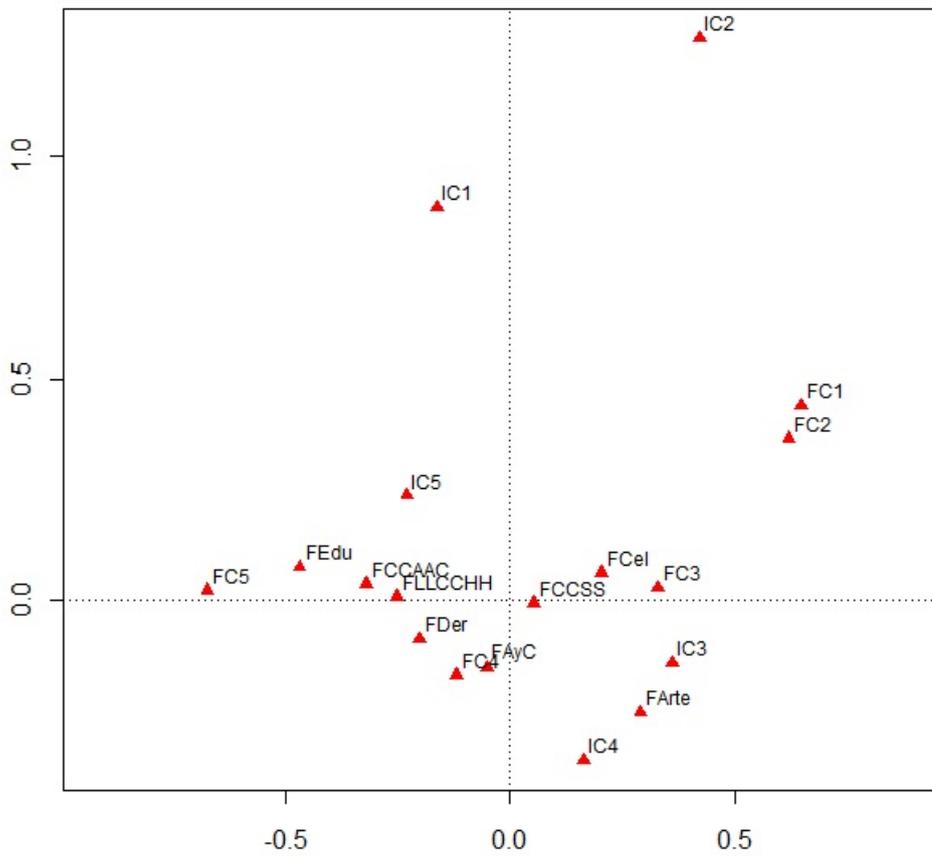


Figura 4.10: ACC - Facultad del egresado contra Percepción del Grado de desarrollo de la Capacidad de comunicación recibida durante la formación contra Percepción del Grado de importancia que el egresado asigna a la Capacidad de comunicación para la obtención del 1er. empleo

Capítulo 5

Conclusiones

5.1. Conclusiones

La propuesta de este trabajo fue la de profundizar en el estudio de las modificaciones al análisis de correspondencias múltiples (ACM): el análisis de correspondencias conjunto (ACC) de Greenacre (1988) y el análisis de correspondencias ajustado (ACMA) de Greenacre (2006), exponer sus características y diferencias aplicando las técnicas a una data real con el apoyo de las implementaciones desarrolladas en R y comparar sus resultados mediante la reproducción de las tablas de contingencias originales en base al planteamiento de Camiz (2010).

Éstas técnicas representan un tema poco explorado en la literatura, más aún en nuestro idioma.

Durante el desarrollo de este trabajo, además se ha enfatizado mucho en la base teórica del ACS sobre la cual se basan las demás técnicas. Se ha hecho un desarrollo exhaustivo y claro de su teoría y esto también representa un valor agregado de este trabajo que no se encuentra a ese nivel en la bibliografía disponible.

Las técnicas del análisis de correspondencias ofrecen un poderoso entorno para facilitar la visualización de datos categóricos y las relaciones de similitud entre ellos, lo que se ha comprobado con la aplicación realizada en esta tesis.

Se puede resumir las principales conclusiones de la siguiente manera:

- Este trabajo pretende mostrar y difundir las técnicas de análisis de correspondencia, presentando sobre todo las características de sus variaciones en el análisis de correspondencia conjunto y múltiple ajustado, no tan difundidos en nuestro medio.
- Se muestran los fundamentos teóricos y los algoritmos utilizados en los cálculos sobre los que se basan las funciones desarrolladas en R para estos casos. Los fundamentos teóricos del ACS están explicados a un nivel de detalle muy específico que hace muy claro su entendimiento.
- Para quien no lo conoce, muchas veces el contar con variables categóricas puede limitar el análisis de los datos. Estas técnicas permiten trabajar con variables categóricas medidas en cualquier escala ya que se basa en tablas de contingencia que solo enfatiza en la frecuencia de ocurrencias de cada respuesta.

- En el caso particular de la encuesta usada para la aplicación en esta tesis, no se hace uso de estas técnicas. Los resultados hallados con la aplicación de éstas técnicas refuerzan y algunos casos dan más información que los resultados oficiales de la encuesta basados simplemente en la observación de las tablas de contingencia originales.
- Se ha comprobado que el ACS es un caso especial de la aplicación del ACC a dos variables.
- En cuanto a la reconstrucción de las tablas, se ha concluido de la misma manera que el estudio de Camiz, 2010 comprobando que el ACC reconstruye de mejor manera la matriz de Burt original. Centrándose en la reconstrucción de las subtablas fuera de la diagonal, se observa que en todos los casos:
 - El ACC reconstruye mejor la tabla original cuando se consideran 2 dimensiones y es la más aproximada a los resultados de un ACS.
 - La reconstrucción brindada por el ACM incluso empeora al pasar de considerar más dimensiones.
 - El ACC reconstruye mejor las tablas originales que el ACMA cuando se consideran 2 dimensiones.
- Tal como cita Camiz (2010), es muy extraño que no existan registros en la literatura sobre este serio inconveniente que presenta el ACM, mientras que por otro lado, el ACC se muestra como el desarrollo más prometedor y sus propiedades merecen estudios más profundos.
- En los resultados obtenidos con la aplicación de la técnica del ACC, se observa que existen principalmente 2 grupos: el primero formado por la facultad de Ciencias e Ingeniería, la facultad de Ciencias Sociales y la facultad de Ciencias Administrativas, que suelen tener mejores sueldos en su primer empleo a pesar de que juzgan que el grado de desarrollo en su capacidad de comunicación en su formación en la PUCP no fue muy alto. El segundo grupo formado por la facultad de Derecho, la facultad de Ciencias y Artes de la Comunicación y la facultad de Letras y Ciencias Humanas que indican tener sueldos más bajos en su primer empleo aunque juzgan que el grado de desarrollo en su capacidad de comunicación en su formación en la PUCP fue muy alto. La facultad de Arte está muy alejada del resto de facultades y es la más cercana a los que no precisan el sueldo en su primer empleo, por lo que no se podría concluir mucho en cuanto a esta facultad. La facultad de Educación podría situarse también en el segundo grupo, a pesar de aparecer bastante alejada del resto. Sin embargo, se debe resaltar que se está comparando las facultades de la PUCP. No puede concluirse que la capacidad de comunicación no es necesaria para obtener un mejor ingreso en el primer empleo para el caso de algunas facultades de egreso PUCP. Lo más probable es que debe haber otras características que influyen en el nivel de ingreso obtenido. En todo caso, para completar una conclusión respecto a este punto, habría que hacer un estudio similar comparando facultades compatibles de diferentes universidades y ver como influye esta capacidad en el nivel de ingreso del primer empleo.

- En los resultados obtenidos en el biplot: Facultad del egresado contra Percepción del Grado de desarrollo de la Capacidad de comunicación recibida durante la formación contra Percepción del Grado de importancia que el egresado asigna a la Capacidad de comunicación para la obtención del 1er. empleo, se muestra en casi todos los casos un desfase de 1 punto entre ambas percepciones, evidenciando que los egresados perciben que necesitan una mejor formación en la capacidad de comunicación.
- Los resultados mostrados en el biplot del análisis de correspondencias de las 3 variables del punto anterior son similares en el ACM y el ACMA. Solamente difieren en la escala, pero la disposición relativa de los puntos es similar. No sucede lo mismo con el biplot del ACC. El biplot del ACC evidencia algunos detalles esperados que no muestra el ACM ni el ACMA.
- Las técnicas de análisis de correspondencias deberían utilizarse siempre en el análisis de los resultados de encuestas cuando se obtienen resultados de variables categóricas que dependen unas de otras.

