Estadística descriptiva multivariada

Estadística descriptiva multivariada

Campo Elías Pardo

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Bogotá, D. C., Colombia, diciembre de 2020

Catalogación en la publicación Universidad Nacional de Colombia

Pardo Turriago, Campo Elías, 1956-
Estadística descriptiva multivariada / Campo Elías Pardo. – Primera edición. – Bogotá : Universidad Nacional de Colombia. Facultad de Ciencias. Coordinación de Publicaciones Facultad de Ciencias, 2023
1 CD-ROM (xv, 239, páginas) : ilustraciones en blanco y negro, diagramas. – (Colección textos)

Incluye referencias bibliográficas e índice analítico
ISBN 978-958-794-344-3 (e-book)

1. Análisis multivariante 2. Análisis cluster 3. R (Lenguaje de programación) 4. Investigación cualitativa 5. Investigación cuantitativa I. Título II. Serie

CDD-23 519.535 / 2023

Primera edición, 2020

ISBN 978-958-794-343-6 (papel)
ISBN 978-958-794-344-3 (digital)

Edición
Angélica María Olaya Murillo
Coordinación de publicaciones - Facultad de Ciencias
coorpub_fcbog@unal.edu.co

Corrección de estilo
Yecid Muñoz y Hernán Rojas

Diseño de la colección
Leonardo Fernández Suárez

Maqueta LaTeX
Camilo Cubides

Impreso y hecho en Bogotá, D. C., Colombia

A la memoria de Soledad, Luis Carlos e Israel.

A Esperanza, David –el científico– y Camilo –el mago–.

Prólogo

El texto del profesor Campo Elías Pardo aborda de forma rigurosa los análisis de componentes principales, de correspondencias simples y múltiples, y algunos de los métodos más importantes de clasificación. Estos temas tienen una vigencia indiscutible y los métodos han sido incorporados a lo que se conoce como minería de datos, donde el resumen y la clasificación cumplen un papel importante para el reconocimiento de patrones.

La exposición teórica es clara y concisa. El aporte del autor se hace presente en la forma como la resume y organiza para que su estudio sea fluido y provechoso con base en los ejemplos. El código programado para los análisis es un factor motivador para seguir el libro de manera activa, y sirve de guía para aplicar los métodos en el ejercicio profesional. Al final de cada capítulo, presenta ejemplos de aplicación resueltos y propone ejercicios y talleres muy originales que reflejan el esfuerzo por ofrecer los resultados de un trabajo personal calificado. Es un valor agregado importante del texto.

Las referencias bibliográficas incluyen un recorrido histórico del desarrollo del tema en sus avances progresivos y su lectura se hace con una perspectiva diferente a la de otros textos conocidos.

En la introducción se encuentra el público a quien va dirigido, los datos que se analizan con los métodos descriptivos multivariados, el software necesario y los requisitos conceptuales para una mejor lectura.

El capítulo 1 orienta al lector en los aspectos principales del contenido del libro, en la preparación de los temas requeridos para entender los conceptos, en la obtención de las herramientas computacionales suficientes para la práctica de los análisis, e insiste convenientemente en la contextualización como pauta metodológica para un buen estudio.

El capítulo 2 es de obligatorio seguimiento, especialmente cuando se trata de investigaciones de carácter socioeconómico. El análisis de las relaciones, focalizado en las variables claves, permite organizar una visión estructurada muy eficiente de los datos.

El tema central del capítulo 3, sobre componentes principales, es la variabilidad de los datos cuantitativos, buscando identificar sus fuentes subyacentes para presentarla de manera resumida. Mediante el ejemplo con los datos de café, ilustra el procedimiento incluyendo la preparación para calcular las distancias entre individuos y la inercia de la nube. El código en R es una invitación a participar en el desarrollo del análisis, lo que se convierte en un trabajo colaborativo muy estimulante para el lector. Los ejes principales, las proyecciones y las ayudas para la interpretación siguen el mismo protocolo, afianzando los conceptos y la capacitación para aplicarlo. Invito

al lector a disfrutar por sí mismo de la estrategia pedagógica utilizada para analizar el conjunto de variables y el engranaje posterior de los resultados con los de los individuos.

Luego de explicar el problema por resolver y los objetivos del análisis, el autor presenta en forma paralela los conceptos y los procedimientos matemáticos, las instrucciones de uso del software y un ejemplo con datos reales para llevar de la mano la teoría y la práctica.

El análisis generalizado de los componentes principales es el tema del capítulo 4, con mucho contenido matemático clásico, donde expone las bases para entender que otros métodos pueden verse como casos particulares. Extiende el uso a datos con ponderaciones no uniformes y a matrices de distancias más generales que la euclidiana clásica. De esta manera se simplifican en buena medida las exposiciones de los métodos tratados en los capítulos siguientes, para verlos como miembros de una familia con propiedades compartidas. Adoptar esta orientación de origen francés es un acierto pedagógico, pues lleva al lector a una posición desde donde puede apreciar la generalidad del tema.

El enfoque del capítulo 5, sobre el análisis de correspondencias simples, es similar al de componentes principales, incorporando el lenguaje de las tablas de contingencia: los perfiles, las distancias Ji-cuadrado, las ponderaciones, etc., pero con la perspectiva generalizada del capítulo anterior. Los conceptos se van ilustrando progresivamente con el ejemplo de los admitidos a la Facultad de Ciencias. Dos talleres, uno con datos sobre las manzanas de Bogotá y otro con los adjetivos y colores, complementan muy bien el entrenamiento para el manejo de la técnica.

Me alegra compartir con el lector la admiración que me revive el tema del capítulo 6 con respecto a dos objetos centrales del análisis de correspondencias múltiples: la tabla disyuntiva completa y la de Burt. La primera ofrece una visión "microscópica" de los individuos en el conjunto de las categorías de todas las variables cualitativas, a la manera de una tabla de chequeo. La segunda es una tabla correlativa de todas las categorías, haciendo abstracción de los individuos. Como en los capítulos anteriores, el desarrollo de la teoría se encuentra acompañado de referencias al ejemplo en lo relativo a las formas prácticas de cálculo de las expresiones que va presentando. Los ejercicios cubren un amplio rango de los procedimientos de análisis incluidos en el capítulo.

El tema de la clasificación está explicado en el capítulo 7 de manera muy intuitiva, con la ayuda de gráficas, y se lee con mucha facilidad. Las diferentes opciones de medición de distancias o similitudes entre los elementos a clasificar abarcan desde las de carácter euclidiano con las variables numéri-

cas, hasta las de asociación entre las que se consideran categóricas. A partir de los resultados de los procedimientos factoriales de los primeros capítulos, abre la posibilidad de utilizar la información resumida en los primeros ejes para identificar y describir los patrones de comportamiento de los datos, eliminando información secundaria, muchas veces ruidosa, contenida en los ejes descartados.

Al final de la lectura, tengo la seguridad de que el lector queda equipado con un conjunto de conocimientos y herramientas que le permiten abordar análisis de información compleja de cuestionarios muy diversos.

Jorge Ortiz Pinilla, Ph.D.
Profesor Pensionado
Universidad Nacional de Colombia - Sede Bogotá
Facultad de Ciencias
Departamento de Estadística

Quiero agradecer a la Universidad Nacional de Colombia, la Facultad de Ciencias y el Departamento de Estadística. A los decanos y vicedecanos de investigación de la Facultad, a los Directores del Departamento, y sus respectivos equipos de publicaciones; de los últimos años, por el apoyo continuado para la publicación de este libro.

Al colega Jimmy Corzo por su concepto, solicitado por el Departamento de Estadística, y sus aportes para mejorar texto. A los profesores, que aceptaron la solicitud de la Facultad de Ciencias para realizar la evaluación académica del texto: Sergio Jorge Bramardi de la Universidad Nacional de Comahue (Argentina) y Jorge Ortiz, quienes hicieron valiosos aportes para mejorarlo. Al profesor Jorge Ortiz, además, la lectura detallada del texto, sus recomendaciones para mejorarlo y el prólogo.

A Angélica María Olaya por sus contribuciones como Editora; a los correctores de estilo Yecid Muñoz y Hernán Rojas; y al diseñador Leonardo Fernández.

A Camilo Cubides por las modificaciones de los archivos fuente del texto para ajustarlos a la plantilla, que diseñó y programó para los libros de la Facultad de Ciencias editados en LaTeX, y por su disposición para la solución de los problemas técnicos.

A mis estudiantes de los cursos de Estadística descriptiva multivariada y a los de los cursos de Análisis Multivariado de Datos, Teoría Estadística Multivariada y Análisis Multivariado Aplicado, de los posgrados en Estadística; por su contribución a la detección de errores del texto de diferente tipo, en sus versiones previas.

Campo Elías Pardo
Profesor Asociado
Universidad Nacional de Colombia - Sede Bogotá
Facultad de Ciencias
Departamento de Estadística

Contenido

Lista de figuras

Lista de tablas

Introducción

Este texto se origina en las notas del curso de Estadística Descriptiva Multivariada de la Carrera de Estadística de la Universidad Nacional de Colombia. También sirve, como referencia para los cursos de análisis de datos multivariados en pregrados y posgrados en Estadística y como libro de consulta para los profesionales de distintas áreas interesados en abordar la descripción y exploración de tablas de datos en sus investigaciones y en su ejercicio profesional.

El objeto de entrada a los métodos estadísticos que se abordan en este texto es una tabla de datos que refleja parcialmente una realidad que se quiere estudiar. Algunas veces los datos son el resultado de un proceso metodológico largo y costoso: concepción de una investigación, definición de variables, diseño de los instrumentos de medición, captura y depuración de los datos, entre otros. Las investigaciones basadas en encuestas son un ejemplo. Otras veces los datos provienen de sistemas de información administrativos o de transacciones (bancarias, de servicios públicos, de grandes superficies de ventas al público, etc.); pero la tabla objeto análisis depende de un proceso metodológico de selección, depuración, concatenación y transformación, y casi siempre, de búsqueda de nuevos datos.

Una tabla de datos básica es un archivo que tiene en filas las unidades estadísticas, que denominaremos "individuos", y en columnas las variables, en general de diferentes escalas de medición: nominal, ordinal, de intervalo, y de razón. Los tipos de variables que se originan con estas escalas se agrupan, para este documento, en dos conjuntos: cualitativas (de escala nominal u ordinal) y continuas (de intervalo o de razón).

Nos situamos en el caso en que algunos de los objetivos del estudio se cumplen realizando análisis descriptivos y exploratorios multivariados de la tabla de datos, que utilizan representaciones gráficas de comprensión más fácil para el cerebro humano. Las descripciones univariadas dependen de las escalas de medición de las variables y ayudan a completar la depuración de los datos, orientar las transformaciones de algunas variables y a tomar decisiones sobre la imputación o no de datos faltantes.

Algunas veces se realizan descripciones bivariadas, según los tipos de las dos variables: ambas continuas, continua y cualitativa, y ambas cualitativas. Las descripciones multivariadas permiten tener en cuenta las relaciones entre variables.

Las descripciones multivariadas que recurren a las gráficas para comprender los datos son mucho más difíciles que las univariadas, porque su interpretación correcta depende del conocimiento de los procedimientos y conceptos para su construcción. Los usuarios de diferentes áreas del conocimiento necesitan al menos una comprensión intuitiva de la lógica de los métodos con el fin de lograr la interpretación correcta de las salidas gráficas y de los índices numéricos que las acompañan. Los científicos y profesionales responsables de la metodología estadística deben conocer los fundamentos de la geometría multidimensional, basados en los conceptos del álgebra lineal que tienen que ver con espacios vectoriales en los reales con producto interno.

Los métodos descriptivos y exploratorios multivariados pretenden encontrar significado en grandes tablas de datos, luego de transformaciones adecuadas según el método, en otras tablas de n filas y p columnas, como dos nubes de puntos: las filas como n vectores en $\mathbb{R}^p$ y las columnas como p vectores en $\mathbb{R}^n$. Estas representaciones permiten obtener gráficas para descubrir el contenido, que se encuentra oculto, dentro de la gran cantidad de cifras de una tabla (Lebart *et al.*, 2006).

En este texto se muestran los principales métodos en ejes principales como aplicación de la geometría euclidiana y del álgebra lineal, que constituye su lenguaje matemático. La simbología que se adopta usual en muchos textos es la siguiente: las letras mayúsculas en negrilla hacen referencia a matrices (**A**), la minúsculas en negrilla, a vectores (**a**); las letras mayúsculas y minúsculas en itálica a, variables (escalares) (A, a). En el caso de conjuntos se utiliza la misma letra mayúscula para indicar el conjunto y su cardinalidad (número de elementos).

El primer capítulo, denominado "Preliminares", se ocupa de mostrar, entre otros, el panorama de los métodos abordados en el curso, el lenguaje estadístico R como "calculadora" gráfica y de álgebra lineal, los principales paquetes de R a utilizar en el curso, y un breve repaso de Estadística descriptiva univariada.

El lector de este texto debe instalar el R (R Core Team, 2020) en su computador y leer el manual *An introduction to R* (Venables *et al.*, 2017), disponible en la consola de R una vez instalado. Además, debe instalar el paquete `FactoClass` (Pardo & Del-Campo, 2007), que complementa este texto: tiene la mayoría de los datos que aquí se utilizan. `FactoClass` carga los

paquetes: `ade4` (Dray & Dufour, 2007), utilizado para realizar los cálculos de los métodos estudiados; `scatterplot3d` (Ligges & Mächler, 2003), para construir gráficas 3D, y `xtable` (Dahl, 2016), para exportar tablas a LaTeX en el entorno *tabular*. Estas notas están editadas en LaTeX (The-LaTeX-Project-Team, 2019). A partir de la versión 1.2.1 de `FactoClass` se cargan, también, los paquetes `ggplot2` (Wickham, 2009) y `ggrepel` (Slowikowski, 2020) para obtener planos factoriales en los que las etiquetas no se sobrepongan.

Para la edición de las gráficas se utiliza el programa de uso libre *xfig* (Sato & Smith, 2018) ya que R permite exportar a ese formato y a su vez *xfig* exporta a los formatos de gráficas más conocidos. Con *xfig* se editan los planos factoriales, para destapar etiquetas que quedan superpuestas y agregar otras o modificar otros elementos de las gráficas. Como complemento y referencia para la ejecución de los métodos se utiliza el programa *DtmVic* (Lebart, 2017) de uso libre académico.

En el curso de Estadística Descriptiva Multivariada, de dieciséis semanas, se sugiere abordar cada uno de los siete capítulos en dos semanas y dejar dos semanas para las presentaciones de los trabajos del curso.

Capítulo
uno

Preliminares

En este capítulo se hace una presentación de los métodos abordados en el texto, el lenguaje estadístico R (R Core Team, 2020) y otros elementos necesarios para estudiar la lógica de los métodos y su aplicación. También se introduce el análisis univariado y la obtención de nuevas variables agrupando categorías en una variable cualitativa y dividiendo en clases a los individuos a partir de una variable continua. Esto se piensa como preparación para análisis multivariados posteriores.

1.1. Introducción a los métodos

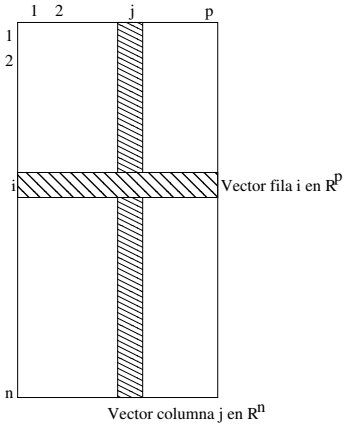
Los métodos de la estadística descriptiva multivariada, abordados en este texto, son de dos tipos: factoriales o en ejes principales y de clasificación o agrupamiento. En ambos casos se hace una representación geométrica de las tablas de datos, transformadas según el método específico.

Para simplificar la introducción a los métodos, pensemos en una tabla de datos de n filas, que denominaremos "individuos", y p variables, que tomaremos como medidas continuas. Una tabla de estas es una matriz numérica de n filas y p columnas y tiene dos representaciones geométricas: 1) n vectores fila en $\mathbb{R}^p$ y 2) p vectores columna en $\mathbb{R}^n$.

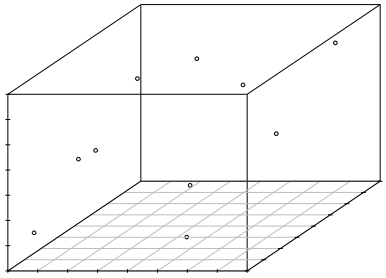
En el primer caso la representación geométrica es un conjunto de n puntos en un sistema de p ejes ortogonales, cada eje asociado a una variable, y las coordenadas de un individuo son los p valores que toma para las variables (figura 1.1). Esta representación es abstracta, pero tiene las mismas propiedades de las representaciones en dos dimensiones, que se denominan *diagramas de dispersión*.

Dos individuos están cercanos en $\mathbb{R}^p$ si tienen más o menos las mismas coordenadas, es decir, valores similares para las p variables. Entonces, la representación geométrica es útil para comparar a los individuos entre sí y observar la estructura de la "nube de individuos" –es decir, los n puntos– en el sentido de observar su forma y detectar patrones que se pueden manifestar como grupos e individuos (figura 1.1).

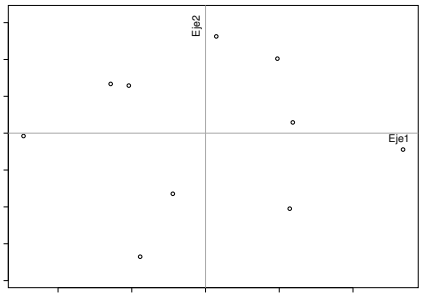
Para detectar patrones se dispone de dos tipos de métodos: 1) los factoriales o en ejes principales, que buscan los mejores ejes y planos de proyección, para observar las nubes de puntos de forma aproximada, y 2) los de agrupamiento o clasificación, que conforman grupos en los datos, buscando que los individuos se parezcan lo más posible cuando pertenecen a un mismo grupo y se diferencien al máximo cuando pertenezcan a diferentes grupos.



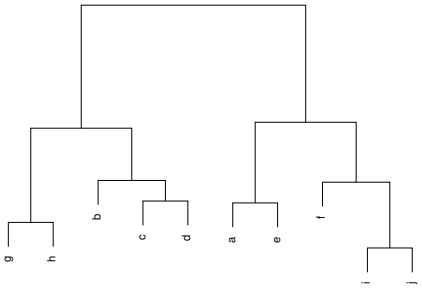
Esquema de la tabla



Esquema de representación en $\mathbb{R}^3$



Proyección de vectores fila en $\mathbb{R}^2$



Árbol de clasificación jerárquica

Figura 1.1. Esquema de una tabla de datos y de los métodos

El conjunto de los p vectores en $\mathbb{R}^n$ se denomina "nube de variables" y se puede pensar que su representación son flechas que empiezan en el origen de los n ejes y terminan en el punto cuyas coordenadas son los n valores que toma la respectiva variable. Cada uno de los n ejes se asocia a un individuo. La longitud de las flechas de los vectores variables representan sus desviaciones estándar y el ángulo entre parejas de ellos, su correlación. Mediante proyecciones sobre planos se pueden ver de manera conjunta las relaciones entre todas las variables.