5.2. Sugerencias para investigaciones futuras

- Se podría desarrollar funciones en R para poder reproducir automáticamente las tablas de contingencia originales a partir de los resultados obtenidos en cada una de las técnicas del análisis de correspondencia e incluso dar alguna medida de aproximación de exactitud.
- Se podría realizar un estudio que compare las diferentes capacidades en facultades de egreso compatibles de diferentes universidades y ver el impacto que tienen en el nivel de ingreso de los primeros empleos.
- Particularmente en el estudio realizado, la aplicación puede ampliarse para el resto de ejes de estudio: situación laboral, valoración de la PUCP y estudios de posgrado. Además del análisis que puede hacerse a cada una de las capacidades planteadas y extenderlo además a nivel de la variable Especialidad del egresado.
- Se recomienda especialmente aplicar la técnica a los resultados que se obtendrán en el IV Estudio institucional de egresados aplicándose actualmente (2011) a los mismos egresados del 2006 y a los egresados del 2009. Las comparaciones darán importantes conclusiones sobre la efectividad de las medidas que las Facultades hayan podido tomar como consecuencia de los resultados obtenidos en el anterior estudio.
- Siendo tan útiles y prácticas estas técnicas, ya que en general pueden utilizarse como parte de la rutina de análisis de resultados de encuestas en general, debería proponerse un curso específico en la PUCP sobre este tema que promueva además nuevas investigaciones relacionadas.

Apéndice A

Resultados teóricos

A.1. Álgebra matricial

Se describirá en este apéndice algunos resultados del álgebra matricial que son utilizados en el trabajo. Mayores detalles de estos resultados así como sus demostraciones pueden encontrarse en el texto de Magnus, J. y Neudecker, H. (2007).

En lo que sigue se entenderá a todo vector en \mathbb{R}^n como un vector columna de orden $n \times 1$. Dada una matriz $A = [a_{ij}]$ de orden $n \times m$, que siempre será asumida con entradas reales, se podría considerar tanto a sus filas como columnas como vectores en \mathbb{R}^m y \mathbb{R}^n , respectivamente. Se llamará el rango fila de A al número de filas linealmente independientes de A y el rango columna de A al número de columnas linealmente independientes de A . Es posible mostrar que tanto el rango fila como el rango columna son siempre los mismos por lo que se llamará a este número el rango de A y se denotará como $r(A)$. Claramente $r(A) \leq \min\{n, m\}$.

Si A es una matriz $n \times n$, que simplemente se denominará, cuadrada de orden n , entonces se dirá que A es no-singular si $r(A) = n$. En caso contrario se dirá que A es singular.

Se dirá que una matriz cuadrada real $A = [a_{ij}]$ de orden n es ortogonal si

$$A^T A = AA^T = I_n$$

y sus vectores columna son ortogonales en \mathbb{R}^n , donde I_n denota a la matriz identidad de orden n .

Una matriz rectangular (no cuadrada) $A = [a_{ij}]$ de orden $n \times m$ puede también tener la propiedad que $A^T A = I_m$ ó $AA^T = I_n$ pero no ambas. Tal matriz es llamada semi-ortogonal.

Dada una matriz $A = [a_{ij}]$ de orden $n \times m$ se puede, similarmente como en el caso de un vector, definir la norma de la matriz $\|A\|$. Estas normas no son únicamente definidas y poseen todas las propiedades conocidas de la norma de vectores:

- a. $A \neq 0 \Rightarrow \|A\| > 0$
- b. $\|\alpha A\| = |\alpha| \|A\|, \forall \alpha \in \mathbb{R}$
- c. $\|A + B\| \leq \|A\| + \|B\|$

con el agregado de que puede cumplir con la propiedad de consistencia:

d. $\|AB\| \leq \|A\|\|B\|$

Todas las normas de matrices de orden finito son equivalentes y en estas se puede citar por ejemplo a:

- La norma inducida por la norma euclideana o también llamada de Frobenius:

$$\|A\|_F = \left(\sum_i^m \sum_j^n |a_{ij}|^2 \right)^{1/2}$$

- La norma infinito o norma fila:

$$\|A\|_\infty = \|A\|_{MFS} = \max_{1 \leq i \leq n} \sum_{j=1}^m |a_{ij}|$$

A.2. Traza de una matriz

La traza de una matriz cuadrada $A = [a_{ij}]$ de orden $n \times n$, denotada por $tr(A)$, es la suma de sus elementos diagonales:

$$tr(A) = \sum_{i=1}^n a_{ii}.$$

Se tiene para ella las propiedades siguientes:

Proposición A.2.1 . Para cualesquiera matrices A, B y un número λ se cumplen que:

i) $tr(A + B) = tr(A) + tr(B)$

ii) $tr(\lambda A) = \lambda tr(A)$

iii) $tr(A)^T = tr(A)$

iv) $tr(AB) = tr(BA)$

La norma de Frobenius se puede escribir alternativamente de la forma siguiente:

$$\|A\|_F = \sqrt{tr(A^T A)}$$

A.3. Autovalores y Autovectores

Sea A una matriz cuadrada de orden n . Los autovalores de A están definidos como las raíces de la ecuación característica:

$$|A - \lambda I_n| = 0$$

Esta ecuación tiene n raíces, en general complejas, donde algunas de ellas pueden repetirse. En este último caso se hará referencia al autovalor como múltiple y en caso contrario como simple.

Si λ es un autovalor de A , entonces existen vectores $x \in \mathbb{R}^n \setminus \{0\}$ tales que:

$$(A - \lambda I_n)x = 0;$$

esto es:

$$Ax = \lambda x.$$

Estos son los llamados autovectores de A asociados al autovalor λ . Note que si se multiplica la ecuación anterior por cualquier constante no nula, ello genera automáticamente un nuevo autovector asociado al mismo autovalor. Para evitar esta falta de unicidad se acostumbra normalizar al autovector, por ejemplo imponiendo que este tenga norma 1; es decir, que $x^\top x = 1$.

Los autovalores y sus correspondientes autovectores comparten una serie de propiedades interesantes que se detalla a continuación.

Proposición A.3.1 *Sea A una matriz cuadrada de orden n , entonces*

- i) Si A es simétrica, entonces sus autovalores son reales.*
- ii) Si G es otra matriz cuadrada no-singular de orden n , entonces A y $G^{-1}AG$ tienen los mismos autovalores.*
- iii) Si A es singular, entonces A debe tener al menos un autovalor nulo.*
- iv) A es una matriz simétrica, entonces esta es positiva definida (positiva semidefinida) si, y solamente si, todos su autovalores son positivos (no-negativos).*
- v) Los autovectores asociados con diferentes autovalores son linealmente independientes.*
- vi) Si A es simétrica y tiene r autovalores no nulos, entonces $r(A) = r$.*
- vii) La traza de A es la suma de sus valores propios y el determinante de A es el producto de sus valores propios.*

A.4. El teorema de descomposición en valores singulares y el teorema de Eckart y Young

Otro resultado fundamental en nuestro desarrollo está constituido por el teorema de descomposición en valores singulares que se detalla a continuación.

Teorema A.4.1 (Descomposición en valores singulares) *Dada un matriz A de dimensiones $n \times m$ y rango r , existen matrices cuadradas ortogonales de ordenes n y m ; U_n y V_m , y una matriz diagonal de orden r , Λ , tales que:*

$$A = U_n \Gamma V_m^T, \quad (\text{A.1})$$

donde $\Gamma = \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix}$ es una matriz $n \times m$ y $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_r \end{bmatrix}$

con $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$.

A los valores no negativos $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$ se les denominan los valores singulares de la matriz A y a los vectores columna en U_n y V_m :

$$U_n = [u_1 \ u_2 \ u_3 \ \dots \ u_r \ u_{r+1} \ u_{r+2} \ \dots \ u_n]$$

y

$$V_m = [v_1 \ v_2 \ v_3 \ \dots \ v_r \ v_{r+1} \ v_{r+2} \ \dots \ v_m]$$

se les denominan los vectores singulares a izquierda y derecha de A . Note que si se considera sólo las primeras r columnas de ambas matrices y se define las matrices:

$$U = [u_1 \ u_2 \ u_3 \ \dots \ u_r]$$

y

$$V = [v_1 \ v_2 \ v_3 \ \dots \ v_r],$$

de ordenes respectivos $n \times r$ y $m \times r$ el teorema de descomposición en valores singulares queda simplemente expresado como:

$$A = U \Lambda V^T .$$

A menos que se indique lo contrario se emplea esta versión.

La descomposición en valores singulares de A revela que esta matriz puede escribirse entonces como:

$$A = \sum_{i=1}^r \lambda_i u_i v_i^T .$$

Cuando la matriz A es cuadrada y simétrica, el teorema de descomposición en valores singulares es conocido también como el teorema de descomposición espectral.

En términos de álgebra matricial, las matrices U y V representan un simple cambio de coordenadas y poseen una estrecha relación con los vectores propios y valores propios de las matrices cuadradas y simétricas AA^T y $A^T A$, respectivamente. En efecto:

$$\begin{aligned} AA^T &= U\Lambda V^T V\Lambda U^T \\ &= U\Lambda^2 U^T \end{aligned}$$

y

$$\begin{aligned} A^T A &= V\Lambda U^T U\Lambda V^T \\ &= V\Lambda^2 V^T. \end{aligned}$$

Los valores singulares de A son las raíces cuadradas de los autovalores de las matrices AA^T ó $A^T A$.

Teorema A.4.2 (Eckart-Young) *Sea X una matriz $n \times m$ y sea la definición de la función de valor real*

$$\phi(A, Z) = \|X - ZA^T\|$$

donde Z es una matriz $n \times r$ y A es una matriz $m \times r$ semi-ortogonal con $A^T A = I_r$. El valor mínimo de ϕ se obtiene tomando $Z = XA$ y a A como la matriz conformada por los r autovectores ortonormales asociados a los r mayores autovalores de $X^T X$. Este valor mínimo viene dado por la suma de los $m - r$ menores autovalores de $X^T X$.

Refraseando, el teorema de Eckart-Young provee esencialmente de la mejor aproximación (en el sentido de la norma de Frobenius) de rango r , $\tilde{X} = ZA^T = XAA^T$, de X . Esta matriz puede también caracterizarse como la parte de la descomposición en valores singulares de X correspondiente a los r mayores autovalores de $X^T X$, tal como lo explicita el resultado siguiente.

Corolario A.4.1 *La mejor aproximación de rango r de una matriz X de orden $n \times m$ viene dada por:*

$$\tilde{X} = U\Lambda^{1/2}V^T,$$

donde U es la matriz $n \times r$ conformada por los r primeros vectores singulares a izquierda de X , Λ es la matriz diagonal de los r primeros autovalores de $X^T X$ y V es la matriz $m \times r$ conformada por los r primeros vectores singulares a derecha de X .

Demostración: Supóngase, sin pérdida de generalidad, que $r(X) = m$ y sean

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m > 0$$

los autovalores de $X^T X$. Sea $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$ y sea

$$X = U_n \Lambda^{1/2} V_m^T$$

la descomposición en valores singulares de X , con $U_n^\top U_n = I_n$ y $V_m^\top V_m = I_m$. Se define

$$\Lambda_1 = \text{diag}(\lambda_1, \dots, \lambda_r) \text{ y } \Lambda_2 = \text{diag}(\lambda_{r+1}, \dots, \lambda_m)$$

y se considera las particiones $U_n = [U_1 \ U_2]$ y $V_m = [V_1 \ V_2]$, donde $U_1 = [u_1 \ u_2 \ \dots \ u_r]$ y $V_1 = [v_1 \ v_2 \ \dots \ v_r]$ están conformados por los primeros autovectores singulares a izquierda y, respectivamente, derecha de X . Entonces

$$X = U_1 \Lambda_1^{1/2} V_1^\top + U_2 \Lambda_2^{1/2} V_2^\top$$

y

$$\begin{aligned} X^\top X &= (U_1 \Lambda_1^{1/2} V_1^\top + U_2 \Lambda_2^{1/2} V_2^\top)^\top (U_1 \Lambda_1^{1/2} V_1^\top + U_2 \Lambda_2^{1/2} V_2^\top) \\ &= (V_1 \Lambda_1^{1/2 \top} U_1^\top + V_2 \Lambda_2^{1/2 \top} U_2^\top) (U_1 \Lambda_1^{1/2} V_1^\top + U_2 \Lambda_2^{1/2} V_2^\top) \\ &= V_1 \Lambda_1^{1/2 \top} U_1^\top U_1 \Lambda_1^{1/2} V_1^\top + V_1 \Lambda_1^{1/2 \top} U_1^\top U_2 \Lambda_2^{1/2} V_2^\top + V_2 \Lambda_2^{1/2 \top} U_2^\top U_1 \Lambda_1^{1/2} V_1^\top + V_2 \Lambda_2^{1/2 \top} U_2^\top U_2 \Lambda_2^{1/2} V_2^\top \\ &= V_1 \Lambda_1 V_1^\top + V_2 \Lambda_2 V_2^\top. \end{aligned}$$

Así,

$$X^\top X V_1 = V_1 \Lambda_1 V_1^\top V_1 + V_2 \Lambda_2 V_2^\top V_1$$

$$X^\top X V_1 = V_1 \Lambda_1$$

y V_1 está conformado por los autovectores asociados a los primeros r autovalores de la matrix $X^\top X$. Por tanto, según el teorema de Eckart-Young V_1 es igual a la matrix A mencionada en este teorema y se tiene que

$$\begin{aligned} \tilde{X} &= X A A^\top = X V_1 V_1^\top = (U_1 \Lambda_1^{1/2} V_1^\top + U_2 \Lambda_2^{1/2} V_2^\top) V_1 V_1^\top \\ &= U_1 \Lambda_1^{1/2} V_1^\top V_1 V_1^\top + U_2 \Lambda_2^{1/2} V_2^\top V_1 V_1^\top = U_1 \Lambda_1^{1/2} V_1^\top. \end{aligned}$$

Proposición A.4.1 *El primer autovalor y los primeros vectores singulares de Q en (2.4) vienen dados por*

$$\lambda_1 = 1, \quad u_1 = D_r^{\frac{1}{2}} \mathbf{1}_I \quad \text{y} \quad v_1 = D_c^{\frac{1}{2}} \mathbf{1}_J,$$

donde $\mathbf{1}_I$ y $\mathbf{1}_J$ denotan, respectivamente, a los vectores columna de 1's de ordenes I y J .

Demostración: Dada una matrix $A = [a_{ij}]$ de orden $n \times m$, se considerará por conveniencia la norma fila de esta matrix, es decir, $\|A\|_{MSF} = \max_{1 \leq i \leq n} \sum_{j=1}^m |a_{ij}|$. Recuérdese que $\|Az\|_{MSF} \leq \|A\|_{MSF} \|z\|_{MSF}$. Ahora, sea t un autovector de una matrix cuadrada A correspondiente al autovalor λ , tal que $At = \lambda t$. Entonces,

$$|\lambda| \|t\|_{MSF} = \|At\|_{MSF} \leq \|A\|_{MSF} \|t\|_{MSF}$$

y

$$|\lambda| \leq \|A\|_{MSF} \dots (1)$$

Sea

$$v_1 = D_c^{-1/2} \mathbf{1}_j.$$

Note que la norma euclidiana de este vector es 1. Como:

$$Q^\top Q = \frac{1}{n^2} D_c^{-1/2} N^\top D_r^{-1/2} D_r^{-1/2} N D_c^{-1/2} = \frac{1}{n^2} D_c^{-1/2} N^\top D_r^{-1} N D_c^{-1/2}.$$

se obtiene al multiplicar por v_1 que:

$$Q^\top Q v_1 = \frac{1}{n^2} D_c^{-1/2} N^\top D_r^{-1} N D_c^{-1/2} D_c^{1/2} \mathbf{1}_j = \frac{1}{n^2} D_c^{-1/2} N^\top D_r^{-1} N \mathbf{1}_j.$$

Puesto que:

$$D_r^{-1} N \mathbf{1}_j = \begin{bmatrix} \frac{n}{n_1} & & \\ & \ddots & \\ & & \frac{n}{n_I} \end{bmatrix} \begin{bmatrix} n_1 \\ \vdots \\ n_I \end{bmatrix} = n \mathbf{1}_I$$

y similarmente $D_c^{-1} N \mathbf{1}_I = n \mathbf{1}_j$ se obtiene reemplazando que:

$$\begin{aligned} Q^\top Q v_1 &= \frac{1}{n^2} D_c^{-1/2} N^\top n \mathbf{1}_I = \frac{1}{n} D_c^{-1/2} N^\top \mathbf{1}_I = \frac{1}{n} D_c^{-1/2} D_c D_c^{-1} N^\top \mathbf{1}_I \\ &= \frac{1}{n} D_c^{1/2} D_c^{-1} N^\top \mathbf{1}_I = D_c^{1/2} \mathbf{1}_J = v_1 \end{aligned}$$