Los métodos en ejes principales o factoriales, abordados en este texto, se muestran esquemáticamente en la figura 1.2. Pero solo se introducen en el capítulo 3, donde se presenta el análisis en componentes principales (ACP), empleado en la exploración de tablas de individuos descritos por variables continuas.

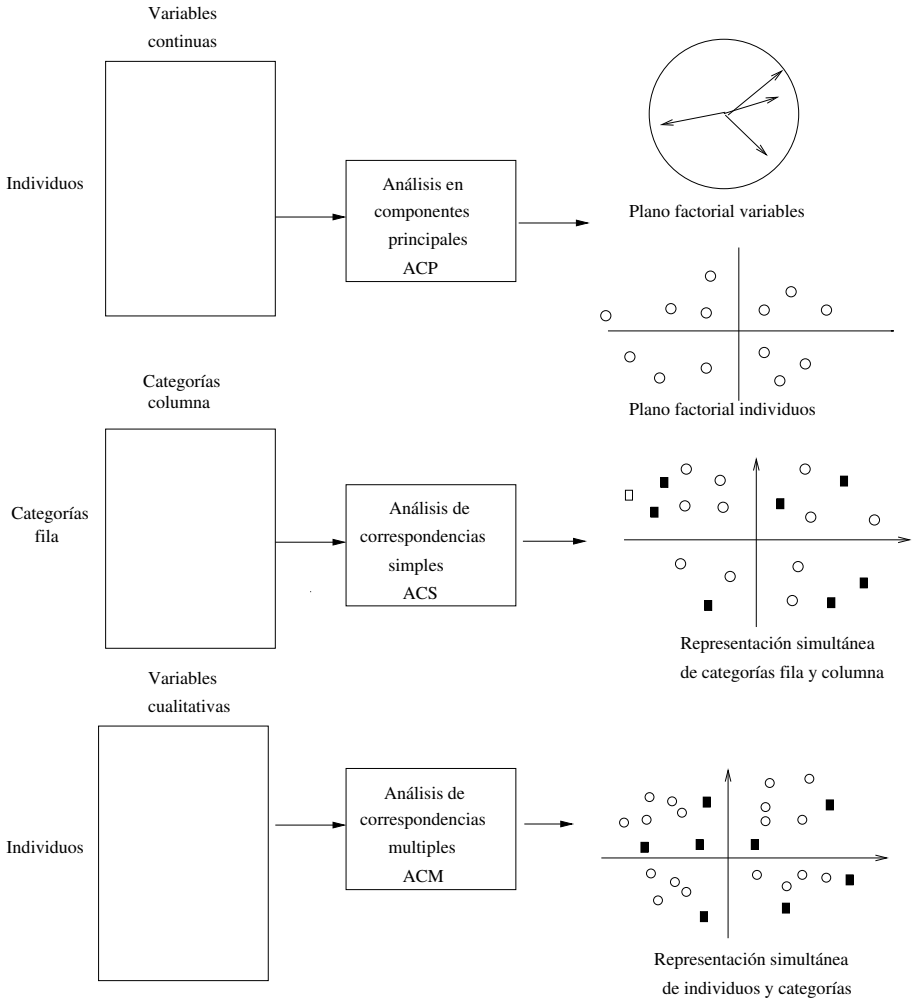


Figura 1.2. Esquema de tres métodos factoriales

En el capítulo 4 se presenta el ACP generalizado o ponderado, denominado también *análisis factorial general*, que es el marco de referencia común para este tipo de métodos. Este capítulo muestra, también la manera de obtener gráficas para las matrices de varianzas y de correlaciones, cuando no se dispone de los datos con los que fueron calculadas; así como el método de análisis en coordenadas principales (ACO), utilizado para obtener imágenes euclidianas para las matrices de distancias entre individuos. El ACO forma parte del los métodos de escalamiento muldimensional, que asocian imágenes geométricas a matrices de similitudes o disimilitudes.

En el capítulo 5 se presenta el análisis de correspondencias simples (ACS), utilizado para la descripción de tablas de contingencia, que dan el número de individuos, en cada una de las clases determinadas por el cruce de las categorías de dos variables cualitativas. El ACS se presenta como dos ACP generalizados, uno de perfiles fila y otro de perfiles columna, punto de vista que es útil para el análisis de los resultados. También se presenta como un ACP generalizado que sirve para la implementación de los cálculos y para derivar las relaciones entre los espacios de filas y columnas.

El capítulo 6 se dedica al análisis de correspondencias múltiples (ACM), usado en la descripción de tablas de individuos por variables cualitativas. El ACM se muestra como una generalización del ACS: análisis de correspondencias de la tabla indicadora de las particiones asociadas a las variables nominales, tabla que tiene n individuos en las filas por p categorías en las columnas.

Los métodos de clasificación automática se presentan en el capítulo 7 y corresponden a la búsqueda de grupos de las nubes de individuos mediante dos tipos de algoritmos, que se combinan entre sí y con los métodos factoriales, para completar la descripción multivariada básica de una tabla de datos.

1.2. El lenguaje estadístico R

El lenguaje estadístico R permite, en primer lugar, hacer los cálculos del álgebra lineal, parte de las matemáticas fundamental en la construcción de la Estadística; en segundo lugar realizar gráficas que están siempre presentes en dicha disciplina, y en tercer lugar llevar a cabo sus métodos específicos. El R es también un lenguaje de comunicación porque podemos utilizar su código para escribir fórmulas matemáticas en forma plana y para identificar métodos estadísticos. Los cálculos y las gráficas presentes en estas notas se obtienen en R, y se presentan en el texto algunas partes del código, primero con el objetivo de ayudar a entender los cálculos de álgebra lineal y luego para ejecutar los métodos con funciones específicas de R.

1.2.1. Obtención e instalación de R

Con el objetivo de que este documento tenga más vida se incluyen instrucciones en R para realizar los cálculos matriciales y obtener las gráficas, ya que todo lector puede instalarlo en su computadora y reproducir lo que se presenta en este texto.

Por economía expresiva en la secuencia de instrucciones de esta sección se utiliza el símbolo ⟶ para indicar que se debe hacer clic en la palabra o frase que aparece a continuación de la flecha en el menú. Obviamente las versiones de R y los paquetes de este documento corresponden a la fecha de su edición y se podrán encontrar otras cuando se esté leyendo. Este documento da un camino posible en cada caso y supone que el lector utiliza el sistema operativo *Windows* y que dispone de una conexión a internet de *banda ancha*.

- Entre a la página `http://www.r-project.org`
- Clic en CRAN ⟶ *Mirror* y escoja uno.
- Clic en *Download R for Windows* ⟶ *base* ⟶ *Download R 3.6.3 for Windows (83 megabytes, 32/64 bit)*

  Los usuarios de Linux y de OS X pueden seguir las instrucciones de la página web de R.
- Guarde el archivo en un directorio ⟶ *Ejecutar*.
- Responda a las preguntas del instalador (aceptando las sugerencias).
- Para ejecutar R se hace clic en el acceso directo R. Aparece la consola de R con una barra de menú en la parte superior. R espera comandos, algunos de los cuales se pueden producir con el menú.
- Para leer el manual de introducción: clic en *Help* ⟶ *An Introduction to R*.

Es muy conveniente, para quien empiece con R, leer *An introduction to R* (Venables *et al.*, 2017), ya que redundará en el mejor uso del lenguaje y a la postre en ahorro de tiempo. Existe una versión en español, aunque es más antigua (González & González, 2000).

R diferencia mayúsculas y minúsculas, <- indica asignación (también se puede utilizar =), -> asignación a la derecha, y el símbolo # se utiliza como comentario, es decir, el texto que aparece a la derecha de # se presenta pero R no lo interpreta.

R utiliza como separadores de comandos tanto el espacio como el punto y coma (;). Por lo tanto, no se requiere poner punto y coma al final de la linea. Sin embargo, en este texto aparece a veces el punto y coma (;) al final de una linea de comando para hacer que funcione cuando el lector la copia en el visor de PDF y lo pega en la consola de R o de RStudio.

El lenguaje R está enmarcado dentro de la programación orientada a objetos. Se recomienda leer sobre estos conceptos en el documento de introducción a R. Todo procedimiento básico, gráfico o estadístico se hace utilizando una función determinada, la cual recibe "parámetros" y entrega un objeto de salida. Las funciones están en librerías denominadas *paquetes*. Las principales quedan disponibles al instalar R, pero otras están en paquetes, desarrollados por investigadores alrededor de todo el mundo, que se pueden instalar. Se utiliza aquí, tal como lo hace la documentación de R, la costumbre de poner entre corchetes, { }, el nombre de la librería o paquete donde está la función. Por ejemplo, `plotct{FactoClass}` indica que la función `plotct` está en el paquete `FactoClass`.

1.2.2. Instalación de paquetes

Para ejecutar los métodos multivaridos de este documento, se recomienda instalar en R el paquete `FactoClass`, procedimiento que además instala: `ade4`, `ggplot2`, `ggrepel`, `xtable` y `scatterplot3d`. Para hacerlo se ejecuta el comando:

```
install.packages("FactoClass").
```

Otra forma de instalar uno o varios paquetes es con la barra de control de la consola:

- Clic en *Paquetes* (en la barra de menú situada en la parte superior) ⟶ *Instalar paquetes*, aparece lista de *Mirrors*.
- Seleccione un *Mirror* ⟶ *OK*, aparece lista de paquetes.
- Selecciones los paquetes, con la tecla *Control* oprimida cuando se quieren varios, por ejemplo: `FactoClass` `ade4`, `ade4TkGUI`, `scatterplot3d`, `ggplot2`, `ggrepel`, `KernSmooth` y `factoextra` ⟶ *OK*.

`FactoClass` utiliza algunas funciones de `KernSmooth` (Wand, 2020). El paquete `factoextra` (Kassambara & Mundt, 2020) toma objetos de salida de `ade4` (Dray & Dufour, 2007), para hacer gráficos, por ejemplo planos factoriales, con `ggplot2` (Wickham, 2009) y `ggrepel` (Slowikowski, 2020). Estos dos últimos paquetes son una implementación en R de una gramática para gráficas (Wilkinson, 2006).

1.2.3. RStudio, Sweave y Markdown

RStudio es el primer IDE (*Integrated Development Environment*) para R (RStudio Team, 2015) y facilita la documentación de los trabajos realizados con R, integrando Sweave y R Markdown.

Sweave (Leisch & R-core, 2020) es una herramienta que permite integrar el código de R para llevar a cabo el análisis de datos dentro de documentos LaTeX (The-LaTeX-Project-Team, 2019). "LaTeX es *de facto* el estándar para la comunicación y publicación de documentos científicos". Este documento se ha editado en LaTeX. Morales (2006) escribió un manual en español para el uso de Sweave.

"*R Markdown* es un formato de edición que permite la creación fácil de documentos dinámicos, presentaciones y reportes desde R".

El uso de estas herramientas permite documentar los trabajos a medida que se van realizando, de modo que se facilita la elaboración del reporte escrito de un trabajo de análisis de datos y la elaboración del material de apoyo para su presentación en público.

1.3. El programa *DtmVic*

El profesor Lebart ha puesto los programas básicos desarrollados en Francia, programados en Fortran, en un entorno que facilita su uso y que ha denominado *DtmVic* (Lebart, 2017). Está disponible bajo *Windows* y es de uso libre para propósitos académicos y de investigación. *DtmVic* se utiliza como programa de referencia para el texto. Se encuentra en página `http://www.dtmvic.com`, se descarga el archivo `dtm_software.zip`, se descomprime la carpeta y se ejecuta haciendo clic en `DtmVic_6.2.exe`. Por otro lado se descarga el archivo de ejemplos: `DtmVic_Examples`.

1.4. Editor para gráficas obtenidas con R

Hay varias posibilidades para editar las gráficas obtenidas con R, pero una buena opción es utilizar el programa *Xfig* (Sato & Smith, 2018), con licencia libre para Linux, que se puede instalar en los computadores Mac, a partir del código fuente.

Xfig es un editor vectorial y, por lo tanto, la calidad de la gráfica se conserva con los cambios de tamaño. Las gráficas de R se pueden exportar directamente a este formato (`.fig`) y desde *Xfig* a casi cualquier formato gráfico. Una manera de hacerlo, es utilizando el comando

`dev.print(device = xfig)`, con el que se graba la gráfica activa en el archivo `Rplot001.fig`, en la carpeta de trabajo.

Este es el editor que se utiliza para las gráficas de este texto, sobre todo para destapar las etiquetas superpuestas en los planos factoriales.

1.5. Conceptos de álgebra lineal

Para repasar los conceptos de álgebra lineal, requeridos para este curso se recomienda un capítulo o anexo de un texto de análisis multivariado de datos, por ejemplo:

- El anexo A de Díaz (2007).
- El capítulo dos de Morrison (1990): *Matrix algebra* .
- El capítulo dos de Hardle & Simar (2007): *A short excursion into matrix algebra*.

1.6. Entorno de una tabla de datos

El objeto que entra a los métodos de estadística descriptiva multivariada es una tabla de datos. Esta se constituye en un producto intermedio dentro de un proyecto de investigación y puede tener distintos orígenes. La tabla por sí sola no es objeto ningún interés de análisis si no forma parte de un contexto investigativo. En algunos casos llegar a ella puede costar el 80 % del presupuesto de una investigación. El análisis de la tabla de datos está orientado por los objetivos y el referido contexto.

El usuario de estos métodos debe situarse dentro del contexto de la investigación de la que la tabla de datos forma parte. Debe por tanto poner en práctica los procedimientos de la metodología de la investigación científica.

Lo que nos ocupa es describir o explorar alguna realidad para conocer un poco más de ella. Cualquier realidad, que se quiera abordar, es compleja y no es posible entenderla en su totalidad. Tenemos que aceptar que lo que observemos será casi siempre parcial y lo que describamos dependerá de los objetivos planteados. Es fundamental entender el contexto de la realidad a estudiar y plantear de manera clara los objetivos deseados.

La información obtenida en una investigación se almacena en una base de datos, acompañada de documentos que dan cuenta del contexto y del

procedimiento que se siguió para llegar a los datos allí guardados. La información sobre los datos (metadatos) puede estar, una parte, en la misma base de datos y otra, en documentos anexos.

De los objetivos de un estudio se derivan los objetivos de análisis, y para cumplirlos se obtienen una o más tablas de datos, que luego se describen con uno o más métodos de los abordados en este texto. Las decisiones que hay que tomar, en el sentido de las técnicas estadísticas a usar y sus aspectos internos, dependen del conocimiento de estas y del contexto en el que se enmarca la tabla de datos.

Esto forma parte de la metodología de la investigación, que tiene que abordar todo profesional, pero que para el estadístico es central, porque los métodos de análisis de datos en general forman también parte de esta metodología.

Las competencias en metodología de la investigación solo se mejoran en la práctica. Sin embargo, en la literatura existen muchos textos guía para ir mejorando esas competencias; por ejemplo, el de Hernández *et al.* (2006) o el de Briones (1996), disponible en la web. En este curso se hace un trabajo con el propósito de mejorar las competencias tanto en metodología como en el uso apropiado de los métodos básicos de la estadística descriptiva multivariada.

1.7. Preparación de los datos para el análisis

Esta sección se incluye para ilustrar algunas transformaciones de variables, que se hacen para mejorar los análisis bivariados y multivariados. Las transformaciones se registran como nuevas variables, para no perder la información original. Entonces, en un análisis estadístico particular, se puede usar la versión original o la transformada.

En la versión descriptiva de los métodos multivariados las variables juegan dos papeles complementarios: se denominan *activas* las que se seleccionan para la construcción de ejes factoriales y clases e, *ilustrativas* las que juegan el papel de explicar o ilustrar los ejes o clases obtenidos. El analista de datos debe prestar atención especial en la preparación de las variables activas para el análisis multivariado porque son estas las que estructuran ejes y clases. Por ejemplo, una variable cualitativa con muchas categorías, comparada con las demás, influirá más en un análisis de correspondencias múltiples.

La fuente original de datos puede ser una base de datos o un archivo con muchas columnas, de donde se debe obtener una tabla para el análisis es-

pecífico que se desea realizar. Como ejemplo para esta y otras secciones del texto se utiliza parte de una consulta del Sistema de Información Académico de la Universidad Nacional de Colombia (SIA), con los admitidos a las carreras de la Facultad de Ciencias, para el primer semestre de 2013.

Ejemplo: "Admitidos a la Facultad de Ciencias"

La Universidad Nacional de Colombia selecciona, a los estudiantes que se admiten en cada semestre mediante la aplicación de un examen de admisión estructurado en cinco áreas: matemáticas, ciencias, sociales, textual e imágenes. La Universidad presenta los resultados de admisión de todos los aspirantes estandarizados con media 10 y desviación estándar 1. El resultado global del examen se estandariza con media 500 y desviación estándar 100.

Para este ejemplo se toman los resultados de los 445 admitidos a las siete carreras de la Facultad de Ciencias –Biología, Estadística, Farmacia, Física, Geología, Matemáticas y Química– del primer semestre de 2013. Este grupo es un conjunto pequeño de todos los que se presentaron. Los resultados de este grupo son mejores que el promedio de todos los que presentaron el examen, porque los admitidos son los de mejores puntajes. Por la misma razón la dispersión disminuye, y el hecho de ser admitidos a las carreras de ciencias, también los puede hacer más homogéneos.

La hoja de datos retenida para el ejemplo tiene en las columnas: la carrera, los resultados del examen en cada área y global, algunas variables sociodemográficas y dos columnas que indican si el admitido debe nivelar (lectoescritura o matemáticas). Como variables sociodemográficas se incluye el género, el estrato socioeconómico, el origen geográfico del estudiante y la edad. A partir de los resultados del examen de admisión algunos estudiantes deben hacer cursos de nivelación en lectoescritura y matemáticas básicas.

La hoja de datos se incluye en el objeto `admi{FactoClass}`, que tiene 445 admitidos y quince columnas. Los nombres abreviados de las variables cualitativas y sus categorías se muestran en las tortas de la figura 1.3, y los de las variables continuas con sus histogramas, en la figura 1.4. Las variables estrato socioenónimo (*estr*) y origen geográfico (*orig*) se recodificaron con menos categorías que las originales y la edad se convirtió en ordinal con cuatro clases. La columna 14, *stra*, tiene el estrato original y la columna 15, *age*, la edad en años, como estaban en los datos originales. La variable *stra* se utiliza para análisis univariados —por ejemplo, el diagrama de barras que muestra su distribución de frecuencias (figura 1.3)— y la variable *estr*, en los análisis bivariados y multivariados.