Es decir 1 es un autovalor de $Q^\top Q$ y v_1 su correspondiente autovector o valor singular a derecha de Q . Análogamente se prueba que 1 es un autovalor de $Q Q^\top$ y $u_1 = D_r^{-1/2} \mathbf{1}_I$ es el correspondiente vector singular a izquierda de Q ($Q Q^\top u_1 = u_1$). De otro lado, como

$$Q^\top Q v_1 = n v_1 \frac{1}{n^2} D_c^{-1/2} N^\top D_r^{-1} N D_c^{-1/2} D_c^{1/2} \mathbf{1}_J = D_c^{1/2} \mathbf{1}_J.$$

Simplificando:

$$\frac{1}{n^2} D_c^{-1/2} N^\top D_r^{-1} N \mathbf{1}_J = D_c^{1/2} \mathbf{1}_J.$$

Multiplicando por $D_c^{-1/2}$ a la izquierda:

$$D_c^{-1/2} D_c^{-1/2} N^\top D_r^{-1} N \mathbf{1}_J = n^2 D_c^{-1/2} D_c^{1/2} \mathbf{1}_J.$$

Simplificando:

$$D_c^{-1} N^\top D_r^{-1} N \mathbf{1}_J = n^2 \mathbf{1}_J.$$

De acá se deduce que todas las filas de $D_c^{-1} N^\top D_r^{-1} N$ suman n^2 . Por lo tanto se cumple que:

$$\|D_c^{-1} N^\top D_r^{-1} N\|_{MSF} = n^2.$$

Sea $M = D_c^{-1}N^\top D_r^{-1}N$, entonces:

$$M = D_c^{-1/2}D_c^{-1/2}N^\top D_r^{-1}ND_c^{-1/2}D_c^{1/2} = n^2 D_c^{-1/2}Q^\top QD_c^{-1/2} = n^2 Q^\top Q.$$

al ser $D_c^{-1/2}$ y $D_c^{1/2}$ matrices diagonales. Así,

$$\|Q^\top Q\|_{MSF} = \frac{1}{n^2}\|M\|_{MSF} = 1$$

Por tanto, como por (1), para cualquier autovalor de $Q^\top Q$, se cumple que $|\lambda| \leq \|Q^\top Q\|_{MSF} = 1$, los autovalores de $Q^\top Q$ cumplen con el teorema de descomposición en valores singulares que $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k > 0$ y 1 es un autovalor de $Q^\top Q$, se tiene que $\lambda_1 = 1$; siendo los primeros vectores singulares a izquierda y derecha de Q los vectores $u_1 = D_r^{1/2}\mathbf{1}_I$ y $v_1 = D_c^{1/2}\mathbf{1}_J$.

◇



Apéndice B

Encuesta y más aplicaciones

B.1. Ficha técnica de la encuesta aplicada en el III Estudio Institucional sobre Egresados PUCP

B.1.1. Objetivo general

Dar a conocer el proceso de inserción y la situación laboral en el año 2008, la valoración de la formación y los estudios post egreso de los egresados de pregrado PUCP 2006-1, 2006-2 y 2007-0.

B.1.2. Objetivos específicos

A partir del estudio se trata de responder a las siguientes preguntas:

- Sobre el proceso de inserción laboral:
 - ¿Cuán frecuente es que nuestros estudiantes trabajen o realicen prácticas pre-profesionales durante sus estudios en la universidad?
 - ¿Cuánto tiempo demoran nuestros estudiantes en obtener su primer empleo en una actividad relacionada con su formación?
 - ¿Qué características tiene el primer empleo relacionado a su formación?
- Sobre la situación laboral actual (año 2008):
 - ¿Cuál es la situación laboral de los egresados PUCP?
 - ¿Qué características tiene el empleo actual del egresado PUCP?
- Sobre la valoración PUCP:
 - ¿Cuán satisfecho está el egresado PUCP con diferentes aspectos de la formación?
 - ¿Qué capacidades desarrolladas por la formación PUCP fueron necesarias para el desempeño en su primer empleo?, ¿cuáles no se desarrollaron?
 - ¿Cuál es la percepción de los egresados con respecto a la mejor universidad en su especialidad?

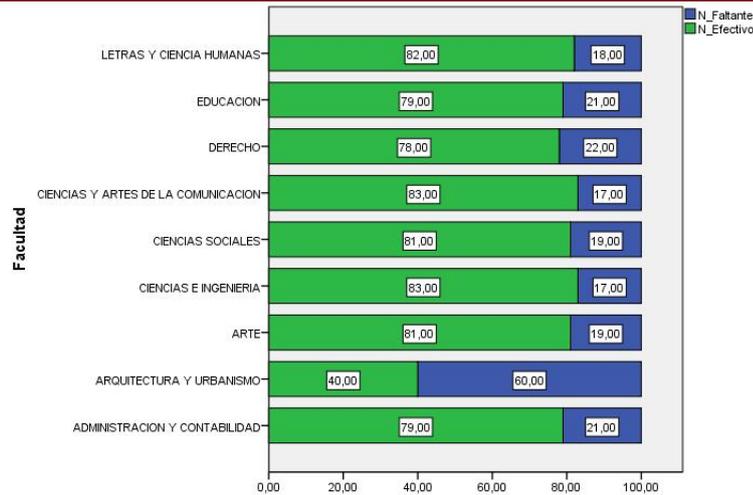


Figura B.1: Cobertura del estudio por facultad

- ¿Cuál es la valoración de la PUCP respecto a otras universidades del país?
- Sobre estudios luego de egresar:
 - ¿Qué tipo de programas educativos cursan nuestros egresados una vez concluidos sus estudios de pregrado en la PUCP?

B.1.3. Método de estudio

El método de estudio empleado es de tipo cuantitativo a nivel descriptivo porque recoge información sobre situaciones y comportamiento tal como se manifiestan, es decir, de manera independiente.

B.1.4. Población de estudio y cobertura

La población de estudio comprende la cohorte de los egresados de pregrado PUCP, semestres 2006-1, 2006-2 y 2007-0. Por tanto, es un estudio censal realizado en el 2008 con egresados que tenían uno y dos años de egreso. Los egresados que cumplen esta característica, según la base de datos institucional de la PUCP son 1660. Al finalizar el estudio se obtuvo una tasa de efectividad alta a nivel institucional (81%). En el caso de facultades PUCP, a excepción de Arquitectura y Urbanismo, se pudo encuestar a más del 78% de los egresados. Ver gráfico 5.1.

B.1.5. Ejes y variables de análisis

El estudio tiene cuatro ejes de análisis: inserción laboral, situación laboral actual (año 2008), valoración PUCP y estudios luego de egresar. Estos ejes se organizan en un conjunto de variables de estudio, como se indica en el cuadro 5.1.

Ejes de análisis		Algunas variables
Inserción laboral	Experiencia laboral antes de egresar	Categoría ocupacional Horas laboradas Ingreso mensual bruto Tiempo entre egresar y obtener el primer empleo Tiempo entre egresar y obtener el primer empleo Mecanismos de información para el primer empleo Actividad económica Categoría ocupacional Tipo de contrato Tipo de empleador Horas laboradas Ingreso mensual bruto Nivel de adecuación del primer empleo relacionado con su formación
	Primer empleo relacionado con su formación luego de egresar	Capacidades importantes para el primer empleo que han sido desarrolladas por la formación Capacidades importantes para el primer empleo que no han sido desarrolladas por la formación Aspectos de la formación que afectaron al desempeño en su primer empleo Aspectos de la formación importantes en el proceso de selección de personal para el primer empleo
Situación laboral actual (año 2008)	Condición de la actividad laboral actual	Condición de la actividad Razones por las cuales los desempleados no buscan empleo Tiempo de búsqueda de empleo Grupo ocupacional
	Características de los ocupados actualmente	Actividad económica Categoría ocupacional Tipo de contrato Tipo de empleador Horas laboradas a la semana Ingreso mensual bruto Afinidad de su ocupación con otras especialidades
Valoración PUCP		Importancia de los estudios generales para la formación Utilidad de los estudios generales para el desempeño en el primer empleo Valoración de diferentes aspectos de la formación PUCP Percepción de la mejor universidad en su especialidad Valoración de la PUCP con respecto a otras universidades nacionales
Estudios luego de egresar		Tipos de programas Lugar de estudios

Cuadro B.1: Ejes y variables del III Estudio de Egresados PUCP

B.1.6. Técnica e instrumento

El instrumento desarrollado tuvo el formato de encuesta telefónica. Constó de 67 preguntas en total. Si bien la mayoría fueron preguntas cerradas (respuesta única o múltiple) hubo 9 preguntas abiertas. Las encuestas fueron realizadas por la empresa IPSOS Apoyo Opinión y Mercado utilizando el sistema CATI (Computer-Assited Telephone Interviewing), el cual permite realizar entrevistas telefónicas y tener respuestas grabadas directamente en una computadora. El tiempo promedio para responder la encuesta fue de 20 minutos.