Para obtener la figura 1.3

```
library(FactoClass) # cargar FactoClass;
data(admi)  # cargar la tabla;
# 6 barplots y etiquetas de las categorias en forma horizontal
par(las=1,mfrow=c(2,3),mai=c(0.6,0.5,0.2,0.1))
for (i in c(1,8,10,14,12,13)){
cat<-attributes(admi[,i])$levels;
  per <- tabulate ( admi[,i])/445*100;
  pl<-plot(admi[,i],horiz=TRUE,col=gray(seq(1.0,0.9,
           length= length(cat))),ylim=c(0,8),
           xlim=c(0,400),xlab=colnames(admi)[i]);
  text(160,pl,round(per,1),cex=0.8,pos=4);
}
# manualmente cuadrar la ventana Plots
# dev.print(device=pdf) # grabar la grafica como Rplots.pdf
```

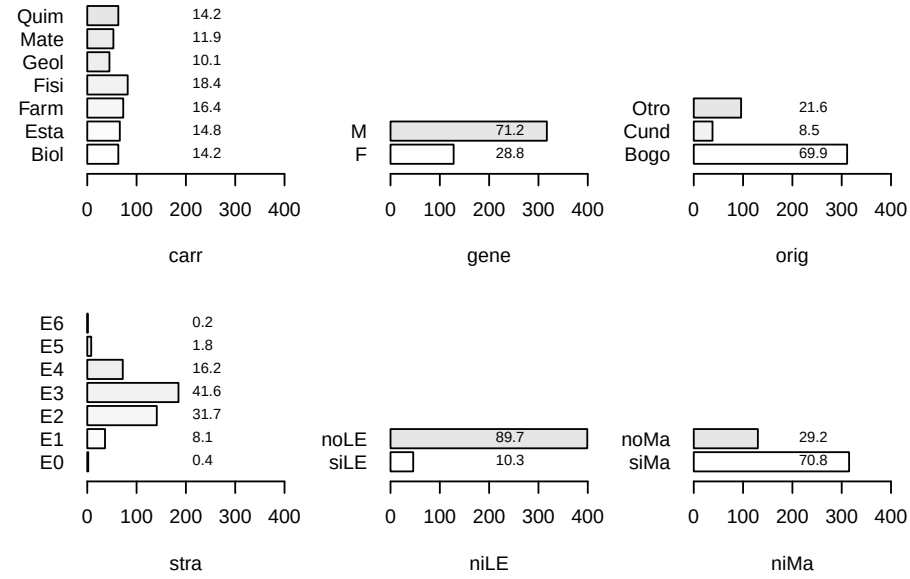


Figura 1.3. Diagramas de barras mostrando la distribución de las categorías de las variables cualitativas de los admitidos a la Facultad de Ciencias. Los números dentro en las barras son porcentajes

1.7.1. Transformación de variables cualitativas

Se pueden lograr mejores descripciones multivariadas con un trabajo previo sobre las variables cualitativas. Los datos faltantes o no respuestas se suelen codificar como una categoría adicional. Lo mismo se puede hacer con los *no aplica* cuando están presentes. En algunas variables se deben agrupar ca-

tegorías para que no haya categorías con frecuencias muy bajas y que todas las variables tengan más o menos el mismo número de categorías.

En los datos del ejemplo ya se realizó una agrupación en la variable origen del admitido, *orig*, porque las categorías correspondían a los departamentos de Colombia, con muy baja frecuencia para casi todos. Se dejaron tres categorías: Bogotá, Cundinamarca y Otro (resto de departamentos).

Código R para obtener la figura 1.4, que incluye los diagramas de caja y bigotes de edad y de puntaje total del examen

```
par(mfrow=c(3,3),mai=c(0.3,0.4,0.3,0.1),las=1);
for(i in c(2:6)) hist(admi[,i],main=names(admi)[i],
       xlim=c(8,18),ylim=c(0,200));
for(i in c(7,15)) hist(admi[,i],main=names(admi)[i]);
boxplot(admi$age,main="age");
boxplot(admi$exam,main="exam");
```

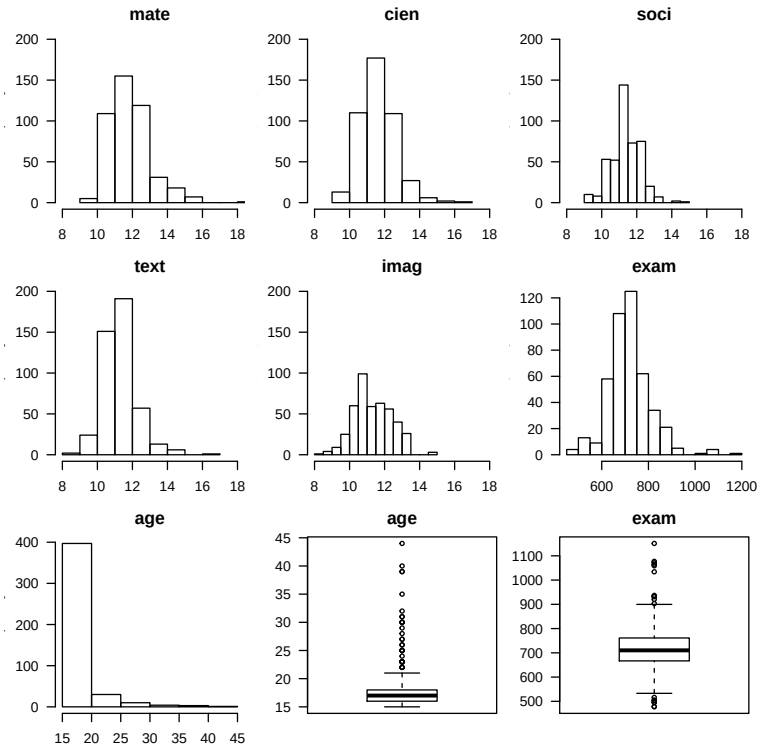


Figura 1.4. Histogramas de los puntajes obtenidos en el examen de los admitidos a la Facultad de Ciencias

La distribución de los admitidos en las tres categorías se muestra en el diagrama de barras *Origen* de la figura 1.3.

La variable estrato del admitido, *stra*, tiene siete categorías, algunas de ellas con frecuencias muy bajas: los estratos 0, 1, 4, 5 y 6 (figura 1.3). Por esta razón y para tener una variable con un número de categorías más parecido a las otras variables sociodemográficas, se decide obtener, a partir de esta, la variable *estr*, con tres categorías: *bajo* (0, 1 y 2), *medio* (3) y *alto* (4, 5 y 6).

En los comandos de R siguientes se muestra una manera de obtener la nueva variable *estr* y su distribución de frecuencias. Esta variable ya está incluida en `admi{FactoClass}`.

Recodificación del estrato

```
estr <-as.integer(admi$stra)-1;
estr[estr<3]<-1;estr[estr==3]<-2; estr[estr>3]<-3;
estr<-factor(estr,labels=c("bajo","medio","alto"));
summary(estr)
## bajo medio  alto
##  179   185    81
```

1.7.2. Codificación en clases de variables continuas

Hay diferentes razones para que algunas veces se decida convertir algunas variables continuas en ordinales. Por ejemplo, para utilizarlas como variables activas en un análisis de correspondencias múltiples. La edad en años se suele pasar a ordinal por su naturaleza de variable sociodemográfica, para analizarla en conjunto con género, estrato, estado civil y otras.

Los criterios guía, para decidir el número de clases y sus límites, provienen del contexto de la investigación, la teoría de la información y las propiedades del análisis de correspondencias múltiples. Se pueden resumir en cuatro (Lebart *et al.*, 2006):

1. Los límites de las clases deben respetar argumentos del contexto, por ejemplo, en el caso de la edad, dieciocho años porque se alcanza la mayoría de edad, sesenta años porque se considera adulto mayor.

2. La frecuencia de las clases debe ser similar porque así se pierde menos información.

3. Evitar clases de baja frecuencia porque son muy influyentes en el análisis de correspondencias múltiples.

4. Buscar que el número de categorías en un análisis de correspondencias múltiples sea más o menos igual en todas las variables activas, ya que las variables con más categorías tienen mayor influencia en este análisis.

El investigador se apoya en estos criterios para tomar las decisiones. Como no todos son compatibles, se busca un compromiso entre ellos. Si se están preparando los datos para un análisis de correspondencias múltiples, es importante tener variables con números de categorías similares y sobre todo evitar categorías de muy baja frecuencia. Un umbral mínimo de 2 % es aconsejable, a menos que el contexto del estudio exija la presencia de categorías de menor frecuencia (Lebart *et al.*, 2006, p. 201).

En el caso de los admitidos, la edad se codificó en cuatro categorías: 16 años o menos, 17, 18 y 19 años o más. En la figura 1.4 tanto en el histograma como en el diagrama de cajas, se observa que la mayor parte de los admitidos son menores de 18 años. La idea para las clases de edad es dejar los años con suficiente frecuencia y unir en los dos extremos. Sin embargo, se obtiene el mismo resultado haciendo la división mediante los cinco números de Tukey (1977, p. 32), que dividen a los individuos en 4 clases buscando que las frecuencias sean similares. Este último procedimiento se muestra en el código de R que aparece a continuación.

División de la edad (`admi$age`) en 4 clases (`admi$edad`)

```
edad<-cut(admi$age,fivenum(admi$age),include.lowest = T,
          labels=c("a16m","a17","a18","a19M"))
summary(edad)
## a16m  a17  a18 a19M
## 118  171   56  100
```

1.8. Ejercicios

1. Realice a mano y verifique con R los ejercicios 1, 2, 4, 5, 6, 7, 12 y 13, del capítulo 2 de Morrison (1990).

2. Instale R y a partir del manual de introducción conteste:

   2.1. ¿En R hay diferencia entre mayúsculas y minúsculas?

   2.2. ¿Con qué se separan las instrucciones de R?

   2.3. ¿Cómo se escriben comentarios en R?

2.4. ¿Qué significa cuando aparece + luego de teclear *Enter*?

2.5. ¿Cómo se recuerdan comandos tecleados previamente en R?

2.6. ¿Qué es el *workspace*?

2.7. ¿Qué se almacena en *.RData*?, ¿qué en *.Rhistory*?

2.8. ¿Cómo se obtiene ayuda en R para una función específica?

2.9. ¿Cuáles son los símbolos de comparación en R: menor que, menor o igual, mayor, mayor o igual, igual y diferente?

2.10. ¿Cuáles son los operadores lógicos: OR, AND y negación?

2.11. ¿Qué efecto tienen `\n` `\t` `\b` al imprimir una cadena de caracteres?

2.12. ¿Cuáles son los principales objetos de R?

2.13. ¿Cómo se define un escalar en R?

2.14. ¿Qué es un factor y qué atributos tiene?

2.15. ¿Qué hace la función `tapply`?

3. Escriba para cada instrucción un comentario resumiendo lo que hace cada función:

3.1. `help.start()` # _____

3.2. `sink("record.lis")` # _____

3.3. `misdatos <-read.table('data.dat')` # _____

3.4. `L2 <- list(A=x, B=y)` # _____

3.5. `ts(1:47, frequency = 12, start = c(1959, 2))` # _____

3.6. `exp1 <-expression(x /(y + exp(z)))` # _____

3.7. `x <- rpois(40, lambda=5)` # _____

3.8. `x[x %%2 == 0]` # _____

3.9. `x <- rnorm(50)` # _____

3.10. `mean(x)` # _____

4. Suponga que usted es la consola de R. Responda al frente a cada uno de los comandos:

4.1. `0/0` _____

4.2. `labs <-paste(c('X','Y'), 1:10, sep='');labs` _____

4.3. `c("x","y")[rep(c(1,2,2,1), times=4)]` _____

4.4. `ls()`

4.5. `apropos("eigen")`

4.6. `x <- 1; mode(x)`

4.7. `seq(1, 5, 0.5)`

4.8. `gl(3, 5)`

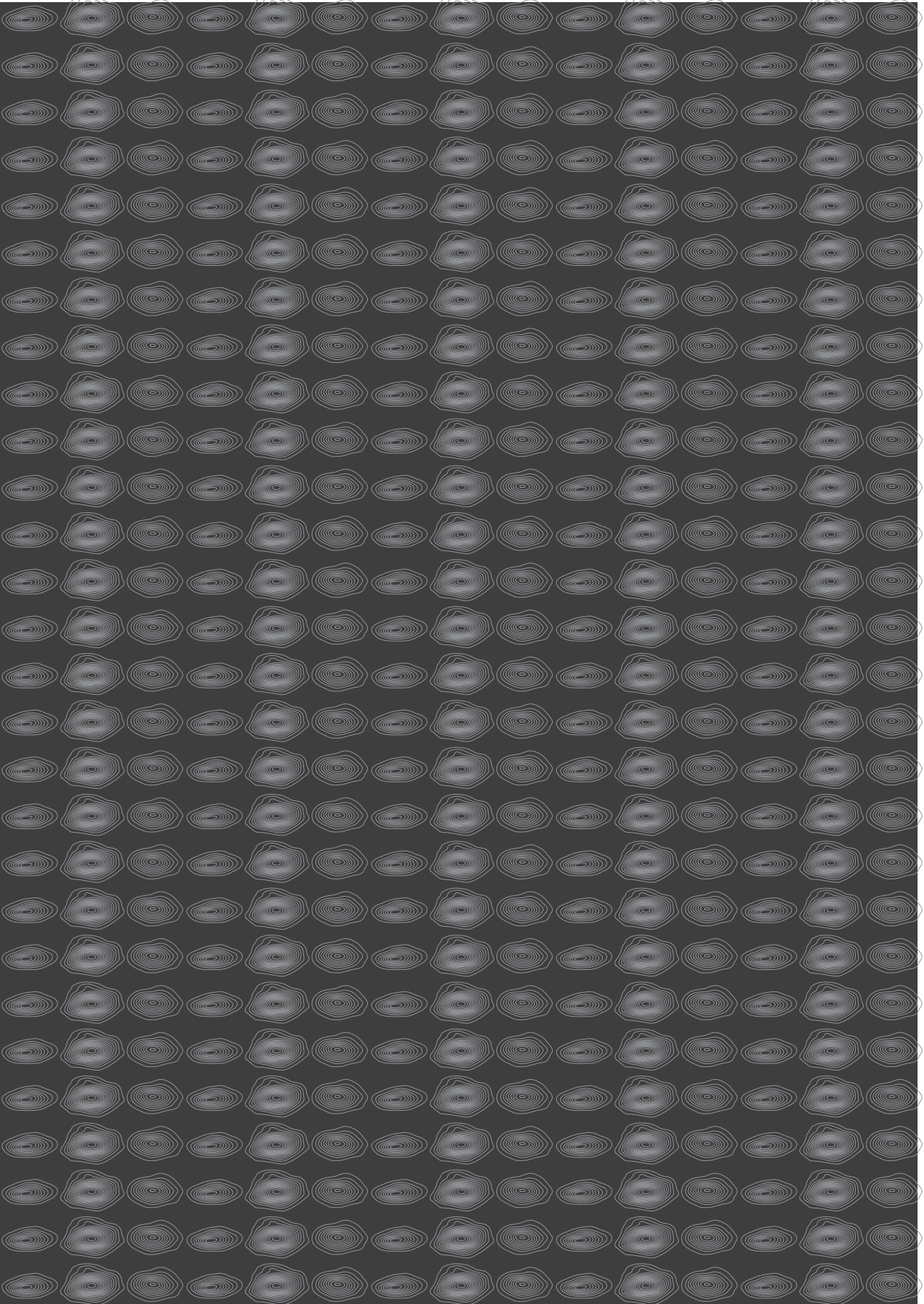
4.9.
```
expand.grid(a=c(60,80), p=c(100, 300),
 sexo=c("Macho", "Hembra"))->trat
dim(trat);class(trat)
```

4.10. `v <- c(10, 20, 30);diag(v)`

Capítulo

dos

Descripción de dos variables



Antes de comenzar con la descripción multivariada de datos, es conveniente repasar la de dos variables. Se consideran dos tipos de variables: continuas y cualitativas, lo que genera tres tipos de descripciones, que se presentan a continuación.

2.1. Descripción de parejas de variables continuas

La descripción de dos variables continuas se logra con los diagramas de dispersión y se complementa con las covarianzas y los coeficientes de correlación. Al tratarlas en conjunto se puede obtener una gráfica que ensambla los diagramas de dispersión dos a dos. La función `plotpairs{FactoClass}` presenta en una gráfica: las densidades *kernel* univariadas en la diagonal, las densidades *kernel* bivariadas en la parte triangular superior y los diagramas de dispersión en la parte triangular inferior.

Las varianzas y covarianzas se arreglan en una matriz simétrica, que tiene en la diagonal las varianzas de las variables y por fuera de la diagonal, las covarianzas. Las correlaciones conforman, también, una matriz simétrica, con unos en la diagonal.

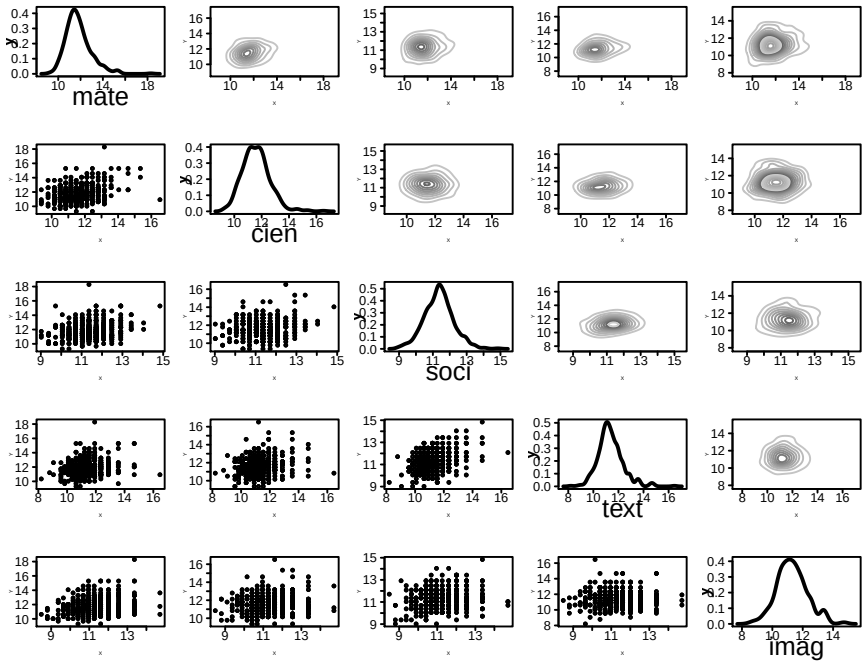
Los diagramas de dispersión de los resultados de los aspirantes admitidos a las carreras de la Facultad de Ciencias, según las áreas –matemáticas, ciencias, sociales, lenguaje e imagen– se muestran en la figura 2.1, junto con la matriz de varianzas y covarianzas *véase* Varianzas y la de correlaciones. Las dos correlaciones más altas son entre sociales y textual, y matemáticas e imagen, lo que se nota en los diagramas de dispersión y densidades *kernel* —ver por ejemplo (Everitt *et al.*, 2011)—. Textual e imagen no están correlacionadas.

La relación entre el puntaje total y los puntajes parciales se da por construcción porque el total es un resumen de los puntajes por áreas:

	mate	cien	soci	text	imag
exam	0.753	0.653	0.593	0.519	0.458

```
xtable(cor(admi$exam,admi[,2:6]),digits=rep(3,6))
```

La componente de matemáticas es la más correlacionada con el puntaje global del examen; la menos correlacionada es imagen.




```
plotpairs(admi[,2:6],col=gray.colors(10,0.8,0.4,2.2));
```

Matriz de varianzas y covarianzas

	mate	cien	soci	text	imag
mate	1.28	0.39	0.24	0.27	0.24
cien	0.39	1.00	0.14	0.20	0.12
soci	0.24	0.14	0.75	0.32	0.09
text	0.27	0.20	0.32	0.98	0.05
imag	0.24	0.12	0.09	0.05	1.00

```
n<-nrow(admi);V<-(n-1)/n*var(admi[,2:6]);
xtable(V,digits=rep(2,6));
```

Matriz de correlaciones

	mate	cien	soci	text	imag
mate	1.00	0.34	0.24	0.24	0.21
cien	0.34	1.00	0.16	0.20	0.12
soci	0.24	0.16	1.00	0.37	0.11
text	0.24	0.20	0.37	1.00	0.05
imag	0.21	0.12	0.11	0.05	1.00

```
R<-cor(admi[,2:6]);
xtable(R,digits=rep(2,6));
```

Figura 2.1. Diagramas de dispersión y densidades *kernel* de los puntajes obtenidos en el examen de los admitidos a la Facultad de Ciencias. Abajo, matrices de covarianzas y correlaciones

2.2. Descripción de una variable continua y una cualitativa

Una variable cualitativa de K categorías establece una partición de los "individuos" en K grupos. Entonces esta descripción se realiza comparando las distribuciones de la variable continua de los K grupos. Los diagramas de cajas y bigotes son apropiados para esto.

Como ejemplo se comparan los resultados del examen de admisión según las carreras en la figura 2.2. Se observa que a las carreras de Geología, Física y Matemáticas se admiten, en promedio estudiantes con mejores resultados en el examen en comparación con las demás carreras. En esta cohorte los admitidos de menores puntajes son los de Química, Farmacia y Estadística. Individualmente los mejores puntajes están en Física, Matemáticas y Geología y los inferiores en Física, Biología y Geología. La distribución más dispersa es la de los admitidos a Física en contraste con las de Estadística y Farmacia.

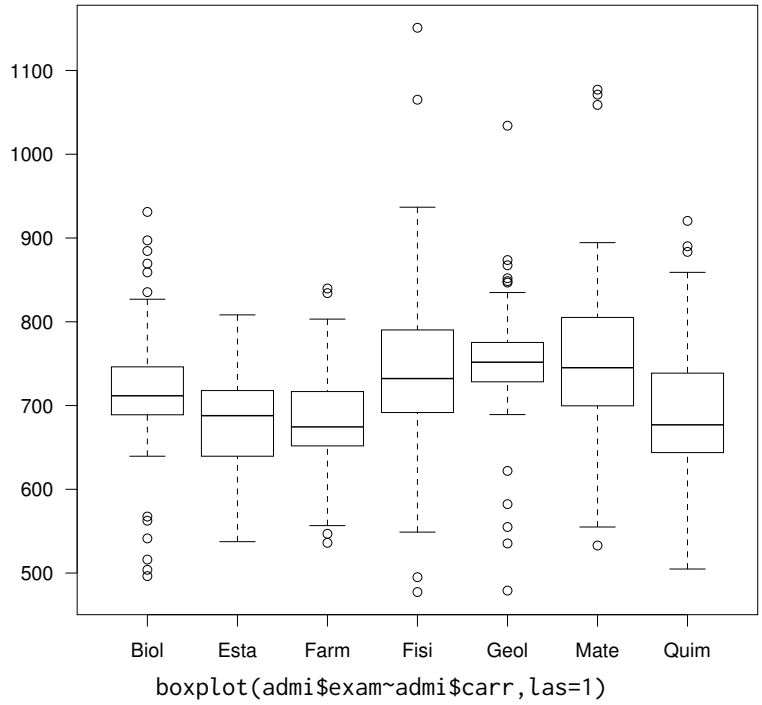


Figura 2.2. Distribuciones del puntaje del examen obtenido por los admitidos según carreras

2.2.1. Razón de correlación

La razón de correlación sirve para medir la relación entre una variable continua X y otra cualitativa Y con K categorías. La partición de los n individuos, inducida por Y, permite descomponer la varianza de X en *varianza entre* y *varianza intra*.

La razón de correlación se define como el cociente entre varianza entre y varianza total:

$$\eta^2_{XY} = \frac{\sigma^2_{entre}}{\sigma^2} \tag{2.1}$$

La varianza entre es:

$$\sigma^2_{entre} = \sum_{k=1}^{K} \frac{n_k}{n} (\overline{X}_k - \overline{X})^2$$

donde $\overline{X}_k = \frac{1}{n_k} \sum_{i \in I_k} X_i$, es decir, el promedio de los individuos que pertenecen a la clase k, y $\overline{X}$ es el promedio de los n individuos, I_k es el conjunto de individuos que pertenecen a la clase k y n_k el número de ellos.

A continuación se calculan las razones de correlación entre las notas de los exámenes y carrera, utilizando la función `centroids{FactoClass}`. Los resultados se expresan en porcentaje:

```
xtable(centroids(admi[,2:7],
admi$carr)$cr*100,digits = rep(2,7));
```

mate	cien	soci	text	imag	exam
15.86	4.31	3.22	2.34	4.42	11.87

Las razones de correlación son bajas, pero se destacan la de matemáticas y la del puntaje global (*exam*) sobre las demás. Entonces, la componente de matemáticas es la que más hace diferencia entre carreras.

2.2.2. Ordenamiento por valores test para describir una variable cualitativa según varias variables continuas

En Morineau (1984) se presentan los denominados *valores test* para buscar las variables continuas que más caracterizan a un grupo de n_k individuos, asociados a una categoría k de una variable cualitativa. El interés es comparar la media del grupo con la media general para cada una de las variables

continuas. El valor test se define como la diferencia de la media del grupo k con respecto a la media global en términos de desviaciones estándar.

Para calcular el valor test se usa como distribución de referencia la de la variable aleatoria $\overline{X}_k$, definida como la media de una muestra aleatoria de tamaño n_k tomada sin reemplazamiento de los n individuos de análisis. La distribución de $\overline{X}_k$ tiene como parámetros, la media de los n datos μ y varianza $\sigma_k^2 = \frac{n-n_k}{n-1}\frac{\sigma^2}{n_k}$, donde $\frac{n-n_k}{n-1}$ es el factor de corrección por tamaño de población finita y σ^2 es la varianza de los n datos. Entonces, el valor test es:

$$t_k = \frac{\bar{x}_k - \mu}{\sigma_k} = \sqrt{\frac{(n-1)n_k}{n-n_k}}\ \frac{\bar{x}_k - \mu}{\sigma} \tag{2.2}$$

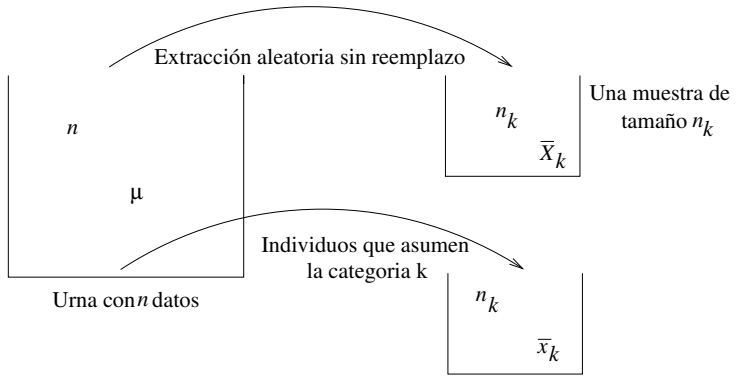


Figura 2.3. Esquema de obtención del valor test: diferencia entre $\overline{x}_k$ y μ, en número de desviaciones estándar

Valores test más grandes que 2 indican que la clase k se caracteriza por tener una media superior a la media global; los inferiores a -2, por tener una media inferior a la global.

Como ejemplo, se calcula el valor test para el puntaje total del examen en la Carrera de Estadística: $n = 445$, $n_{esta} = 66$, $\bar{x}_{esta} = 680.2$, $\mu = 718.4$, $\sigma^2 = 8039$, entonces:

$$t_{esta} = \sqrt{\frac{444 * 66}{445 - 66}}\ \frac{680.2 - 718.4}{\sqrt{8039}} = -3.746$$

La función `cluster.carac{FactoClass}` ordena las categorías según el valor test y solo presenta a las que en valor absoluto son mayores que 2, parámetro que se puede cambiar al llamar la función.

En la tabla 2.1 se muestra la caracterización de las carreras según los resultados por áreas y puntaje total obtenidos por sus admitidos en el examen. Al pie de la tabla se muestra el código R para obtenerla.

Código para obtener la tabla 2.1

```
cluster.carac(admi[,2:7],admi$carr,tipo.v="co", dm=1)->
   desCarrExamen;
xtable(list.to.data(desCarrExamen),digits=c(0,0,3,1,0,1));
```

Tabla 2.1. Caracterización de las carreras según los resultados por áreas y global del examen de admisión

Categoría	Carrera	V.test	Media clase	Frecuencia clase	Media global
mate	Biol	-2.258	11.5	63	11.8
text	Esta	-2.598	11.1	66	11.4
soci		-2.839	11.1		11.4
cien		-3.576	11.2		11.6
exam		-3.745	680.2		718.4
imag	Farm	-2.945	11.0	73	11.3
exam		-3.399	685.7		718.4
mate		-4.472	11.3		11.8
mate	Fisi	3.374	12.2	82	11.8
exam		3.316	748.0	82	718.4
cien		2.482	11.8	82	11.6
exam	Geol	2.467	749.6	45	718.4
mate		2.045	12.1		11.8
mate	Mate	5.816	12.6	53	11.8
exam		3.909	763.5		718.4
imag		3.128	11.7		11.3
mate	Quim	-2.100	11.5	63	11.8

En la tabla 2.1 se puede observar que los admitidos a Biología obtienen en promedio menores puntajes en matemáticas; los de Estadística, en las áreas textual, de sociales y de ciencias; los de Farmacia, en imagen y matemáticas y los de Química en matemáticas. Los admitidos a Física obtienen mayores puntajes promedio en matemáticas y ciencias; los de Geología, mayores en matemáticas, y los de Matemáticas, en matemáticas e imagen.

2.3. Descripción de dos variables cualitativas

Para describir la asociación de dos variables cualitativas se construye una tabla de contingencia (TC), que clasifica a los individuos por las categorías de las dos variables simultáneamente.

Por ejemplo, la tabla 2.2 muestra la asociación entre las variables cualitativas edad y estrato de los admitidos, donde se incluyen los totales fila y columna, y la tabla de frecuencias relativas, con sus marginales.

La tabla de frecuencias relativas es la distribución conjunta de probabilidad (empírica) de *edad* y *estrato*. La marginal fila (sumas por filas) es la distribución de probabilidad de la *edad* y la marginal columna, la distribución de *estrato*. Estas distribuciones univariadas se denominan también *distribuciones marginales*.

Código para obtener la tabla 2.2

```
table(admi$edad,admi$estr)->tc;
tabtc<-cbind(tc,totF=rowSums(tc));
tabtc<-rbind(tabtc,totC=colSums(tabtc));
xtable(cbind(tabtc,round(tabtc/445*100,1)),digits=c(rep(0,5),
       rep(1,4)));
```

Tabla 2.2. Tabla de contingencia edad × estrato de los admitidos, tabla de frecuencias relativas

Edad	Estrato							
	Frecuencia				Porcentaje			
	bajo	medio	alto	totF	bajo	medio	alto	totF
a16m	44	47	27	118	9.9	10.6	6.1	26.5
a17	58	74	39	171	13.0	16.6	8.8	38.4
a18	22	26	8	56	4.9	5.8	1.8	12.6
a19M	55	38	7	100	12.4	8.5	1.6	22.5
totC	179	185	81	445	40.2	41.6	18.2	100.0

Al dividir una fila de la tabla de contingencias o de la tabla de frecuencias por su suma (marginal), se obtiene la distribución condicional de la fila. Por ejemplo, la distribución para el grupo de los admitidos de 17 años se obtiene dividiendo la fila 2 por 171. Del mismo modo, al dividir una columna por su suma se obtiene su condicional. Por ejemplo, dividiendo la primera columna por 179 se obtiene la distribución condicional de estrato bajo.

En la tabla 2.3 se presentan las distribuciones condicionales fila y columna, denominadas también *perfiles*.

Tabla 2.3. Perfiles fila y columna de la tabla edad × estrato

Perfiles fila

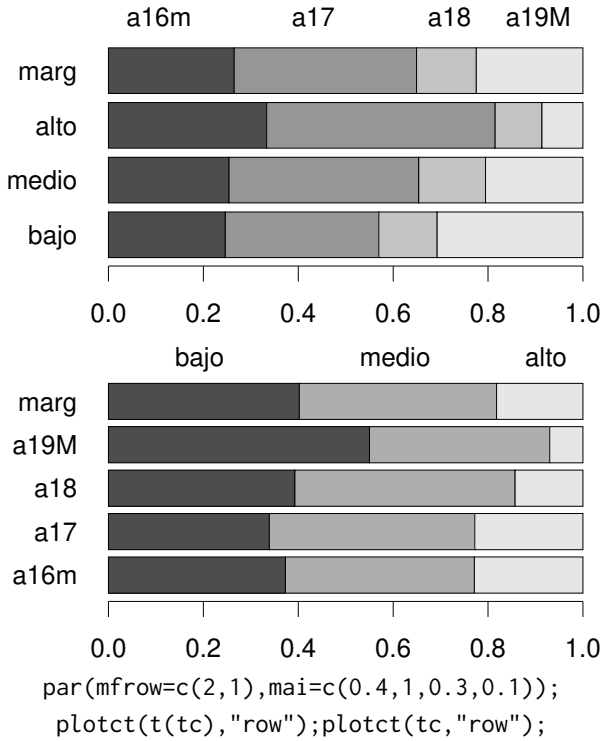
	bajo	medio	alto	Suma
a16m	37.3	39.8	22.9	100.0
a17	33.9	43.3	22.8	100.0
a18	39.3	46.4	14.3	100.0
a19M	55.0	38.0	7.0	100.0

Perfiles columna

	bajo	medio	alto
a16m	24.6	25.4	33.3
a17	32.4	40.0	48.1
a18	12.3	14.1	9.9
a19M	30.7	20.5	8.6
Suma	100.0	100.0	100.0

Código para obtener la tabla 2.3 (editada)

```
prop.table(tc,1)->pf;
prop.table(tc,2)->pc;
xtable(addmargins(pf,margin=2)*100,digits=rep(1,5));
xtable(addmargins(pc,margin=1)*100,digits=rep(1,4))
```




```
par(mfrow=c(2,1),mai=c(0.4,1,0.3,0.1));
plotct(t(tc),"row");plotct(tc,"row");
```

Figura 2.4. Perfiles fila y columna de la TC edad × estrato

La asociación entre categorías fila y columna se visualiza con los histogramas de perfiles fila y columna, puestos como barras del 100 % cuyas franjas representan el porcentaje de cada categoría en el histograma de la fila o columna.

Se observa, en la figura 2.4 que los admitidos de mayor edad tienen más porcentaje de estrato bajo y los de 17 años y menos un poco más de estrato alto. Por otro lado, los estudiantes de estrato bajo tienen mayor porcentaje de 19 o más años de edad.

En tablas de datos es común interesarse por una variable cualitativa que se desea describir por otras variables también cualitativas. Como ejemplo podemos ver los perfiles de las carreras según las demás variables cualitativas en la figura 2.5.

En esa misma figura se puede ver, por ejemplo: Farmacia es la carrera donde se admitieron más mujeres, Química es la que tuvo más porcentaje de estrato bajo, a Farmacia y Estadística se admitió más porcentaje de bogotanos, Química fue la de mayor porcentaje de admitidos de 16 años o menos, Estadística tuvo mayor porcentaje de los que debieron nivelar lecto-escritura, y en Farmacia y Biología hubo mayor porcentaje de los que tuvieron que nivelar matemáticas.

2.3.1. Dos medidas de asociación entre variables cualitativas

En una tabla de contingencia, denotada N, con J categorías en fila y K categorías en columna, n_{jk} es el número de individuos que asumen simultáneamente las categorías j y k; $n_{j.}$, el número de los que asumen la categoría j (marginal fila); $n_{.k}$, el de los que asumen la categoría k (marginal columna), y n, el total de la tabla (número de individuos).

La estadística χ^2 utilizada para contrastar la hipótesis de independencia se puede usar como medida de asociación entre dos variables cualitativas y se expresa como (Canavos, 1988, p. 370):

$$\chi^2 = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{jk} - \dfrac{n_{j.}n_{.k}}{n}\right)^2}{\dfrac{n_{j.}n_{.k}}{n}} \tag{2.3}$$

Para probar la independencia en una tabla de contingencia, se utiliza la distribución asintótica de χ^2 que es χ^2 con $(J-1)(K-1)$ grados de libertad. La aproximación se considera buena si ninguna celda de la tabla

bajo el supuesto de independencia, de término general $\frac{n_{j.}n_{.k}}{n}$, es inferior a 5 (Agresti, 2002, p. 78). Aquí se utiliza como índice descriptivo.

Se calcula además su valor p, que es igual a $P(\chi^2 \geq \chi_c^2)$, donde χ_c^2 es el valor calculado en la tabla de contingencia. Al valor p se le asocia el cuantil de la normal estándar, denominado valor test:

$$t \ \ tal\ que\ P(Z \geq t) = valor\ p, \ \ donde\ \ Z \sim N(0,1)$$

Código para obtener la figura 2.5 (está editada)

```
par(mfrow=c(3,2),mai=c(0.3,1,0.2,0.1),las=1,cex=0.7);
for (i in 8:13){
  tc<-unclass(table(admi$carr,admi[,i]));
  plotct(tc,"row");
  title(main=names(admi)[i]);
}
```

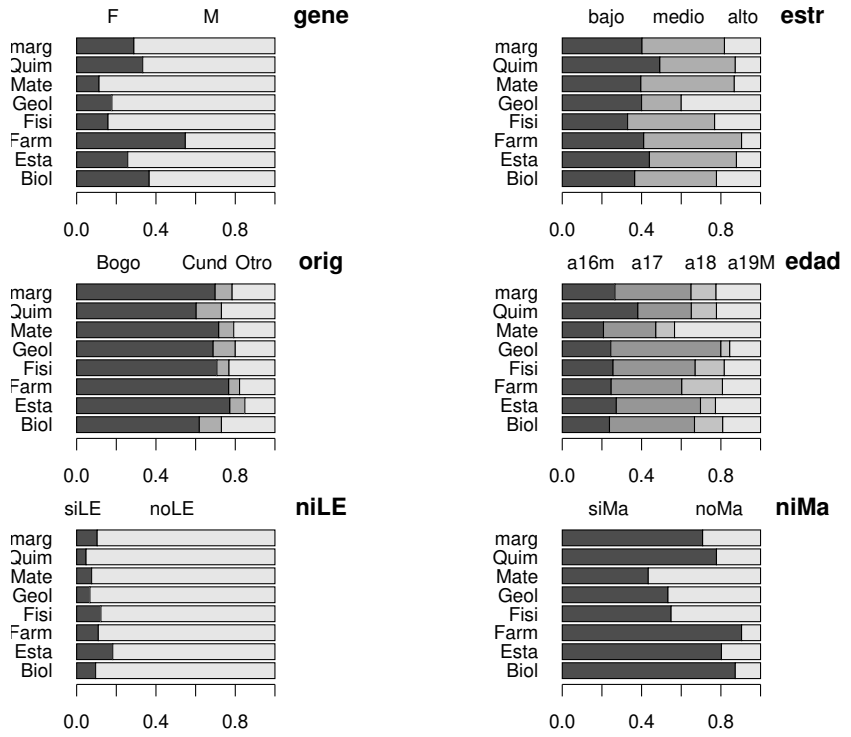


Figura 2.5. Perfiles de las carreras según variables cualitativas

La estadística χ^2 depende del total de la tabla n. Por lo tanto, también se utiliza el índice de asociación ϕ^2:

$$\phi^2 = \frac{\chi^2}{n} \tag{2.4}$$

Con la función `chisq.carac{FactoClass}` se pueden obtener los índices de asociación entre una variable cualitativa (a caracterizar) y otras variables cualitativas. Se muestra, como ejemplo, la asociación de las carreras con las variables sociodemográficas de los admitidos: *gene*, *estr*, *orig* y *edad*, en la tabla 2.4. Solo entre carreras y origen de los admitidos no se muestra asociación. En la sección siguiente se muestran las categorías responsables de las relaciones entre carreras y las otras variables a través de sus categorías.