B.1.7. Procedimiento

Las encuestas telefónicas se realizaron entre el 13 de agosto y el 10 de diciembre del 2008. El promedio de intento de llamadas a los egresados fue 19.2 y las horas con mayor número de llamadas efectivas fueron de 6 a 9pm. Finalizado el trabajo de campo se logró 81 % de entrevistas efectivas a nivel institucional. El 19 % restante corresponde a egresados que no aceptaron la entrevista, entrevistas que quedaron incompletas o egresados que nunca respondieron las llamadas.

B.2. Mas Aplicaciones

B.2.1. Aplicación del ACS

ACS - Facultad del egresado Vs Categoría ocupacional del primer empleo relacionado con su formación

dim	value	%	cum %	scree plot
1	0.058693	47.7	47.7	*****
2	0.034167	27.7	75.4	*****
3	0.024367	19.8	95.2	*****
4	0.004001	3.2	98.4	**
5	0.001737	1.4	99.8	*
6	0.000196	0.2	100.0	
Total:		0.123161	100.0	

El porcentaje de inercia representado por las dos primeras dimensiones es del 75.4 %. Las conclusiones evidentes en el biplot de la figura 4.3 indican que los egresados de Ciencias e Ingeniería, Administración y Contabilidad y Educación suelen tener empleos dependientes. Los de Derecho y Ciencias y Artes de la Comunicación, tienden más a ser practicantes en su primer empleo y en siguiente lugar a ser independientes. Del resto de facultades no se podría dar mucha información ya que están influenciados por su falta de precisión en sus respuestas o porque no aplicaban.

ACS - Facultad del egresado Vs Grado de importancia que la Capacidad de comunicación tiene para realizar el 1er empleo

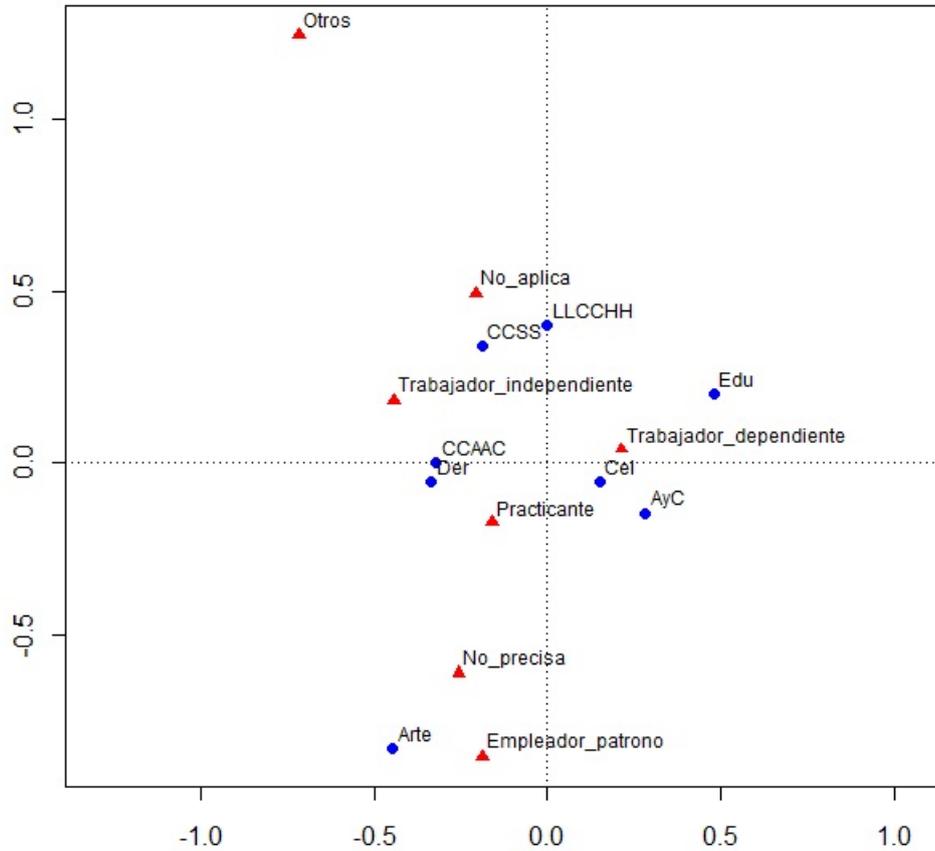


Figura B.2: Facultad del egresado Vs Categoría ocupacional del primer empleo relacionado con su formación

Principal inertias (eigenvalues):

dim	value	%	cum %	scree plot
1	0.030807	50.6	50.6	*****
2	0.013648	22.4	73.0	*****
3	0.008687	14.3	87.2	*****
4	0.005545	9.1	96.3	****
5	0.001737	2.9	99.2	*
6	0.000508	0.8	100.0	
Total:		0.060931	100.0	

El porcentaje de inercia representado por las dos primeras dimensiones es del 73%. En este caso la primera dimensión aporta el 50.5% por lo que la segunda dimensión tiene una contribución considerable en la misma.

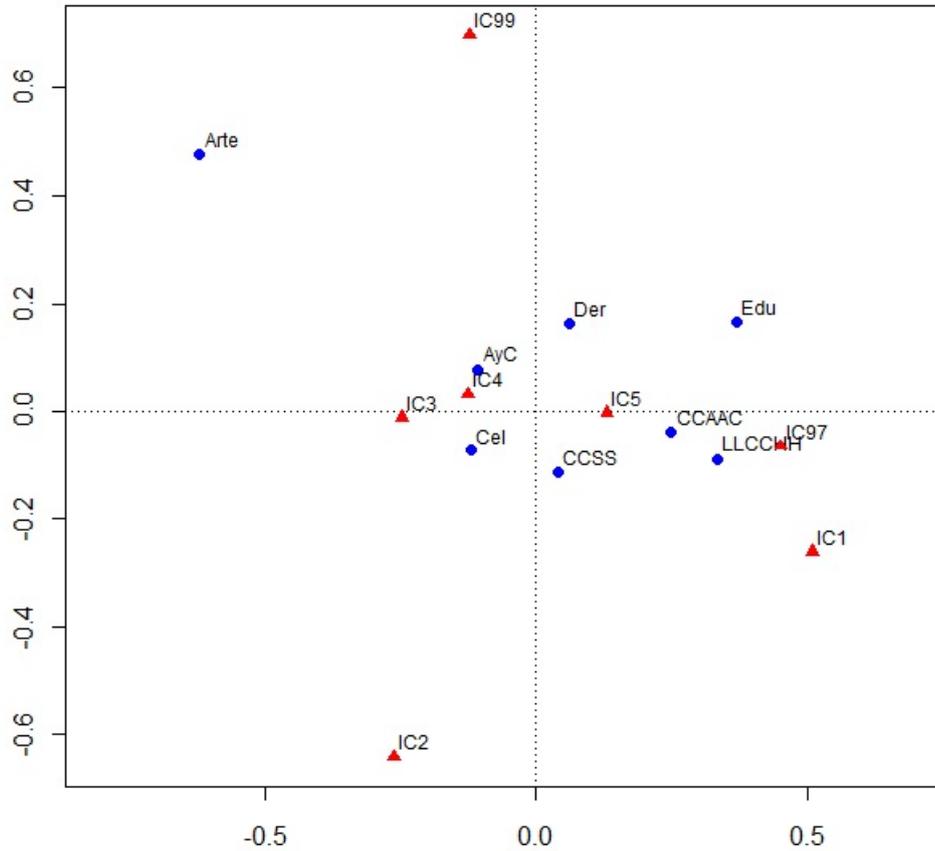


Figura B.3: Facultad del egresado Vs Grado de importancia que la Capacidad de comunicación tiene para realizar el 1er empleo

En el caso mostrado en la figura 4.5, los egresados de todas las facultades están próximos a determinar que el grado de importancia de la capacidad de comunicación es por lo menos 4. El caso particular de Arte está fuertemente influenciado por quienes no precisaron su respuesta por lo que no se podría concluir nada en especial.

La evidente brecha que hay entre lo que los egresados perciben como el grado de desarrollo de la capacidad de comunicación recibido en su formación con la gran importancia que le asignan para la realización de su 1er empleo, debe marcar una pauta en los contenidos curriculares con el fin de ir reduciendo dicha brecha. Se podría hacer el mismo análisis con el resto de capacidades indicadas al inicio de esta sección.

Para avanzar con el análisis visto en la aplicación del ACS, se planteará un análisis de correspondencia múltiple en el que se pueda visualizar la brecha correspondiente a la capacidad de comunicación para cada facultad y relacionándola además con la variable sueldo del 1er empleo para determinar en qué medida esta brecha influye en

un indicador de éxito de los egresados en el mercado laboral. Entonces se usará las siguientes variables:

- Facultad: Facultad del egresado.
- Ingreso1erEmp: Ingreso del primer empleo relacionado con su formación.
- FComunica: Grado de desarrollo de la Capacidad de comunicación recibida durante la formación.
- IComunica: Grado de importancia que la Capacidad de comunicación tiene para realizar el 1er empleo.

Entonces, tal como se indicó anteriormente, nuestra matriz de entrada para la función `mjca()` tendrá en cada fila los valores de las variables indicadas y habrá tantas filas como casos disponibles hay en el estudio.

Nuevamente, se copiará la tabla al clipboard y se ejecutará en R:

```
R > multiple <- read.table("clipboard", header = TRUE)
```

B.2.2. Aplicación del ACM

Como se sabe, hay dos formas de ejecutar el ACM. Uno es con la matriz de indicadores `Z` y la otra es con la matriz de Burt. Primero se ejecutará con la matriz de indicadores. Para esto se usará el parámetro `lambda = "indicator"`:

```
R > mjca(multiple[, 1 : 4], lambda = "indicator")
```

Y obtenemos:

Eigenvalues:

	1	2	3	4	5	6	7	...
Value	0.391187	0.326064	0.313425	0.2908	0.281035	0.275646	0.271143	...
Percentage	8.24 %	6.86 %	6.6 %	6.12 %	5.92 %	5.8 %	5.71 %	...

El resultado completo muestra 19 dimensiones. Solamente se está mostrando las primeras 7.

Columns:

	FArte	FAyC	FCCAAC	FCCSS	FCeI	FDer	FEdu	...
Mass	0.003106	0.021351	0.017663	0.014363	0.116266	0.050272	0.006793	...
ChiDist	8.916277	3.272475	3.626823	4.050359	1.072497	1.993232	5.983310	...
Inertia	0.246894	0.228649	0.232337	0.235637	0.133734	0.199728	0.243207	...
Dim. 1	0.877822	-0.069020	-1.678196	0.474442	1.087200	-1.147094	-3.140499	...
Dim. 2	-0.522830	1.784689	0.057249	0.069282	-0.191493	0.186583	-2.013187	...

En este caso todas las categorías de las variables analizadas son tomadas como columnas. Hay una columna para cada categoría. Solamente se está mostrando las primeras correspondientes a la variable Facultad.

El comando R:

R > summary(ca(table))

retorna el resumen del ACM:

Principal inertias (eigenvalues):

dim	value	%	cum %	scree plot
1	0.391187	8.2	8.2	*****
2	0.326064	6.9	15.1	*****
3	0.313425	6.6	21.7	*****
4	0.290800	6.1	27.8	*****
5	0.281035	5.9	33.7	*****
6	0.275646	5.8	39.5	*****
7	0.271143	5.7	45.2	*****
8	0.267045	5.6	50.9	*****
9	0.257602	5.4	56.3	*****
10	0.245415	5.2	61.5	*****
11	0.244044	5.1	66.6	*****
12	0.233551	4.9	71.5	*****
13	0.223567	4.7	76.2	*****
14	0.219900	4.6	80.9	*****
15	0.212713	4.5	85.3	*****
16	0.199087	4.2	89.5	*****
17	0.185158	3.9	93.4	***
18	0.162343	3.4	96.8	*
19	0.150274	3.2	100.0	
Total:		4.750000	100.0	

Se comprueba como las dos primeras dimensiones solamente representan el 15.1% de la inercia total. No se podrían obtener conclusiones muy certeras de este resultado.

Columns:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	FArte	3	5	52	549	4	1	-299	1	0
2	FAyC	21	97	48	-43	0	0	1019	97	29
3	FCCAAC	18	84	49	-1050	84	25	33	0	0
4	FCCSS	14	5	50	297	5	2	40	0	0
5	FCeI	116	412	28	680	402	68	-109	10	2
6	FDer	50	132	42	-717	130	33	107	3	1
7	FEdu	7	145	51	-1964	108	33	-1150	37	12
8	FLLCCHH	20	66	48	-800	56	16	-337	10	3
9	IC1	1	6	52	54	0	0	-1375	6	2
10	IC2	5	265	52	2160	100	31	-2779	165	54
11	IC3	33	74	46	668	68	19	-205	6	2
12	IC4	88	390	34	229	29	6	816	361	77
13	IC5	123	339	27	-435	184	29	-401	156	26
14	FC1	7	133	51	1091	33	10	-1893	100	32
15	FC2	23	202	48	1153	138	39	-789	64	19
16	FC3	79	133	36	527	129	28	-99	5	1
17	FC4	90	359	34	-173	17	3	781	342	72
18	FC5	51	459	42	-1197	364	92	-611	95	25
19	Ing1	19	148	49	-1231	128	37	-491	20	6
20	Ing2	93	63	33	-154	14	3	289	49	10
21	Ing3	94	21	33	180	20	4	45	1	0
22	Ing4	32	56	46	618	56	15	38	0	0
23	IngNP	12	180	50	125	1	0	-1882	179	56

Se puede visualizar los valores obtenidos para cada columna. Como se indicó, cada columna representa a cada categoría de las variables analizadas.

Para poder obtener el biplot del resultado ejecutamos:

```
plot(mjca(multiple[, 1 : 4], lambda = "indicator"), what = c("passive", "active"))
```

A partir de lo mostrado en la figura A.4, se puede ver como se agrupan las facultades:

- Ciencias e Ingeniería, Arte y Ciencias Sociales. Cercanas a una apreciación de 3 tanto en importancia como en percepción de desarrollo de la capacidad de comunicación en la formación recibida. Las tres además a un ingreso en el primer empleo mayor que las demás.
- Ciencias de la Comunicación, Educación y Letras y Ciencias Humanas. Cercanas a una apreciación de 5 tanto en importancia como en percepción de desarrollo de la capacidad de comunicación en la formación recibida. Las tres además con un rango menor de ingresos en su primer empleo.

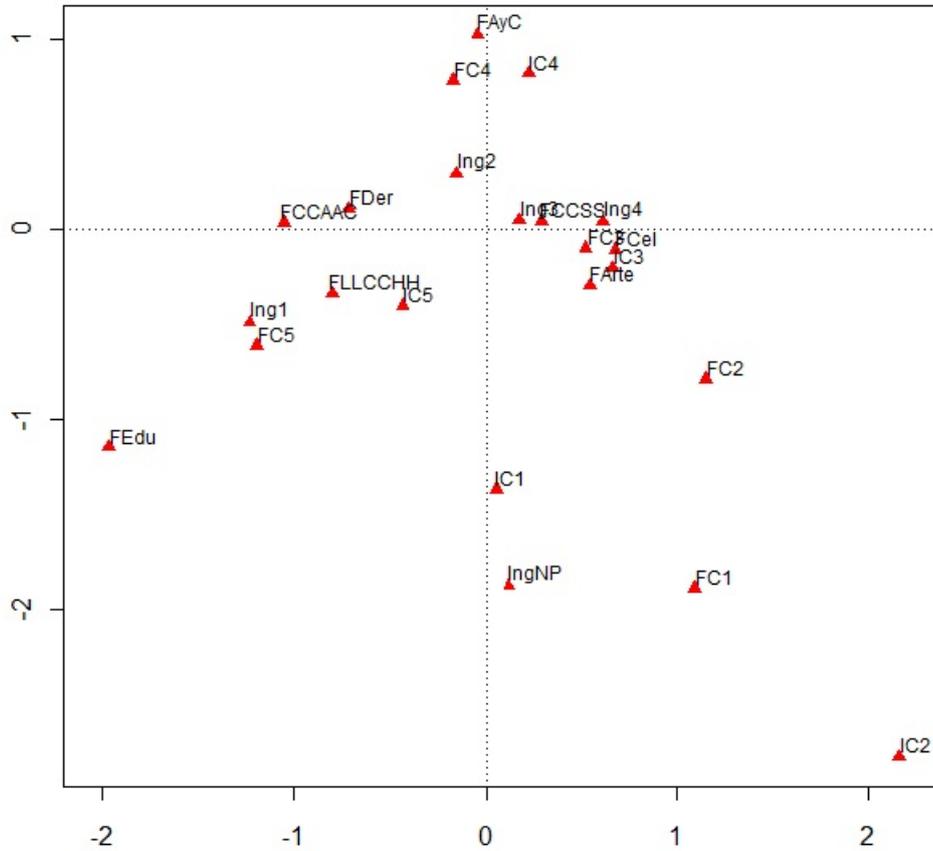


Figura B.4: ACM - Matriz indicador

- Administración y Contabilidad. Cercana a una apreciación de 4 tanto en importancia como en percepción de desarrollo de la capacidad de comunicación en la formación recibida. También con un rango menor de ingresos en su primer empleo.