2.3.2. Ordenamiento por valores test para describir una variable cualitativa según las categorías de varias variables cualitativas

Para ordenar automáticamente las categorías más características en los perfiles de la variable cualitativa, que se está describiendo (categoría k), se recurre a calcular los valores test (Morineau, 1984).

La proporción de individuos que asumen una categoría j (de otra variable que caracteriza) dentro del grupo k: $\frac{n_{jk}}{n_k}$ se compara con la proporción de la categoría j en los n individuos: $\frac{n_j}{n}$. Una manera de definir esa diferencia, cuando es positiva, es calculando la probabilidad de obtener n_{jk} individuos o más, en una muestra aleatoria de tamaño n_k sin reemplazamiento. El esquema de comparación se muestra en la figura 2.6.

Tabla 2.4. Estadísticas χ^2 entre carreras y variables sociodemográficas

variable	χ^2	dfr	pval	tval	ϕ^2
gene	44.1	6	0.000	5.264	0.099
estr	29.2	12	0.004	2.679	0.066
orig	9.7	12	0.644	-0.370	0.022
edad	33.6	18	0.014	2.189	0.075

```
xtable(chisq.carac(admi[,8:11],admi$carr,decr = FALSE),
                digits=c(0,1,0,3,3,3))
```

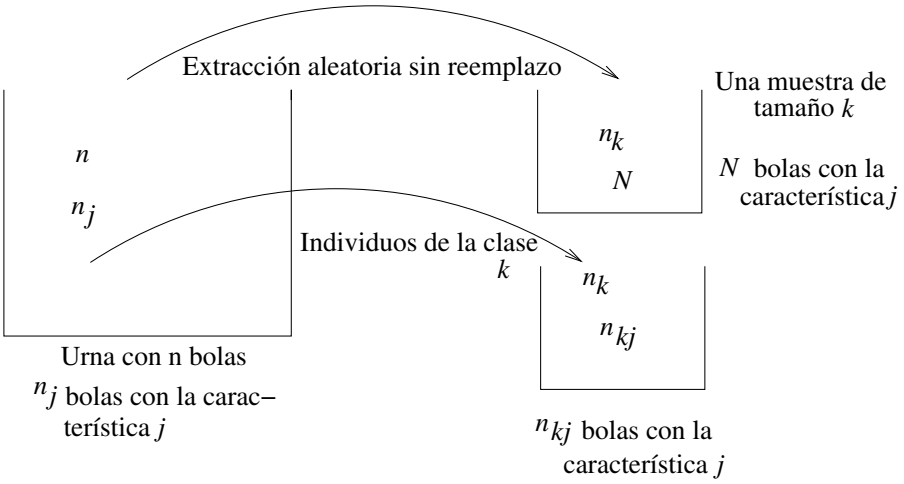


Figura 2.6. Esquema para obtener el valor p como índice de comparación entre la proporción de la categoría j dentro de la clase k y su proporción global

En la urna hay n "bolas" (que representan a los individuos) y entre estas hay algunas (n_j) que tienen la característica j. De la urna se extrae una muestra al azar y sin reemplazo de tamaño n_k.

Sea la variable aleatoria N = "número de bolas con la característica j al extraer n_k bolas al azar y sin reemplazo". La variable aleatoria N tiene distribución hipergeométrica de parámetros n, n_j, n_k. La probabilidad que se desea calcular, notada *valor p*, es:

$$valor\ p = \begin{cases} P(N \geqslant n_{kj}) & \text{si , } \dfrac{n_{kj}}{n_k} \geqslant \dfrac{n_j}{n} \\ P(N \leqslant n_{kj}) & \text{en otro caso.} \end{cases}$$

Si la probabilidad es muy baja –por ejemplo menor que 0.05–, la categoría j caracteriza a la clase k por tener mayor o menor proporción de la categoría dentro de la clase comparada con su proporción global. El valor p no es práctico porque es decimal y puede tener muchos ceros a la izquierda. Por eso Morineau (1984) propone transformarlos en valores test, que son cuantiles de la normal estándar. El valor test es el cuantil de la normal estándar que corresponde al valor p:

$$valor\ test = t \text{ tal que } \begin{cases} P(Z \geqslant t) = \frac{valor\ p}{2}) & \text{si } \dfrac{n_{kj}}{n_k} \geqslant \dfrac{n_j}{n} \\ P(Z \leqslant t) = \frac{valor\ p}{2}) & \text{en otro caso.} \end{cases} \tag{2.5}$$

La división del valor p sobre 2 se introduce porque se consideran las dos diferencias, es decir, el valor test está asociado a una prueba de hipótesis bilateral o de dos colas. Cuando la proporción de la característica j dentro del grupo k es menor que la global, el valor test es negativo.

La función `cluster.carac{FactoClass}` calcula los valores p y los valores test asociados. En el objeto de salida, para cada una de las categorías de la variable que se está caracterizando (clase k) se presentan las categorías de las demás variables que la caracterizan (categorías j), ordenadas por los valores test. `cluster.carac` muestra solamente las categorías que tienen valores test superiores, en valor absoluto, al umbral $v.lim = 2$, parámetro que se puede cambiar (ver ayuda de la función con el comando `?cluster.carac`).

En la tabla 2.5 se muestra la salida que caracteriza las carreras según las cuatro variables sociodemográficas y las dos variables que indican si los admitidos tienen que nivelar o no. Esta tabla se puede considerar un resumen de la figura 2.5.

La tablas: de contingencia, que cruza $niMa \times carr$; de frecuencias relativas; de perfiles fila y de perfiles columna se presentan en la tabla 2.6 y de ellas se deriva el resumen de la tabla 2.5.

Código para obtener la tabla 2.6

```
tcCarrNiMa<-unclass(table(admi$niMa,admi$carr));
tabs<-plotct(tcCarrNiMa,tables=TRUE);
xtable(tabs$ctm,digits=rep(0,9));
xtable(tabs$ctm*100/445,digits=rep(1,9));
xtable(tabs$perR,digits=rep(1,8));
xtable(tabs$perC,digits=rep(1,9))
```

Considérense, por ejemplo, algunos de los valores de la primera fila de la tabla 2.5: el 87.3 % de la celda (1, 1) de los perfiles columna; el 70.8 % corresponde al marginal de la fila $siMa$ de la tabla de frecuencias relativas; el 17.5 % de la celda (1, 1) de la tabla de perfiles fila y el 315 marginal $siMa$ de la tabla de contingencia.

Para mostrar la lectura y un cálculo, tomemos la primera fila de la tabla 2.5, corresponde a la categoría *siMa*, dentro del grupo de los admitidos a Biología. El valor test 3.332, que corresponde a un valor p de 0.001, indica que el porcentaje de 87.3 % (columna *Cat/Cl*) de los que tienen que nivelar matemáticas dentro de Biología es característico, comparado con el 70.8 % (penúltima columna) de todos los admitidos.

El 17.5 % de la columna *Cl/Cat* indica que de los 315 (última columna) admitidos que tienen que nivelar matemáticas el 17.5 % están en los admitidos a Biología. Como *niMa*, solo tiene dos categorías, el valor test para

los que no tienen que nivelar es -3.332, es decir, Biología se caracteriza por tener menos proporción que el global de los que no tienen que nivelar matemáticas. Sin embargo, este resultado complementario y no siempre se utiliza en el análisis.

A continuación se ilustra el cálculo del valor p y del valor test para la categoría *siMa* dentro de Biología.

La variable aleatoria N = *número de admitidos que tienen que nivelar matemáticas en la muestra de tamaño 63, tomada sin reemplazo de los 445 admitidos a Ciencias, donde hay 315 que tienen que nivelar matemáticas*, tiene dis-

Tabla 2.5. Caracterización de las carreras según algunas variables cualitativas

Categoría	Carrera	V.test	Valor p	Cl/Cat	Cat/Cl	Global	n_{cat}
niMa.siMa	Biol	3.332	0.001	17.5	87.3	70.8	315
niMa.noMa		-3.332	0.001	6.2	12.7	29.2	130
niLE.siLE	Esta	2.235	0.025	26.1	18.2	10.3	46
niMa.siMa		2.034	0.042	16.8	80.3	70.8	315
niMa.noMa		-2.034	0.042	10.0	19.7	29.2	130
niLE.noLE		-2.235	0.025	13.5	81.8	89.7	399
gene.F	Farm	5.152	0.000	31.2	54.8	28.8	128
niMa.siMa		4.355	0.000	21.0	90.4	70.8	315
edad.a18		2.252	0.024	26.8	20.5	12.6	56
estr.alto		-2.281	0.023	8.6	9.6	18.2	81
niMa.noMa		-4.355	0.000	5.4	9.6	29.2	130
gene.M		-5.152	0.000	10.4	45.2	71.2	317
niMa.noMa	Fisi	3.475	0.001	28.5	45.1	29.2	130
gene.M		3.045	0.002	21.8	84.1	71.2	317
gene.F		-3.045	0.002	10.2	15.9	28.8	128
niMa.siMa		-3.475	0.001	14.3	54.9	70.8	315
estr.alto	Geol	3.677	0.000	22.2	40.0	18.2	81
niMa.noMa		2.706	0.007	16.2	46.7	29.2	130
edad.a17		2.554	0.011	14.6	55.6	38.4	171
niMa.siMa		-2.706	0.007	7.6	53.3	70.8	315
estr.medio		-3.242	0.001	4.9	20.0	41.6	185
niMa.noMa	Mate	4.467	0.000	23.1	56.6	29.2	130
edad.a19M		3.683	0.000	23.0	43.4	22.5	100
gene.M		3.218	0.001	14.8	88.7	71.2	317
edad.a17		-2.089	0.037	8.2	26.4	38.4	171
gene.F		-3.218	0.001	4.7	11.3	28.8	128
niMa.siMa		-4.467	0.000	7.3	43.4	70.8	315
edad.a16m	Quim	2.320	0.020	20.3	38.1	26.5	118
edad.a17		-2.189	0.029	9.9	27.0	38.4	171

```
desCarrSocio<-cluster.carac(admi[,8:13],admi[,1])
xtable(list.to.data(desCarrSocio),digits=c(0,0,3,3,1,1,1,0))
```

Tabla 2.6. TC de nivela matemáticas × carreras y tablas de perfiles fila y columna, incluyendo marginales

Tabla de contingencia

niMa	Biol	Esta	Farm	Fisi	Geol	Mate	Quim	marR
siMa	55	53	66	45	24	23	49	315
noMa	8	13	7	37	21	30	14	130
marC	63	66	73	82	45	53	63	445

Frecuencias relativas en porcentaje

siMa	12.4	11.9	14.8	10.1	5.4	5.2	11.0	70.8
noMa	1.8	2.9	1.6	8.3	4.7	6.7	3.1	29.2
marC	14.2	14.8	16.4	18.4	10.1	11.9	14.2	100.0

Perfiles fila

siMa	17.5	16.8	21.0	14.3	7.6	7.3	15.6	100
noMa	6.2	10.0	5.4	28.5	16.2	23.1	10.8	100

Perfiles columna

siMa	87.3	80.3	90.4	54.9	53.3	43.4	77.8	
noMa	12.7	19.7	9.6	45.1	46.7	56.6	22.2	
	100	100	100	100	100	100	100	

tribución hipergeométrica, con parámetros $n = 445$, $n_k = n_{biol} = 63$, $n_j = n_{siMa} = 315$, es decir: $N \sim H(445, 315, 63)$.

El valor de los admitidos a Biología que tienen que nivelar matemáticas, $n_{kj} = 55$, se toma como una muestra específica. Entonces, hay que calcular primero la probabilidad

$$P(N \geq 55) \text{ con el modelo } N \sim H(445, 315, 63)$$

Esa probabilidad es el valor p y el cuantil de la normal estándar que le corresponde es el valor test, calculado como se muestra en el código siguiente:

Código para calcular el valor p y el valor test, y obtener la figura 2.7

```
vp<-phyper(54, 315, 130,63,lower.tail=FALSE);vp;
# [1] 0.000862394
qnorm(vp/2,lower.tail=FALSE)
# [1] 3.331951
curve(dnorm(x),xlim=c(-5,5),las=1) # densidad N(0,1)
abline(v=seq(-5,5),h=seq(0,0.4,0.1),col="gray") # grilla
abline(v=c(-3.332,3.332),lty=2) # cuantiles
# valores de cuantiles y areas  agregados con xfig
```

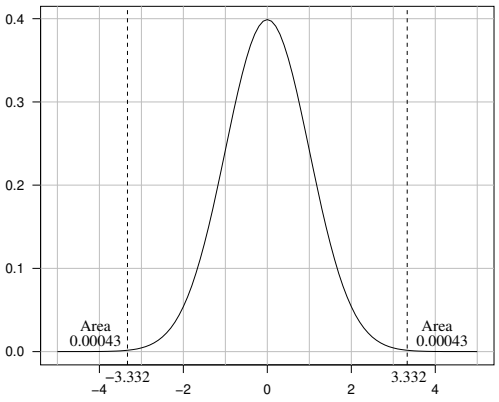


Figura 2.7. Ilustración de la obtención del valor test a partir de una probabilidad (área de las dos colas bajo la curva normal estándar)

En la figura 2.7 se muestra la recodificación de la probabilidad obtenida al valor test como cuantil de la normal estándar. El área que representa la probabilidad se reparte en los extremos de la distribución es decir, $0.00086/2 = 0.00043$, lo que permite establecer un umbral de 2 para el valor test.

2.4. Ejercicios

1. Deduzca la razón de correlación (2.1).
2. Deduzca el valor test (2.2).
3. Use (2.2) para verificar el valor test de *imag* para *Mate* en la tabla 2.1.
4. Obtenga todos los valores de la fila *gene* en la tabla 2.4.
5. Deduzca la ecuación (2.5).
6. Obtenga todos los valores de la fila *estr.alto* para *Geol* en la tabla 2.5.

2.5. Taller: caracterización de la función de razas de perros

En un estudio de 27 razas de perros, se registran seis variables que miden las cualidades físicas o psíquicas de la raza y una variable que indica la función para la que se suele emplear la raza:

Variables	Categorías		
Tamaño	Pequeño	Medio	Grande
Peso	Liviano	Medio	Pesado
Velocidad	Baja	Media	Alta
Inteligencia	Pequeña	Media	Grande
Afectividad	Pequeña	Grande	
Agresividad	Pequeña	Grande	
Función	Compañía	Caza	Utilidad

Las razas de función utilidad incluyen: salvamento, defensa, perro lazarillo, perro de policía, etc.

El objetivo del ejercicio es describir las características físicas y psíquicas de las razas más relacionadas con cada una de las categorías de la función.

Los datos se encuentran en el paquete `FactoClass` como `DogBreeds`. Para ver la asociación entre cada una de las variables físicas y psíquicas con la utilidad, use la función `chisq.carac{FactoClass}` y para ordenar las categorías según su caracterización, use `cluster.carac{FactoClass}`

Preguntas

Complete

1. Las tres variables que más caracterizan a la función para la cual se utilizan las razas son: _____
2. La estadística χ^2 asociada a la tabla de contingencia $peso \times funcion$ es: _____
3. Para encontrar el valor p se utiliza la distribución: _____ con _____ grados de libertad.
4. El valor test se puede obtener con el comando de R:

   _____

Conteste *falso* o *verdadero* a las afirmaciones siguientes:

5. ☐ De las razas utilizadas para compañía el 71.4 % es de afectividad alta.
6. ☐ Todas las razas utilizadas para compañía son de afectividad alta.
7. ☐ Todas las razas de afectividad alta se utilizan para compañía.
8. ☐ Hay catorce razas utilizadas para compañía.
9. ☐ Hay catorce razas de afectividad alta.

10. ☐ En las razas utilizadas para compañía hay catorce de afectividad alta.
11. ☐ Todas las razas pesadas son de utilidad.
12. ☐ Todas las razas de utilidad son de tamaño grande.
13. ☐ La velocidad no caracteriza a las razas de caza.
14. ☐ Hay cinco razas de utilidad.

Capítulo

tres

Análisis en componentes principales

El análisis en componentes principales (ACP) se utiliza para describir tablas que tienen en las filas las unidades estadísticas, generalmente denominadas, "individuos", y en las columnas las variables de tipo continuo que se han medido sobre los individuos. Los objetivos del ACP son:

1. Comparar los individuos entre si. Las gráficas que se obtienen permiten observar la forma de la "nube de individuos", lo que a su vez permite detectar patrones en ellos.

2. Describir las relaciones entre las variables.

3. Reducir la dimensión de la representación. A mayor relación entre las variables mayor es la capacidad de síntesis del ACP y unos pocos ejes factoriales podrán resumir las variables originales.

A continuación se desarrolla el ACP paso a paso utilizando un ejemplo muy pequeño, lo que permite observar datos y gráficos de manera completa, entender los pasos lógicos y las claves para su lectura y aprender a usar las ayudas numéricas para interpretarlas.

En el ejemplo de aplicación de la sección 3.6 se utilizan funciones específicas de R para llevar a cabo un ACP que sirve de guía para los trabajos propios de los lectores.

3.1. Ejemplo "Café"

En Duarte *et al.* (1996) se presenta un experimento en el que se preparan tazas de café para detectar la influencia de la contaminación del grano con maíz y cebada. La tabla de datos está incluida en `cafe{FactoClass}` y tiene doce filas y diesiséis columnas.

El experimento considera tres factores: agregado (sin o excelso, maíz, cebada), porcentaje del agregado (20 % y 40 %) y grado de tostación (clara y oscura). Con el café molido de cada uno de los diez tratamientos se preparan tazas de café para medir propiedades químicas, físicas y sensoriales.

En este ejemplo se utilizan solamente las variables físicas —*color*, DA: densidad aparente, EA: extracto acuoso (contenido de sólidos solubles)— y las diez primeras filas, que corresponden a los tratamientos del experimento.

Los valores obtenidos para los diez tratamientos se muestran en la figura 3.1. Los identificadores de los ocho tratamientos que corresponden a los cafés contaminados con maíz y cebada se construyen así: la primera letra indica el grado de tostación (C = claro, O = oscuro); los dos números, el porcentaje de contaminación (20,40) y la última letra el contaminante

(M = maíz, C = cebada). Por ejemplo, C20M quiere decir tostación clara, con 20 % de agregado de maíz. Los dos cafés excelsos se identifican como ExCl y ExOs.

Código cargar los datos y construir la gráfica 3D con los cafés e imprimir la matriz **Y** (figura 3.1).

```
library(FactoClass); data(cafe); Y <- cafe[1:10,1:3];
par(las=1); # grafica;
Y3D<-scatterplot3d(Y,main="Y",type="h",color="black",box=FALSE,
       las=1);
Y3D$points3d(Y,pch=1);
addgrids3d(Y, grid = c("xy","xz","yz"));
cord2d<-Y3D$xyz.convert(Y) # convertir cordenadas 3D a 2D;
# poner etiquetas;
text(cord2d,labels=rownames(Y), cex=0.8,col="black",pos=3);
xtable(Y,digits=c(0,0,1,0)) # para tabular de LaTeX;
```

La matriz **Y** contiene los datos "activos" del ejemplo. Las 10 filas ($n = 10$) de **Y** se representan como puntos en $\mathbb{R}^3$ ($p = 3$), imagen que se denomina *nube de individuos* (figura 3.1).

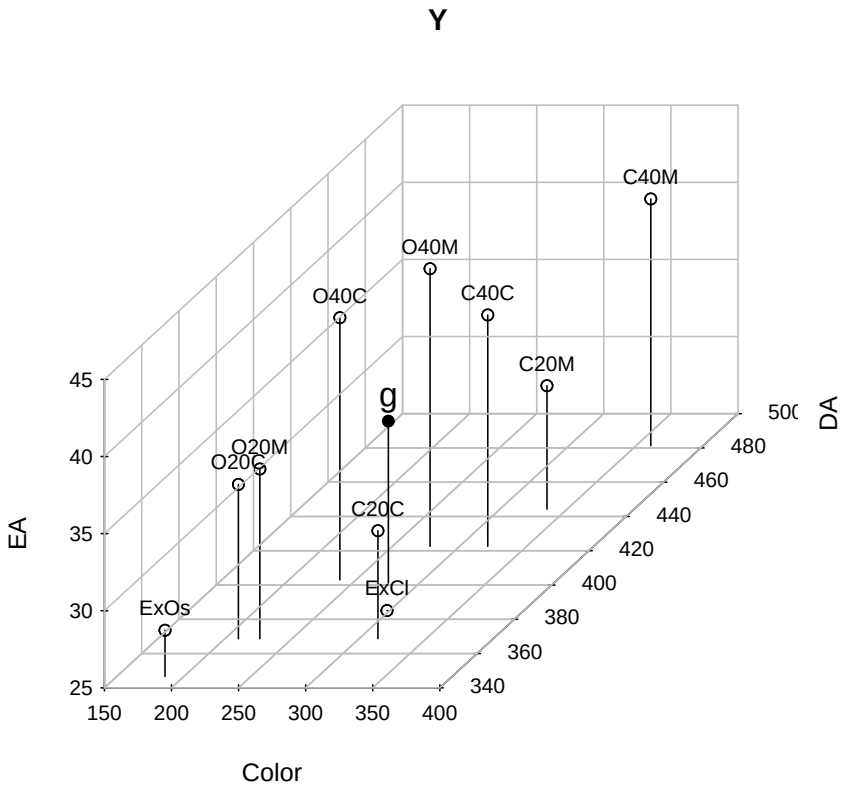
Las columnas de **Y** representan a las variables. Cada una se puede ver como un vector en $\mathbb{R}^{10}$. Esta geometría es abstracta, pero tiene las mismas propiedades de la geometría en 3D ($\mathbb{R}^3$). Los tres vectores (*color*, DA y EA) constituyen la *nube de variables*.

3.2. Nube de individuos N_n

En la nube de los n individuos en $\mathbb{R}^p$ los ejes son las variables y las coordenadas de cada punto son los valores de las variables que asume el individuo de la fila de **Y**. En la figura 3.1, se muestra en 3D la nube de los diez individuos del ejemplo "Café".

3.2.1. Centro de gravedad

Sobre la nube de individuos se define el *centro de gravedad*, notado **g**, que generaliza el concepto de *media* como una medida de localización multivariada. Cada individuo se dota de un peso p_i, tal que $\sum_{i=1}^{n} p_i = 1$.



Matriz **Y**

Cafe	Color	DA	EA
ExCl	298	385.1	25
C40M	361	481.3	41
C40C	321	422.6	40
C20M	335	444.3	33
C20C	314	368.7	32
ExOs	186	346.6	28
O40M	278	422.6	43
O40C	238	403.0	42
O20M	226	368.7	36
O20C	210	368.7	35

Figura 3.1. Representación de la tabla de datos del ejemplo Café en 3D. Se muestra el centro de gravedad **g**.

Cuando los individuos tienen el mismo peso $\frac{1}{n}$, el centro de gravedad es la suma de los n vectores individuo, notados $\mathbf{y}_i$, multiplicada por el escalar $1/n$:

$$\mathbf{g} = \sum_{i=1}^{n} p_i \mathbf{y}_i = \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i \tag{3.1}$$

El centro de gravedad constituye un individuo artificial denominado *típico* porque es el punto de referencia para comparar a los demás.

Código para calcular y adicionar el centro de gravedad a la figura 3.1

```
g <- colMeans(Y) #centro de gravedad;
Y3D$points3d(t(g),pch=19,col="darkgreen",type = "h");
text(Y3D$xyz.convert(t(g)),labels="g",pos=3,col="black",
     cex=1.3);
```

3.2.2. Centrado de la nube de individuos

El centrado de los individuos permite trasladar el cero de la representación al centro de gravedad. En la gráfica centrada se pierden las coordenadas del centro de gravedad, por lo tanto, es necesario registrar esos valores, que son los promedios de las variables.

La coordenada de un individuo centrado $\mathbf{y}_{Ci}$, se obtiene restándole las coordenadas del centro de gravedad $\mathbf{g}$:

$$\mathbf{y}_{Ci} = \mathbf{y}_i - \mathbf{g} \tag{3.2}$$

La matriz de datos centrados $\mathbf{Y}_\mathbf{C}$ se obtiene mediante:

$$\mathbf{Y}_\mathbf{C} = \mathbf{Y} - \mathbf{1}_n \mathbf{g}' \tag{3.3}$$

donde $\mathbf{1}_n$ es el vector de n unos.

Al representar $\mathbf{Y}_\mathbf{C}$ en $\mathbb{R}^p$ el origen se traslada al centro de gravedad de la nube N_n, hecho que se muestra en el esquema de la figura 3.2.

En la figura 3.2, se muestra la representación en 3D de la nube de puntos centrados del ejemplo "Café". La representación centrada tiene la misma forma que la original, pero las coordenadas de los cafés han cambiado.

Las coordenadas en cada eje representan la diferencia de un café con respecto al café típico, por ejemplo, el café excelso con tostación clara (ExCl) tiene 21.30 unidades de color más que el café típico, 16.06 unidades menos de densidad aparente y 10.05 unidades menos de extracto acuoso.

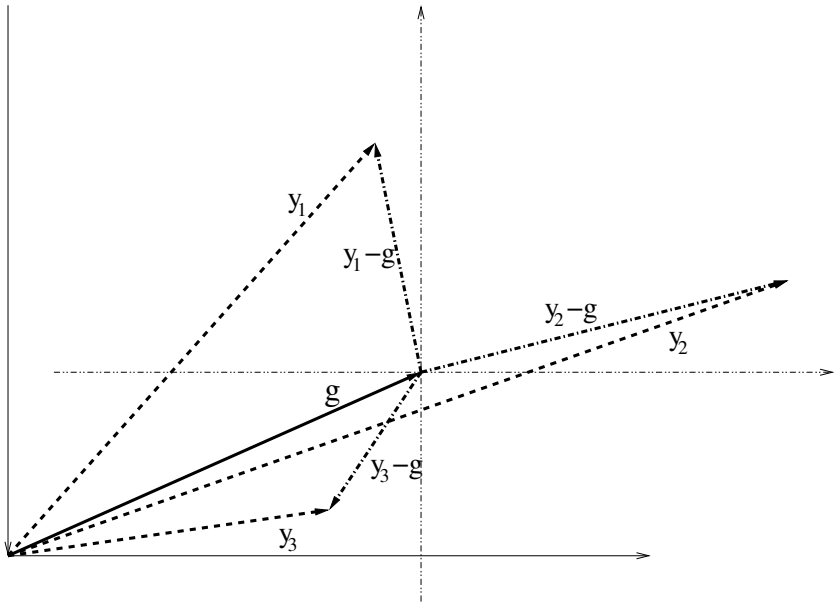
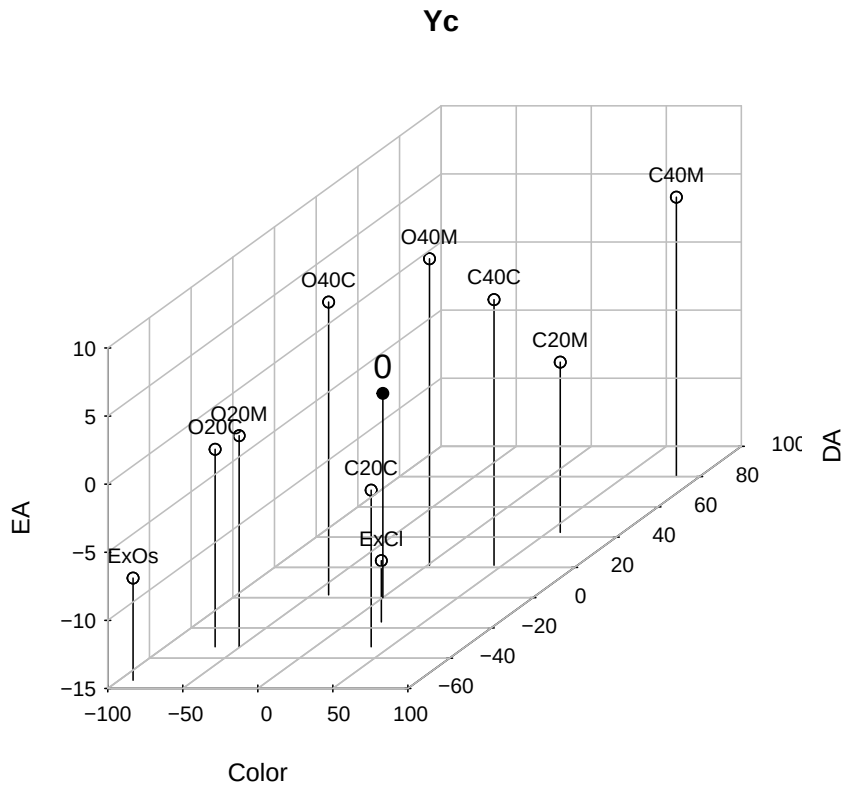


Figura 3.2. Centrado de los individuos en ACP: para representar los puntos $\mathbf{y}_i - \mathbf{g}$, el cero del sistema de coordenadas se traslada a $\mathbf{g}$

Los valores para el café típico son 276.70, 401.16 y 35.50, respectivamente. Esas son las coordenadas del centro de gravedad en la representación sin centrar. Esta información hay que guardarla porque al centrar los datos se pierde.

Centrar los datos, graficarlos en 3D e imprimir la matriz $\mathbf{Y}_c$ (figura 3.3)

```
par(las=1);
n<-nrow(Y); unos<-rep(1,n); # vector de n unos
Yc<-Y-unos %*%t(g);Yc;
# grafica de datos centrados
Yc3D<-scatterplot3d(Yc,main="Yc",type="h",color="black",
                box=FALSE,las=1);
Yc3D$points3d(Yc,pch=1);
addgrids3d(Yc,grid=c("xy", "xz", "yz"));
text (Yc3D$xyz.convert(Yc),labels=rownames(Yc),cex=0.8,
   col="black",pos=3);
Yc3D$points3d(t(c(0 ,0 ,0)),pch=19 , col ="black",type = "h");
text(Yc3D$xyz.convert (t(c(0 ,0 ,0))),labels ="0",pos =3,
     col ="black",cex =1.3);
```



Cafe	Color	DA	EA
ExCl	21.3	-16.1	-10.5
C40M	84.3	80.1	5.5
C40C	44.3	21.4	4.5
C20M	58.3	43.1	-2.5
C20C	37.3	-32.5	-3.5
ExOs	-90.7	-54.6	-7.5
O40M	1.3	21.4	7.5
O40C	-38.7	1.8	6.5
O20M	-50.7	-32.5	0.5
O20C	-66.7	-32.5	-0.5

Figura 3.3. Representación de la tabla de datos centrados del ejemplo Café en 3D. Visualmente esta figura es igual a la figura 3.1, pero cambian las coordenadas sobre los ejes, cuyos valores están en las tablas. Las dos gráficas están representando también las distancias entre los cafés, que aparecen en la tabla 3.1

3.2.3. Distancia entre individuos

El parecido entre individuos en la tabla de datos se traslada a la representación geométrica como una distancia (figura 3.4).

La distancia euclidiana al cuadrado entre dos individuos es:

$$d^2(i,l) = \sum_{j=1}^{p} (y_{ij} - y_{lj})^2 \tag{3.4}$$

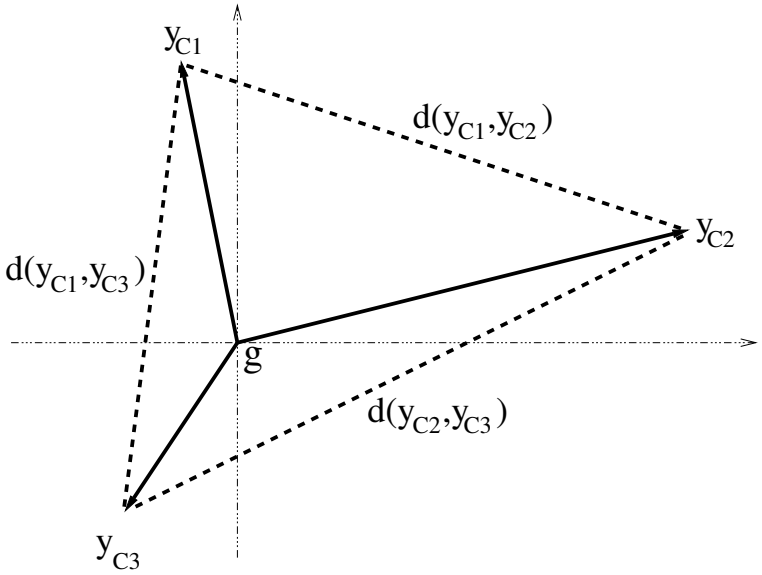


Figura 3.4. Distancias entre individuos

Las distancias son iguales si se calculan a partir de la matriz de los datos originales o de los centrados.

En la tabla 3.1 aparecen las distancias entre los diez cafés.

Una distancia de 0 indicaría que los dos cafés tienen los mismos valores para las variables.

En la figura 3.3, la pareja de cafés más alejados son ExOs y C40M, la distancia entre ellos es 221, la mayor en la tabla. Los más próximos son O20C y O20M, con una distancia de 16.

Tabla 3.1. Distancias entre cafés

	ExCl	C40M	C40C	C20M	C20C	ExOs	O40M	O40C	O20M
C40M	116								
C40C	46	71							
C20M	70	46	27						
C20C	24	122	55	78					
ExOs	118	221	155	178	130				
O40M	46	102	43	62	66	120			
O40C	65	146	85	106	84	78	45		
O20M	75	176	109	133	88	46	75	37	
O20C	90	188	123	146	104	33	87	45	16

```
round(as.dist(dist(Y)),0)
```

3.2.4. Inercia de la nube de individuos N_n

La noción física de *momento de inercia alrededor del centro de gravedad* se utiliza como medida de dispersión de la nube de puntos y se denomina *inercia*.

La inercia de la nube de individuos es:

$$Inercia(N_n) = \sum_{i=1}^{n} p_i d^2(i, \mathbf{g}) \tag{3.5}$$

En el caso de pesos iguales para todos los individuos $p_i = 1/n$, la inercia, calculada a partir de los datos centrados, es:

$$Inercia(N_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} y_{Cij}^2 = \sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n} y_{Cij}^2 = \sum_{j=1}^{p}\sigma_j^2 \tag{3.6}$$

Al centrar los datos, la información que queda en la nube de puntos es su forma, la cual se logra describir con el ACP que descompone la inercia en nuevos ejes de coordenadas. La fórmula 3.6 permite ver que la inercia es la suma de las varianzas de las variables, por lo que estas influyen en el análisis en proporción a su varianza.

Las varianzas dependen de las unidades de medida de las variables. Por lo tanto, al cambiar la escala cambia su varianza. La influencia de esas unidades de medida se elimina con la operación de reducido, que consiste en dividir cada columna de la matriz de datos centrados por la desviación estándar de la variable correspondiente.

3.2.5. Reducción de la nube de puntos

La matriz de varianzas y covarianzas $\mathbf{V}$ asociada a la tabla $\mathbf{Y}$ es: $\mathbf{V} = \frac{1}{n}\mathbf{Y}'_{\mathbf{C}}\mathbf{Y}_{\mathbf{C}}$. En la diagonal de $\mathbf{V}$ se tienen las varianzas, de modo que la suma de varianzas es igual a $traza(\mathbf{V})$.

La matriz normalizada $\mathbf{X}$, matriz de datos $\mathbf{Y}$ centrada y reducida, tiene término general:

$$x_{ij} = \frac{y_{ij} - \bar{y}_j}{\sigma_j}$$

donde $\bar{y}_j$ y σ_j son respectivamente la media y la desviación estándar de la variable j. En términos matriciales $\mathbf{X}$ se obtiene mediante:

$$\mathbf{X} = \mathbf{Y}_{\mathbf{C}}\mathbf{D}_{\sigma}^{-\mathbf{1}} \tag{3.7}$$

donde $\mathbf{D}_{\sigma} = diag(\sigma_j)$.

El valor que un individuo asume para una variable es la diferencia con respecto al promedio, pero medida en el número de desviaciones estándar.

Al reducir los datos, la información de las varianzas de las variables se pierde en las gráficas, pero los programas que realizan ACP los reportan en sus salidas. En el ejemplo "Café": $\sigma_{color} = 55.7$, $\sigma_{DA} = 39.5$ y $\sigma_{EA} = 5.8$.

Es claro que los datos iniciales se pueden recuperar a partir de los datos centrados y reducidos (estandarizados), si disponemos de los valores de las medias y varianzas (o desviaciones estándar).