Lo que se concluye es que el grado de importancia o de desarrollo de la capacidad de comunicación no influye de manera directa en el nivel de ingreso obtenido en el primer empleo. Otro punto importante en este gráfico es que no se nota la brecha que en el ACS se obtuvo entre la importancia y el desarrollo de la capacidad de comunicación. Hay que recordar el bajo porcentaje de inercia obtenido en la aplicación por lo que las conclusiones no tendrían que ser tan certeras.

En las siguientes aplicaciones, se mostrará las inercias principales de las dimensiones y los biplots correspondientes.

El segundo caso de ACM es utilizando la matriz de Burt. Se ejecuta el comando en R:

```
R > summary(mjca(multiple[, 1 : 4], lambda = "Burt"), what = c("passive", "active"))
```

Principal inertias (eigenvalues):

dim	value	%	cum %	scree plot
1	0.153027	12.3	12.3	*****
2	0.106318	8.5	20.8	*****
3	0.098235	7.9	28.6	*****
4	0.084565	6.8	35.4	*****
5	0.078981	6.3	41.7	*****
6	0.075981	6.1	47.8	*****
7	0.073518	5.9	53.7	*****
8	0.071313	5.7	59.4	*****
9	0.066359	5.3	64.7	*****
10	0.060229	4.8	69.5	*****
11	0.059558	4.8	74.3	*****
12	0.054546	4.4	78.7	*****
13	0.049982	4.0	82.7	*****
14	0.048356	3.9	86.5	*****
15	0.045247	3.6	90.2	****
16	0.039636	3.2	93.3	***
17	0.034283	2.7	96.1	**
18	0.026355	2.1	98.2 v *	
19	0.022582	1.8	100.0	
Total:	1.249071	100.0	v	

En este caso se obtiene un porcentaje de inercia del 20.8% con las dos primeras dimensiones, el cual sigue siendo muy bajo. Recordar acá que lo que aumenta la inercia es la inclusión de las matrices diagonales de la matriz de Burt. Para poder obtener el biplot del resultado se ejecuta:

```
plot(mjca(multiple[, 1 : 4], lambda = "Burt"), what = c("passive", "active"))
```

Observando el gráfico A.5, se puede ver que aunque las conclusiones son similares que en el caso anterior, el espacio esta rotado.

B.2.3. Aplicación del ACC

Se ejecuta:

```
summary(mjca(multiple[, 1 : 4], lambda = "JCA"), what = c("passive", "active"))
```

Se obtiene:

Principal inertias (eigenvalues):

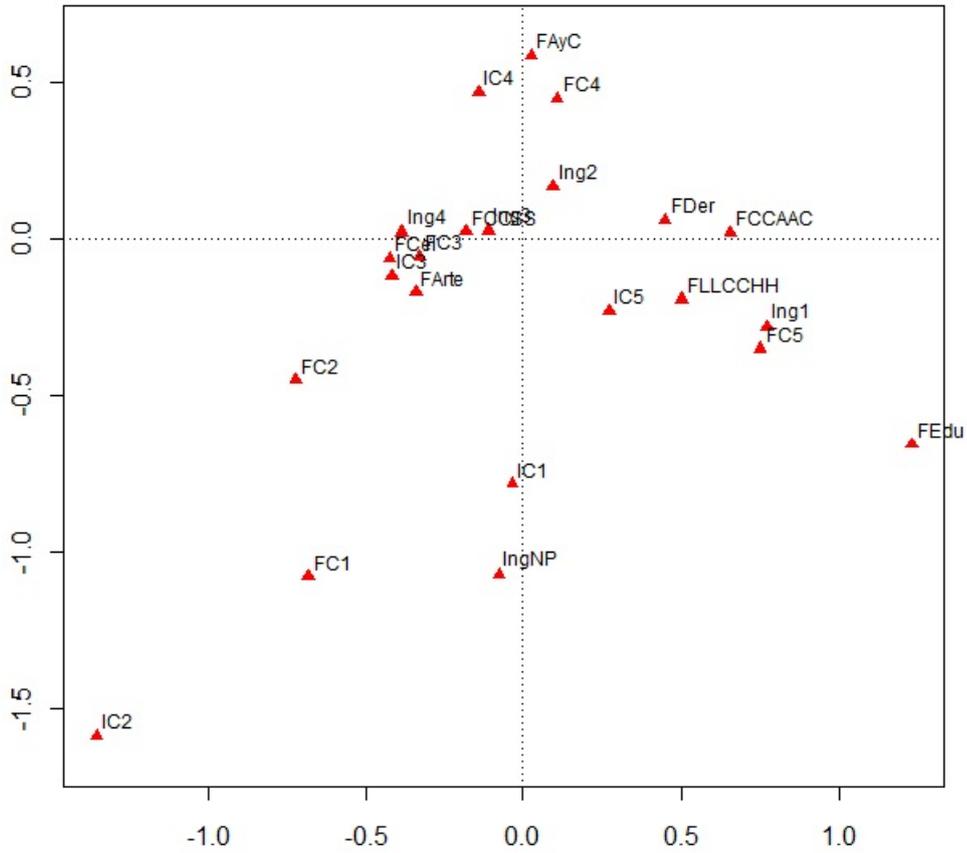


Figura B.5: ACM - Matriz de Burt

dim	value
1	0.055939
2	0.031932
3	0.006860
4	0.001920
5	0.001702
6	0.000848
7	0.000559
8	0.000352
9	0.000104
<hr/>	
Total:	0.110063

Diagonal inertia discounted from eigenvalues: 0.048492 Percentage explained by JCA in 2 dimensions: 64% (Eigenvalues are not nested) [Iterations in JCA: 50 , epsilon =

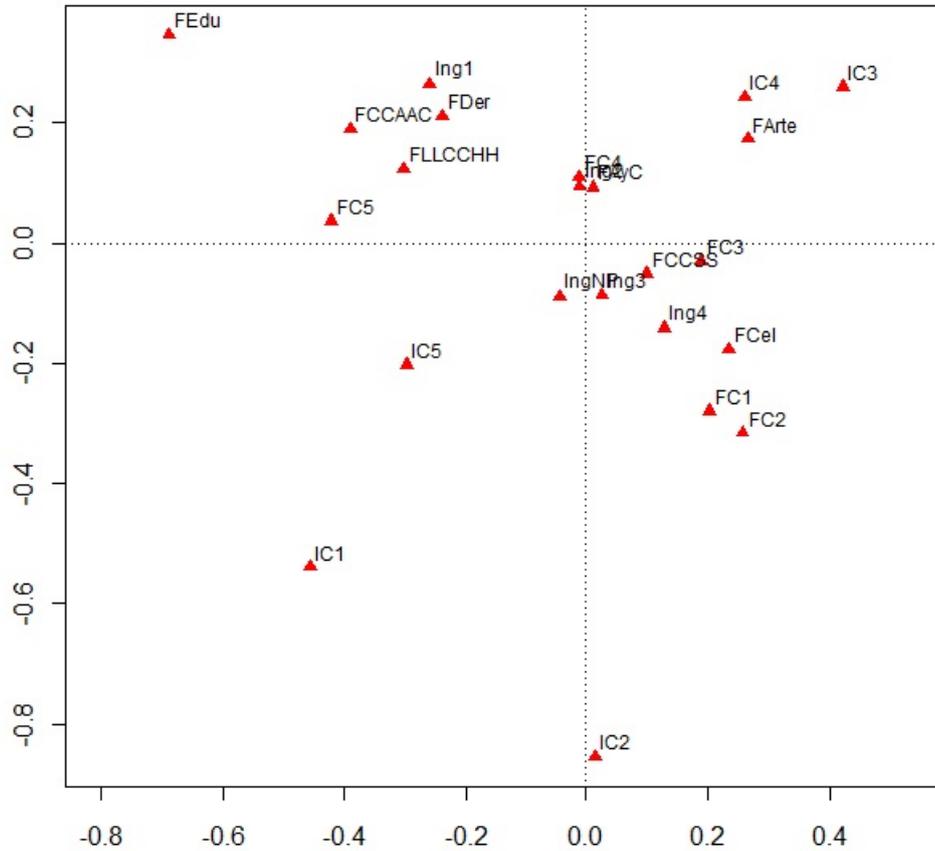


Figura B.6: ACC

2.0865462]

Notar que, tal como se indica, se descontó 0.048492 de la inercia total atribuida a las matrices de las diagonales de la matriz de Burt. La nueva inercia obtenida es de 0.110063 y las dos primeras dimensiones representan el 64% de ella. No se presentan los valores acumulados por cada dimensión ya que el resultado no es anidado. Se debería esperar obtener mejores conclusiones a partir de esta representación.