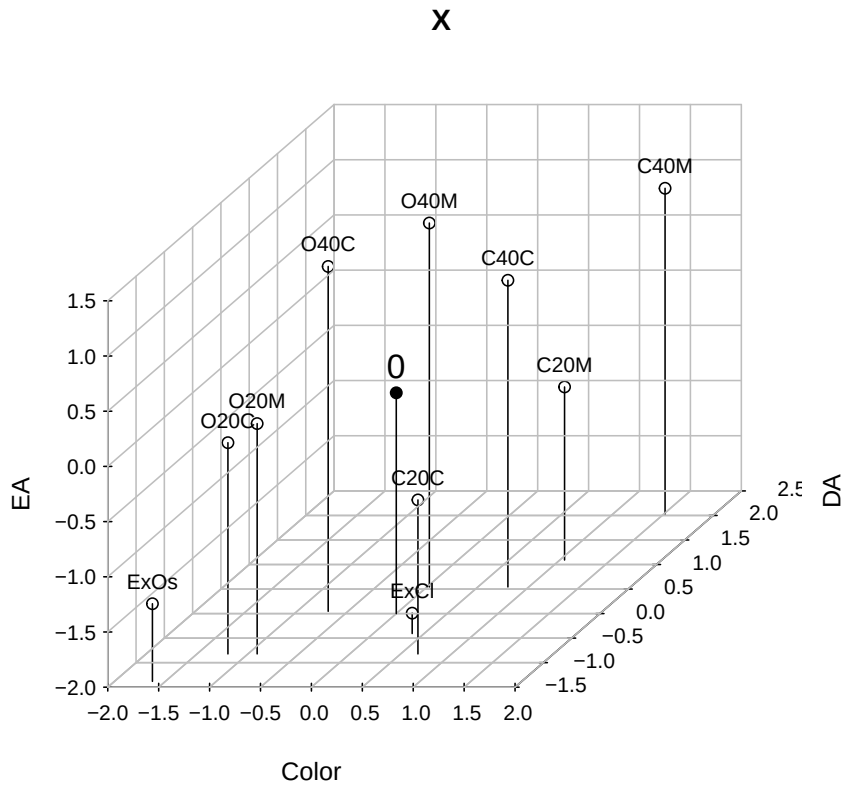
El análisis en componentes principales que se realiza casi todas las veces se denomina *normado* y se hace con la matriz $\mathbf{X}$, que contiene los datos estandarizados, es decir, centrados y reducidos.

En la figura 3.5 se muestra la gráfica 3D para el ejemplo "Café", junto con los valores de $\mathbf{X}$, y en la tabla 3.2 aparecen las distancias entre cafés. Nótese que se conserva un orden similar al de los datos sin reducir. Por ejemplo, los cafés más cercanos siguen siendo O20C y O20M, con una distancia de 0.3; y los más alejados, ExOs y CM40, con una distancia de 4.9.

La matriz de correlaciones de las variables iniciales, registradas en la tabla $\mathbf{Y}$, es la matriz de varianzas y covarianzas de $\mathbf{X}$:

$$\mathbf{V}_X = \frac{1}{n}\mathbf{X}'\mathbf{X}$$

Para el ejemplo, la matriz de correlaciones se puede ver en la figura 3.12, abajo a la izquierda.



Coordenadas de cafés estandarizados

Café	Color	DA	EA
ExCl	0.38	−0.41	−1.82
C40M	1.51	2.03	0.95
C40C	0.79	0.54	0.78
C20M	1.05	1.09	−0.43
C20C	0.67	−0.82	−0.61
ExOs	−1.63	−1.38	−1.30
O40M	0.02	0.54	1.30
O40C	−0.69	0.05	1.12
O20M	−0.91	−0.82	0.09
O20C	−1.20	−0.82	−0.09

Figura 3.5. Nube de individuos asociada a los datos estandarizados del ejemplo "Café". Las coordenadas sobre los ejes representan el número de desviaciones estándar que el café se desvía de la media de la respectiva variable

Código para obtener **X** a partir de los datos centrados, hacer la gráfica 3D e imprimir la matriz (figura 3.5)

```
par(las=1) # etiquetas de los dos ejes sean horizontales;
V<-t(Yc) %*% as.matrix(Yc)/n; V # = var(Y)*(n-1)/n;
Dsigma<-diag(sqrt(diag(V)));round(diag(Dsigma),1);
X<-as.matrix(Yc) %*% solve(Dsigma); colnames(X) <- colnames(Y);
X3D<-scatterplot3d(X,main="X",type ="h",box=FALSE);
X3D$points3d(Yc,pch=1);
addgrids3d(X,grid=c("xy","xz","yz"));
text (X3D$xyz.convert(X),labels=rownames(X),cex=0.8,pos=3);
X3D$points3d (t(c(0,0,0)),pch=19,col="black",type="h");
text(X3D$xyz.convert(t(c(0,0,0))),labels="0",pos=3,col="black"
  ,cex=0.8);
xtable (X,digits=rep (1,4)) # tabla para LaTeX;
```

Cuando los datos están estandarizados, la inercia de la nube de puntos es igual al número de variables, ya que cada una de ellas contribuye con 1 a la inercia total. Esto implica que la inercia, en el ACP normado deja de tener significado estadístico porque no depende de los valores de la tabla que se está analizando, sino del número de variables que contenga. En el ejemplo la inercia es 3.

Tabla 3.2. Distancias entre cafés cuando los datos están estandarizados (centrados y reducidos)

	ExCl	C40M	C40C	C20M	C20C	ExOs	O40M	O40C	O20M
C40M	3.7								
C40C	2.6	1.6							
C20M	2.0	1.6	1.3						
C20C	1.2	3.2	1.8	1.9					
ExOs	2.2	4.9	3.5	3.6	2.3				
O40M	3.1	2.0	0.9	2.0	2.3	3.4			
O40C	3.0	2.8	1.5	2.4	2.2	2.8	0.8		
O20M	2.2	3.6	2.2	2.6	1.6	1.6	1.9	1.3	
O20C	2.3	3.9	2.4	2.8	1.8	1.3	2.2	1.5	0.3

```
X<-scale(Y); round(as.dist(dist(X)),1)
```

3.2.6. Búsqueda de nuevos ejes: cambio de base

El objetivo geométrico de los métodos en ejes principales es buscar un nuevo sistema de ejes, de tal manera que la mayoría de la inercia se concentre en

los primeros ejes. Es decir, se trata de descomponer la inercia de la nube de puntos en ejes ortogonales ordenados, de modo que en el primer eje esté la mayor inercia posible; en el segundo, la mayor inercia residual posible, etc. Se busca primero el eje de máxima inercia proyectada. Si se denota $\mathbf{u}$ al vector unitario que da la dirección del eje, la coordenada de un vector individuo $\mathbf{x_i}$ sobre el eje es el producto punto (figura 3.6):

$$< \mathbf{x}_i, \mathbf{u} >= \mathbf{x}_i'\mathbf{u}$$

y la contribución a la inercia del individuo i sobre el eje $\mathbf{u}$ es $\frac{1}{n}(\mathbf{x}_i'\mathbf{u})^2$.

La inercia total de la nube de individuos, proyectada sobre el eje $\mathbf{u}$ es entonces:

$$\sum_{i=1}^{n} \frac{1}{n}(\mathbf{x}_i'\mathbf{u})^2 = \frac{1}{n}(\mathbf{Xu})'\mathbf{Xu} = \mathbf{u}'\frac{1}{n}\mathbf{X}'\mathbf{Xu} \tag{3.8}$$

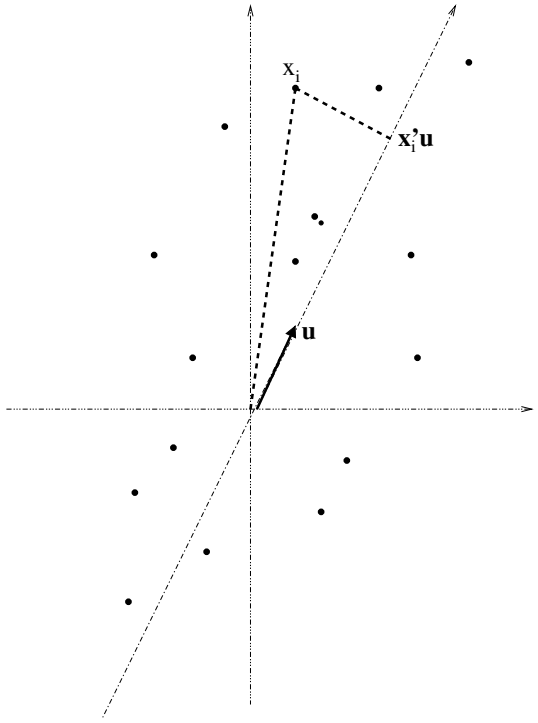


Figura 3.6. Proyección sobre la recta generada por $\mathbf{u}$. Se busca la dirección de la recta que tenga la mayor suma de los cuadrados de las proyecciones sobre ella.

Encontrar el eje de mayor inercia proyectada equivale a encontrar la dirección $\mathbf{u}$ que maximice (3.8) sujeto a la restricción $\mathbf{u}'\mathbf{u} = 1$. Una manera de resolver este problema es introduciendo un multiplicador de Lagrange λ. Entonces se debe maximizar:

$$f(\mathbf{u}) = \mathbf{u}'\frac{1}{n}\mathbf{X}'\mathbf{X}\mathbf{u} - \lambda(\mathbf{u}'\mathbf{u} - 1) \tag{3.9}$$

Los puntos críticos, en este caso, los puntos máximos de la función, son la solución de:

$$f'(\mathbf{u}) = 2\frac{1}{n}\mathbf{X}'\mathbf{X}\mathbf{u} - 2\lambda\mathbf{u} = \mathbf{0}$$

Es decir:

$$\frac{1}{n}\mathbf{X}'\mathbf{X}\mathbf{u} = \lambda\mathbf{u} \tag{3.10}$$

Las soluciones de (3.10) son los vectores propios unitarios asociados a los valores propios de $\frac{1}{n}\mathbf{X}'\mathbf{X}$. ¿Cuál de los p valores propios escoger? Premultiplicando (3.10) por $\mathbf{u}'$ se obtiene la cantidad que se quiere maximizar (3.8):

$$\mathbf{u}'\frac{1}{n}\mathbf{X}'\mathbf{X}\mathbf{u} = \lambda\mathbf{u}'\mathbf{u} = \lambda$$

y entonces las soluciones son los dos vectores propios unitarios asociados al valor propio mayor de $\frac{1}{n}\mathbf{X}'\mathbf{X}$. El primer valor propio se denota λ_1 y el vector propio unitario asociado que se escoja se nombra $\mathbf{u}_1$. El vector $-\mathbf{u}_1$ también es solución y las coordenadas sobre la recta generada por este vector son de signo contrario a las que se obtienen sobre $\mathbf{u}_1$.

Las coordenadas de los individuos sobre el eje generado por $\mathbf{u}_1$, denominado *primer eje principal*, se denotan por $\mathbf{F}_1$ y son: $\mathbf{F}_1 = \mathbf{X}\mathbf{u}_1$

Para obtener el mejor plano de proyección de la nube de puntos se busca un segundo eje generado por un vector unitario $\mathbf{u}$ ortogonal a $\mathbf{u}_1$ y que maximice la inercia (3.8). El problema ahora es maximizar $\mathbf{u}'\frac{1}{n}\mathbf{X}'\mathbf{X}\mathbf{u}$ sujeto a las restricciones $\mathbf{u}'\mathbf{u} = 1$ y $\mathbf{u}'\mathbf{u_1} = 0$. Entonces se introducen dos multiplicadores de Lagrange y la función a maximizar es:

$$f(\mathbf{u}) = \mathbf{u}'\frac{1}{n}\mathbf{X}'\mathbf{X}\mathbf{u} - \lambda(\mathbf{u}'\mathbf{u} - 1) - \mu(\mathbf{u}'\mathbf{u_1})$$

que tiene como primera derivada:

$$f'(\mathbf{u}) = 2\frac{1}{n}\mathbf{X}'\mathbf{X}\mathbf{u} - 2\lambda\mathbf{u} - \mu\mathbf{u}_1 \tag{3.11}$$

Los puntos críticos se obtienen igualando (3.11) a $\mathbf{0}$. Entonces el segundo multiplicador μ debe ser 0, lo que se puede ver premultiplicando (3.11) por $\mathbf{u}_1'$. Así se obtiene de nuevo la ecuación (3.10) y la solución es ahora el vector propio, notado $\mathbf{u}_2$, asociado al segundo valor propio más grande λ_2.

Encontrar un subespacio 3D para proyectar la nube de puntos es, por el mismo procedimiento, introducir un tercer eje ortogonal a los dos primeros, que es el tercer vector propio $\mathbf{u}_3$ asociado al tercer valor propio más grande λ_3 de $\frac{1}{n}\mathbf{X}'\mathbf{X}$.

El rango r de $\frac{1}{n}\mathbf{X}'\mathbf{X}$ es el número de vectores columna linealmente independientes de $\mathbf{X}$ y da el número de valores propios diferentes de cero, generalmente $r = p$, si $n > p$, más filas que columnas. Los p vectores propios $\{\mathbf{u_1}, \cdots, \mathbf{u_s}, \cdots, \mathbf{u_p}\}$ constituyen una base ortonormal para el espacio de los "individuos", con las propiedades que permiten obtener las mejores proyecciones de la nube de puntos. Es decir, el mejor eje para proyectar los puntos es el eje 1, generado por $\mathbf{u_1}$, y las coordenadas sobre él se constituyen en los valores del mejor índice de ordenamiento que se puede lograr. El mejor plano de proyección es el generado por los ejes 1 y 2.

Las n coordenadas de los individuos sobre un eje factorial s, $\mathbf{F}_s$, constituyen los valores de una variable nueva denominada *componente principal*. Su varianza es:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{F}_s(i))^2 = \frac{1}{n}\mathbf{F}_s'\mathbf{F}_s = \mathbf{u}_s'\frac{1}{n}\mathbf{X}'\mathbf{X}\mathbf{u}_s = \lambda_s$$

La obtención de valores y vectores propios es un problema básico de álgebra lineal, pero para su cálculo es necesario utilizar métodos numéricos. Los programas de cálculo matemáticos y estadísticos incluyen funciones para obtener los valores y vectores propios de una matriz. En R se calculan con la función `eigen`.

En el ACP del ejemplo "Café", los valores propios son (tomados de la tabla 3.3): $\lambda_1 = 2.067$, $\lambda_2 = 0.822$, y $\lambda_3 = 0.111$; y los vectores propios: $\mathbf{u_1} = (0.58 \quad 0.67 \quad 0.46)'$, $\mathbf{u_2} = (-0.57 \quad -0.07 \quad 0.82)'$ y $\mathbf{u_3} = (0.58 \quad -0.74 \quad 0.35)'$; o sus opuestos, ya que hay dos soluciones $\mathbf{u_1}$ y $-\mathbf{u_1}$, etc.

El primer plano factorial recoge $2.067 + 0.822 = 2.889$ de inercia, que es el $2.889 * 100/3 = 96.3\,\%$. Es decir que, casi nada se pierde al leer el primer plano factorial, en lugar de la representación en 3D, pero en cambio la lectura se hace mucho más fácil.

Tabla 3.3. Obtención de los valores y vectores propios del ACP del ejemplo "Cafe"

```
des<-eigen(V);des # calculo de valores y vectores propios
lambda<-des$values
U<-des$vectors #matriz con vectores propios en columnas
rownames(U)<-rownames(V)
colnames(U)<-c("Eje1","Eje2","Eje3");round(U,3)
lambda; U
## [1] 2.0670307 0.8216466 0.1113227
##            Eje1         Eje2        Eje3
## Color 0.5794934 -0.57140813  0.5811025
## DA    0.6728898 -0.06680772 -0.7367197
## EA    0.4597898  0.81794222  0.3457801
```

En el ejemplo "Café", la primera componente principal es una variable nueva, que resume las tres propiedades físicas y cuya expresión es:

$$F_1 = 0.58X_{color} + 0.67X_{DA} + 0.46X_{EA}$$

$$F_1 = 0.58\left(\frac{Y_{color} - \overline{Y}_{color}}{\sigma_{color}}\right) + 0.67\left(\frac{Y_{DA} - \overline{Y}_{DA}}{\sigma_{DA}}\right) + 0.46\left(\frac{Y_{EA} - \overline{Y}_{EA}}{\sigma_{EA}}\right)$$

$$F_1 = 0.58\left(\frac{Color - 276.7}{55.7}\right) + 0.67\left(\frac{DA - 401.2}{39.5}\right) + 0.46\left(\frac{EA - 35.5}{5.8}\right)$$

$$F_1 = 0.0104Color + 0.0170DA + 0.0795EA - 12.5$$

Para el café excelso claro (`ExCl  298 385.1 25`) el valor es:

$$F_1(ExCl) = 0.0104 * 298 + 0.0170 * 385.1 + 0.0795 * 25 - 12.5 = -0.87$$

F_1 es un índice que permite ordenar las diez preparaciones de cafés. En la figura 3.7 se puede ver el orden de los cafés de izquierda a derecha. En R se puede obtener como se muestra a continuación:

```
F <- X %*% U
round(sort(F[,1]),2)
#  ExOs  O20C  O20M  ExCl  C20C  O40C  O40M  C20M  C40C  C40M
# -2.47 -1.29 -1.04 -0.89 -0.44  0.15  0.98  1.14  1.18  2.68
# La diferencia entre -0.89 y -0.87 del cafe ExCl
# se debe a errores de redondeo.
```

Sentido de los ejes. Cada eje factorial se puede generar por uno de los dos vectores propios normados que definen su dirección $\mathbf{u_s}$ o $-\mathbf{u_s}$. El significado del sentido de los ejes se busca a partir de las variables, ya que el signo de

las coordenadas depende del vector propio seleccionado. Esto implica que para un mismo análisis se pueden tener planos rotados, según el paquete y el procedimiento utilizado y que el analista puede cambiar el signo de todas las coordenadas sobre un eje cuando le convenga.

3.2.7. Gráficas y ayudas para su interpretación

El primer plano factorial se construye buscando las coordenadas de los individuos sobre los ejes 1 y 2. El vector de todas las coordenadas sobre un eje s se nota F_s y es: $F_s = \mathbf{X}u_s$. Si se arreglan los vectores propios como columnas en una matriz $\mathbf{U}$, la tabla de las coordenadas sobre los nuevos ejes es $\mathbf{F} = \mathbf{XU}$.

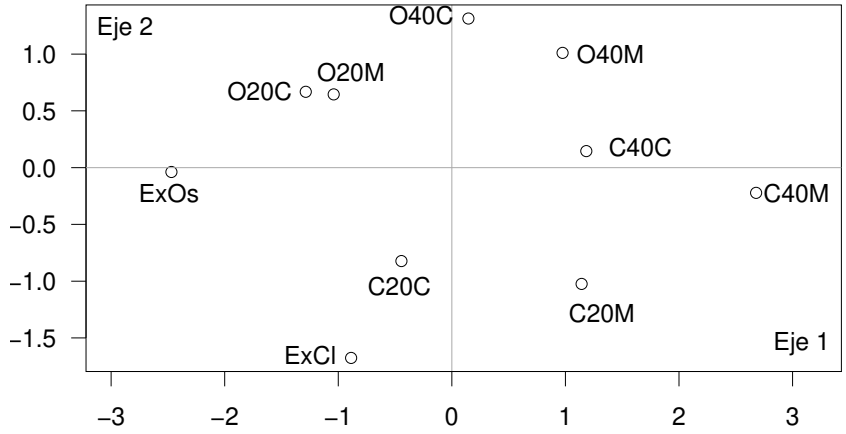
El primer plano factorial del ACP normado del ejemplo "Café" se muestra en la figura 3.7, donde se incluyen los valores de las coordenadas y ayudas a la interpretación. La lectura del plano se hace teniendo en cuenta los vectores propios: al lado positivo del primer eje se sitúan los cafés con mayores valores en las tres variables; al lado positivo del segundo eje, los de mayor valor en EA y al lado negativo los de mayor valor en *Color*.