Se obtiene el biplot, con el comando:

```
plot(mjca(multiple[, 1 : 4], lambda = "JCA"), what = c("passive", "active"))
```

Lo observado en la figura A.6, es coherente con lo previsto en el ACS. Se reitera la brecha entre la importancia que se le otorga a la capacidad de comunicación y el grado de desarrollo recibido en la formación, específicamente en las facultades de Ciencias e Ingeniería y Ciencias Sociales. En menor medida se nota en las facultades de Derecho y

Administración y Contabilidad. En los casos de Ciencias y Artes de la Comunicación, y Educación sí hay correspondencia entre ambas variables como también se había notado antes. Es importante como Arte ahora aparece alejado, explicándose esto por su falta de representatividad relativa en la población.

B.2.4. Aplicación del ACMA

Se ejecuta:

```
summary(mjca(multiple[, 1 : 4]), what = c("passive", "active"))
```

Se obtiene:

Principal inertias (eigenvalues):

dim	value	%	cum %	scree plot
1	0.035438	43.2	43.2	*****
2	0.010286	12.5	55.7	*****
3	0.007152	8.7	64.4	*****
4	0.002959	3.6	68.0	**
5	0.001712	2.1	70.1	*
6	0.001169	1.4	71.5	*
7	0.000795	1.0	72.5	*
8	0.000516	0.6	73.1	
9	0.000103	0.1	73.2	
Total:		0.082095		

El porcentaje de inercia explicado por el modelo asciende a 55.7%. Notar que los valores mostrados no suman el 100%.

Se halla el biplot:

```
plot(mjca(multiple[, 1 : 4]), what = c("passive", "active"))
```

En el gráfico A.7, nuevamente se pierde la visualización de la brecha entre el grado de importancia y el grado de desarrollo recibido en la formación de la capacidad de comunicación. Nuevamente aparecen juntas las Facultades de Ciencias e Ingeniería, Arte y Ciencias Sociales que asignan un grado de importancia 3 o 4 a dicha capacidad y un grado de desarrollo de 3. Por el otro lado están Educación, Ciencias y Artes de la Comunicación y Letras y Ciencias Humanas que asignan un alto grado de importancia y de desarrollo de dicha capacidad. Y en un tercer grupo está Administración y Contabilidad que en ambos aspectos asignan un nivel de 4. Arte no presenta el alejamiento que mostraba en el ACC. Es importante resaltar algunos autores (Greenacre, 2007) afirman que el ACMA es la mejor opción dado que conserva las óptimas propiedades de escalamiento del ACM y eleva los porcentajes de inercia y el cuadrado de las correlaciones.

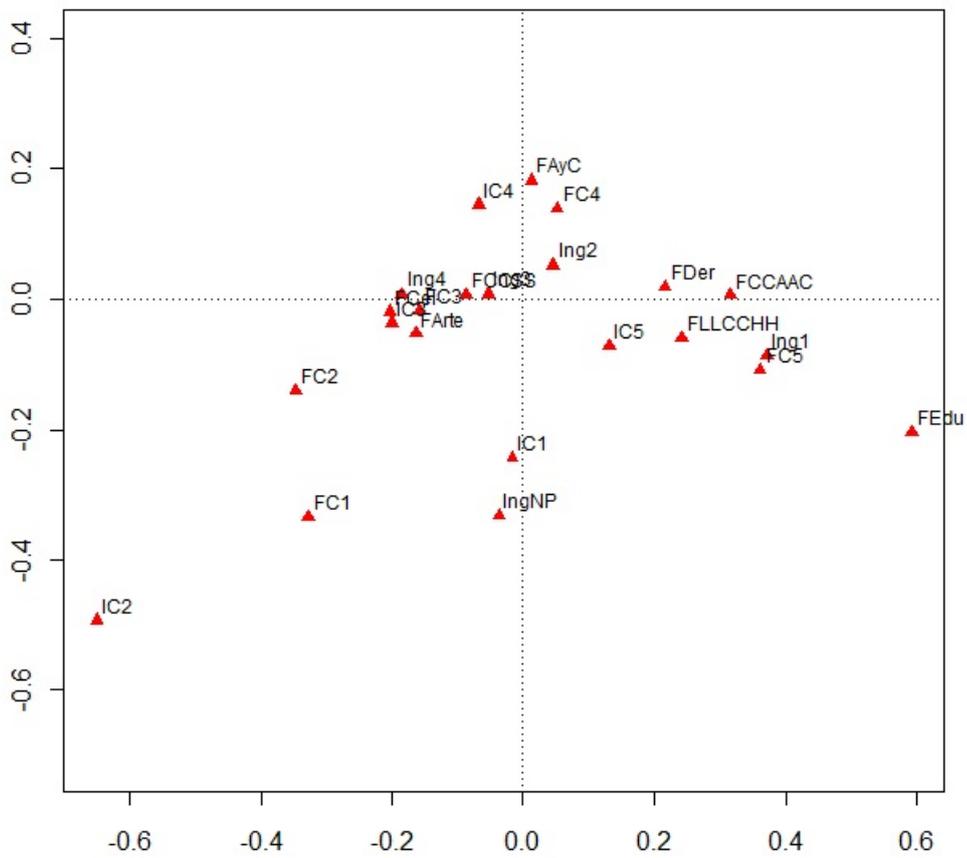


Figura B.7: ACMA

Bibliografía

- Ben Ammou, S. y Saporta, G. (1998). Sur la normalité asymptotique des valeurs propres en acm sous l'hypothèse d'indépendance des variables, *Revue de Statistique Appliquée* **46(3)**: 21–35.
- Ben Ammou, S. y Saporta, G. (2003). On the connection between the distribution of eigenvalues in multiple correspondence analysis and log-linear models, *REVSTAT-Statistical Journal* **1(0)**: 42–79.
- Benzecri, J. (1973). *L'Analyse des Données*, Vol. Tome 2: L'Analyse des Correspondances, Paris: Dunod.
- Benzecri, J. (1978). Sur l'analyse des tableaux binaires associés à une correspondance multiple, *Cahiers de l'Analyse des Données* 2 pp. 55–71.
- Camiz, S. (2010). On the problems of multiple correspondence analysis and its use.
- Camiz, S. y Gomes, G. (2012). Multiple and joint correspondence analysis: Testing the true dimension of a study.
- DAPE-PUCP (2009). *III Estudio Institucional Sobre Egresados PUCP*, PUCP.
- Gifi, A. (1981). *Nonlinear Multivariate Analysis*, Leiden: DSWO Press.
- Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*, London: Academic Press.
- Greenacre, M. J. (1988). Correspondence analysis of multivariate categorical data by weighted least squares, *Biometrika* **75**: 457–467.
- Greenacre, M. J. (2006). *From Simple to Multiple Correspondence Analysis*. In M. J. Greenacre, J. Blasius (eds.), *Multiple Correspondence Analysis and Related Methods*, Chapman and Hall/CRC, Boca Raton.
- Greenacre, M. J. (2008). *La Práctica del análisis de correspondencias*, Fundación BBVA.
- Hirschfeld, H. (1935). *A connection between correlation and contingency*, Proc. Cambridge Philosophical Society.
- Hwang, H., T. M. y Y., T. (2008). Correspondence analysis, multiple correspondence analysis and recent developments.
- Magnus, J. y Neudecker, H. (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley and Sons.
- Malinvaud, E. (1987). *Data analysis in applied socio-economic statistics with special consideration of correspondence analysis*, Marketing Science Conference, Joy en Josas: HEC-ISA.

- Nenadic, O. y Greenacre, M. (n.d.). Computation of multiple correspondence analysis, with code in r.
- Nenadic, O. y Greenacre, M. J. (2007). Analysis in r, with two- and three-dimensional graphics: The ca package, *Journal of Statistical Software* **20**.
- Orlói, L. (1978). *Multivariate Analysis in Vegetation Research*, 2nd ed.. Den Haag: Junk.
- UP y PRESTA (n.d.). *Seminario de Métodos Estadísticos Aplicados a la Investigación en Ciencias Sociales - Análisis de Correspondencias*.
- W.Hardle y L.Simar (2007). *Applied Multivariate Statistical Analysis*, Springer.