Código para obtener el primer plano factorial y la tabla que se muestra en la figura 3.7

```
F <- X %*% U; round(F,2) #coordenadas sobre los nuevos ejes
plot(F[,1:2],las=1,asp=1)  # plano 12
text(F[,1:2],label=rownames(F),col="black",pos=2) # etiquetas
abline(h=0,v=0,col="darkgrey")   # ejes
rowSums(F^2)->d2;d2 # distancias
1/n*F^2%*%diag(1/lambda)*100->cont #contribuciones
F^2/d2*100->cos2 #cosenos cuadrados
#tabla de ayudas para la interpretacion
Ayu<-cbind(dis2=d2,F1=F[,1],F2=F[,2],cont1=cont[,1],cont2=
     cont[,2],    cos21=cos2[,1],cos22=cos2[,2],
     cosp=rowSums(cos2[,1:2]))
round(Ayu,2)  # ayudas en consola
xtable(Ayu,digits=rep(2,9))  # salida para LaTeX
```

3.2.7.1. Distancia al origen

La distancia de un punto al origen en el espacio completo es un buen complemento en la lectura de los ejes factoriales y está dada por la norma del vector-individuo en $\mathbb{R}^p$. En las salidas de algunos programas se presenta la distancia al cuadrado: $d^2(i, \mathbf{g}) = d^2(i, \mathbf{0}) = \|\mathbf{x}_i\|^2$



Coordenadas y ayudas a la interpretación

		Coordenadas		Contribuciones		Cosenos^2		
Café	d^2	F1	F2	F1	F2	F1	F2	Plano
ExCl	3.61	-0.89	-1.68	3.80	34.19	21.80	77.89	99.69
C40M	7.31	2.68	-0.22	34.72	0.60	98.20	0.67	98.87
C40C	1.53	1.18	0.15	6.78	0.26	91.45	1.39	92.84
C20M	2.47	1.14	-1.02	6.31	12.77	52.75	42.41	95.16
C20C	1.49	-0.44	-0.82	0.95	8.23	13.20	45.42	58.62
ExOs	6.24	-2.47	-0.04	29.49	0.02	97.70	0.02	97.73
O40M	1.98	0.98	1.01	4.60	12.44	48.09	51.72	99.81
O40C	1.75	0.15	1.31	0.10	20.98	1.21	98.65	99.86
O20M	1.51	-1.04	0.65	5.24	5.07	71.67	27.58	99.25
O20C	2.12	-1.29	0.67	8.00	5.43	78.22	21.10	99.32

Figura 3.7. Primer plano factorial del ACP normado del ejemplo "Café". En la columna d^2 se observan las distancias al origen en el espacio completo $\mathbb{R}^3$, figura 3.5. El más cercano es C20C y el más alejado C40M. Las coordenadas son las usadas para la gráfica. El café que más contribuye a la varianza del primer eje es C40M. En el plano están bien representados los diez cafés; en el primer eje el O40C está mal representado

3.2.7.2. Calidad de la representación o coseno cuadrado

Un plano factorial es una aproximación de la nube de puntos y como tal tendrá puntos bien y mal representados.

La calidad de la proyección o representación sobre un eje se mide con el coseno cuadrado, que se define para un punto como el cuadrado de la relación entre la norma de la proyección y la norma en el espacio completo (distancia del punto al origen).

El coseno cuadrado de la proyección de un punto sobre un plano es la suma de los cosenos cuadrados del punto sobre los ejes que conforman el plano.

La suma de los cosenos cuadrados de las proyecciones de un punto sobre todos los ejes factoriales es 1.

$$Cos_s^2(i) = \frac{F_s^2(i)}{\|\mathbf{x}_i\|^2}; \quad \sum_s Cos_s^2(i) = 1$$

Por ejemplo el coseno cuadrado del café C40M sobre el primer eje es:

$$Cos_1^2(C40M) = \frac{2.68^2}{7.31} = 0.98$$

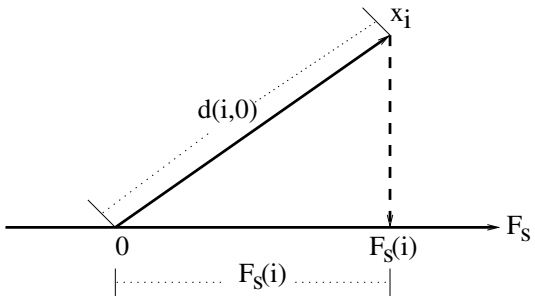


Figura 3.8. Calidad de la proyección sobre un eje *s*: cociente al cuadrado entre $F_s(i)$ y $d(i, 0)$. Cuando se acerca a uno, la longitud de la proyección se aproxima a la distancia original; si se acerca a cero, la proyección conserva muy poco de la distancia original

3.2.7.3. Contribución absoluta

La varianza o inercia proyectada sobre un eje *s* es:

$$Inercia_s(N_n) = \sum_{i=1}^{n} p_i F_s^2(i) = \lambda_s$$

Si los pesos de los individuos son iguales $p_i = \frac{1}{n}$.

Cada sumando es la contribución de un individuo a la inercia proyectada (varianza) sobre el eje *s*.

Para conocer los individuos que más influyen sobre la dirección de un eje factorial se utiliza el cociente de la contribución a la inercia del individuo

sobre la inercia total del eje (valor propio), que se denomina *contribución absoluta* $Ca_s(i)$:

$$Ca_s(i) = \frac{p_i F_s^2(i)}{\lambda_s}$$

La suma de las contribuciones de todos los individuos es: $\sum_i Ca_s(i) = 1$

Por ejemplo, la contribución del café C40M a la varianza del primer eje es: $Ca_1(C40M) = \frac{1}{10} * \frac{2.68^2}{2.067} = 0.347 = 34.7\,\%$. En la figura 3.7, abajo, se encuentran los valores de las ayudas a la interpretación para el ejemplo "Café".

3.2.8. Individuos ilustrativos o suplementarios

En el espacio de los individuos se pueden proyectar individuos nuevos para relacionarlos con los que participaron en el análisis. Cuando en un ACP se encuentran individuos atípicos se puede repetir el análisis sin ellos y luego proyectarlos como ilustrativos. De esta manera no influyen en la conformación de los ejes, pero se observa su posición con respecto a los individuos activos. El cálculo de las coordenadas se hace realizando las mismas transformaciones que para los individuos activos y proyectándolas sobre la recta generada por el vector propio $\mathbf{u}_s$. Es decir $F_s(i^+) = \mathbf{x}'_{i^+}\mathbf{u}_s$ (el signo + para indicar que es un individuo suplementario).

En el ejemplo "Café", se prepararon tazas con dos cafés comerciales y se les hicieron las mismas mediciones de las tazas originadas en el diseño experimental. La posición de los dos cafés comerciales permite ver su relación con los del diseño experimental (figura 3.9): el comercial 2 se situó muy cerca del café excelso claro y el comercial 1, entre los cafés excelsos y los que tienen menos agregados de granos. Con esto se pueden describir los cafés comerciales como de buena calidad. Su proyección se hace realizando sobre sus vectores las mismas transformaciones que para los cafés activos: centrado y reducido utilizando la media y varianza de los cafés activos y su proyección.

3.2.9. Variables cualitativas ilustrativas

Una variable cualitativa de K categorías establece una partición del conjunto de individuos en K clases o grupos. Los centros de gravedad de las clases se pueden proyectar como ilustrativos, sobre los ejes factoriales obtenidos, utilizando las mismas transformaciones y la misma fórmula de proyección.

Sin embargo, esas proyecciones son equivalentes a los centros de gravedad de las coordenadas factoriales de cada uno de los grupos.

En el ejemplo "Café" vamos a proyectar cada categoría de la variable tipo de contaminación: excelso (sin contaminación), con cebada y con maíz (figura 3.9).

3.2.9.1. Valores test para categorías de las variables cualitativas suplementarias

Para los centros de gravedad de las categorías se puede calcular la distancia al origen y los cosenos cuadrados, pero las contribuciones absolutas son cero porque no participan en la obtención de los ejes factoriales.

Adicionalmente se utiliza el valor test, que se definió en la sección 2.2.2, pues se trata aquí de la descripción de una variable continua (el componente principal s) y una variable cualitativa. La media de un componente principal F_s es cero porque es centrado y su varianza es el valor propio λ_s. Entonces el valor test $t_s(k)$ para una categoría k, asumida por n_k individuos, se obtiene de la fórmula (2.2):

$$t_s(k) = \sqrt{\frac{(n-1)n_k}{(n-n_k)\lambda_s}}\, F_s(k) \tag{3.12}$$

3.3. La nube de variables N_p

La nube de variables está constituida por p puntos en $\mathbb{R}^n$. Las coordenadas de cada punto son las columnas de la matriz $\mathbf{Y}$. Las estadísticas de resumen –medias, varianzas, covarianzas, correlaciones– tienen significado geométrico en el espacio de las variables. Las transformaciones de la matriz $\mathbf{Y}$ presentadas en el espacio de los individuos –operaciones de centrado y reducido– tienen otro significado en este espacio.

3.3.1. Significado de la media y del centrado en $\mathbb{R}^n$

Sea $\mathbf{Y}_j$ el vector columna asociado de la variable j –es decir, la columna j de la matriz $\mathbf{Y}$– y sean $\mathbf{Y}_{Cj}$ y $\mathbf{X_j}$ las columnas j de las matrices $\mathbf{Y_C}$ y $\mathbf{X}$, respectivamente.

Código para calcular las coordenadas factoriales de los dos cafés comerciales y su proyección sobre el primer plano factorial (3.9)

```
comer<-as.matrix(cafe[11:12,1:3]);comer
comc<-comer-rep(1,2)%*%t(g);comc # centrado
comcr <- comc%*%solve(Dsigma) # reducido
colnames(comcr)<-colnames(comer); comcr
Fsup <- comcr%*%U; Fsup
# primer plano factorial
plot(F[,1:2],las=1,asp=1)
text(F[,1:2],label=rownames(F),col="black",pos=1)
abline(h=0,v=0,col="darkgrey")
points(Fsup,col="black",pch=20)  # cafes comerciales
text(Fsup,labels=c("Com1","Com2"),col="darkgreen",pos=2)
```

Código para calcular las coordenadas de las categorías de *contaminación* y proyectarlas sobre el primer plano factorial (3.9)

```
conta<-factor(c("exce","maiz","ceba","maiz","ceba","exce",
       "maiz","ceba","maiz","ceba"))
centroids(F,conta)$centroids->Fconta; Fconta
points(Fconta,col="brown",pch=20)
text(Fconta,col="brown",labels=rownames(Fconta),pos=2)
```

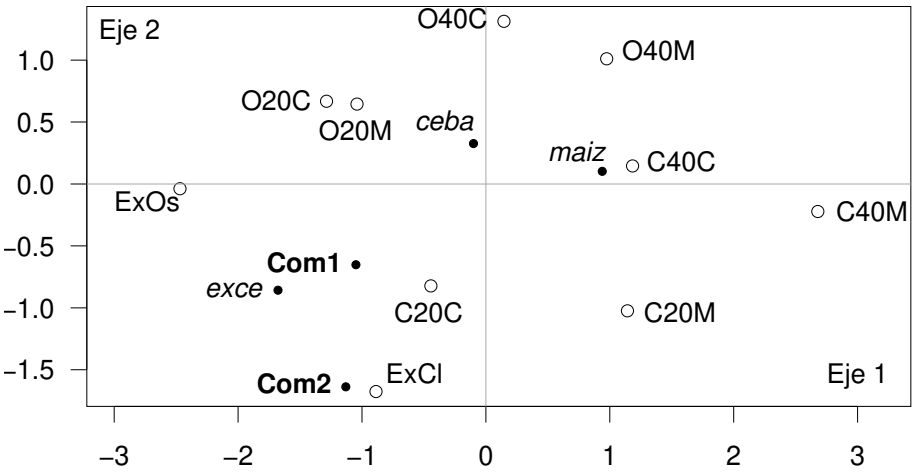


Figura 3.9. Primer plano factorial del ACP del ejemplo "Café", que muestra dos cafés comerciales y las categorías del tipo de contaminante. Los cafés comerciales se ponen en el marco de referencia del experimento y su posición permite decir que son de buena calidad (están cerca de los cafés no contaminados). La posiciones de los centros de gravedad –*excelso*, *cebada* y *maiz*–, muestran que el maíz afecta más la calidad del café

Utilizando en este espacio el producto interno definido mediante la matriz $\frac{1}{n}\mathbf{I}_n$, la norma del vector de n unos $\mathbf{1}_n$ es 1 y las medidas estadísticas adquieren significado geométrico:

$$\|\mathbf{1}_n\|^2 = \langle \mathbf{1}_n, \mathbf{1}_n \rangle_{\frac{1}{n}\mathbf{I}_n} = \frac{1}{n}\sum_{i=1}^{n} 1 = \frac{1}{n}n = 1$$

3.3.1.1. Significado de la media de una variable *j*

En la ecuación (3.13) se puede observar que la media $\bar{Y}_j$ es la coordenada de la proyección de la variable sobre la primera bisectriz, es decir, la recta generada por el vector $\mathbf{1}_n$ (figura 3.10).

$$\bar{Y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} = \frac{1}{n}\mathbf{Y'}_j\mathbf{1}_\mathbf{n} = \left\langle \mathbf{Y}_j, \mathbf{1}_\mathbf{n} \right\rangle_{\frac{1}{n}\mathbf{I}_n} \tag{3.13}$$

Se puede definir un vector que repite el valor de la media n veces: $\bar{\mathbf{Y}}_j = \bar{Y}_j\mathbf{1}_\mathbf{n}$.

3.3.1.2. Significado del centrado de una variable

El centrado de un vector $\mathbf{Y_j}$ se logra mediante: $\mathbf{Y}_{\mathbf{C}j}$=$\mathbf{Y}_j$ - $\bar{\mathbf{Y}}_j$. Entonces una variable centrada es la proyección de la variable sobre el subespacio ortogonal a la primera bisectriz, lo que implica que en el proceso de centrado se pierde una dimensión (figura 3.10).

3.3.2. Significado de las varianzas y covarianzas

En (3.14) se observa que la varianza de una variable j igual a la norma al cuadrado del vector variable centrado y, por lo tanto, la desviación estándar de j es su norma: $\sigma_j = \|\mathbf{Y}_{\mathbf{C}j}\|_{\frac{1}{n}\mathbf{I}_n}$.

$$var(Y_j) = \frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \bar{Y}_j)^2 = \left\langle \mathbf{Y}_{\mathbf{C}j}, \mathbf{Y}_{\mathbf{C}j} \right\rangle_{\frac{1}{n}\mathbf{I}_n} \tag{3.14}$$

La covarianza entre dos variables Y_j y Y_k es el producto punto entre los dos vectores que representan a las variables centradas:

$$cov(Y_j, Y_k) = \frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \bar{Y}_j)(y_{ik} - \bar{Y}_k) = \left\langle \mathbf{Y}_{\mathbf{C}j}, \mathbf{Y}_{\mathbf{C}k} \right\rangle_{\frac{1}{n}\mathbf{I}_n} \tag{3.15}$$

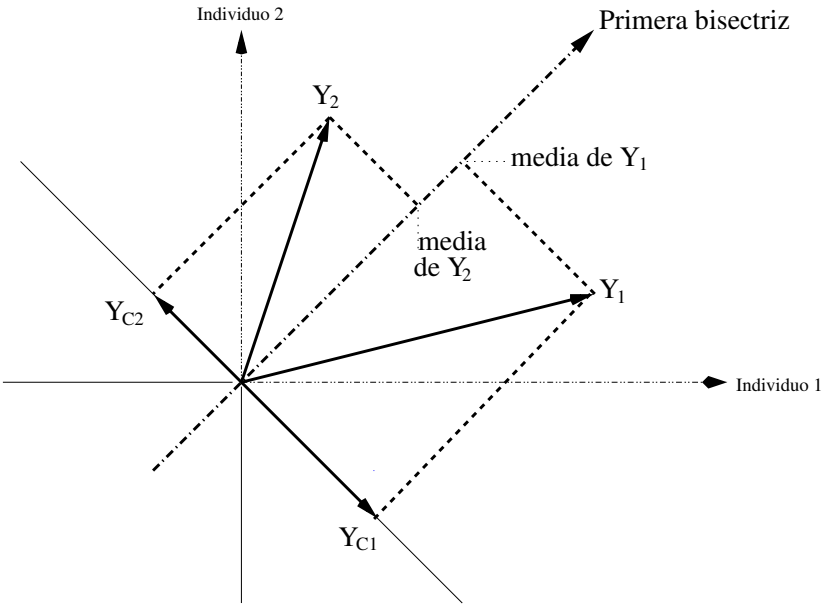


Figura 3.10. Significado geométrico de las medias y del centrado de las variables: el valor de la media de una variable es la coordenada de su proyección sobre la primera bisectriz. El vector centrado de la variable está contenido en el subespacio $\mathbb{R}^{n-1}$ ortogonal a la primera bisectriz. En está gráfica en $\mathbb{R}^2$ ese subespacio es la recta que pasa por (-1,1).

3.3.3. Significado del reducido de una variable en $\mathbb{R}^n$

Una columna j de $\mathbf{X}$ (fórmula 3.7) se obtiene mediante $\mathbf{X_j} = \frac{1}{\sigma_j}\mathbf{Y}_{C_j}$. Reducir una variable es, entonces, multiplicarla por el inverso de su norma, y el vector variable reducido queda con norma 1.

La varianza de una variable centrada y reducida es:

$$< \mathbf{X_j}, \mathbf{X_j} >_{\frac{1}{n}\mathbf{I_n}} = \mathbf{X_j}' \left(\frac{1}{n}\mathbf{I}_n\right) \mathbf{X}_j = \frac{1}{n}\mathbf{X_j}'\mathbf{X_j} = 1$$

y entonces, la representación de las variables estandarizadas se puede ver como flechas que terminan en el cascarón hiperesférico de radio 1 y centro en el origen.

3.3.4. Significado de la correlación entre dos variables

En (3.16) se ve que la correlación entre dos variables j y k es igual al coseno entre los dos vectores de las variables centradas y es también el coseno entre los dos vectores variables centradas y reducidas, puesto que $\|\mathbf{X_j}\|_{\frac{1}{n}\mathbf{I_n}} = 1$.

$$cor(Y_j, Y_k) = \frac{cov(Y_j, Y_k)}{\sigma_j \sigma_k} = \frac{< \mathbf{Y_{C_j}}, \mathbf{Y_{C_k}} >_{\frac{1}{n}\mathbf{I_n}}}{\|\mathbf{Y_{C_j}}\|_{\frac{1}{n}\mathbf{I_n}} \|\mathbf{Y_{C_k}}\|_{\frac{1}{n}\mathbf{I_n}}} =< \mathbf{X_j}, \mathbf{X_k} >_{\frac{1}{n}\mathbf{I_n}} \quad (3.16)$$

Entonces el espacio de las variables de un ACP normado es una representación de la matriz de correlaciones, porque la norma de todas las variables es uno y el coseno entre dos vectores variables es igual a la correlación entre estas.

Si dos vectores variables tienen ángulo pequeño, su correlación es alta; dos vectores variables ortogonales indican que las variables no están correlacionadas.

3.3.5. Inercia en el espacio de las variables

En el ACP centrado las variables se representan como flechas de longitud igual a la desviación estándar y con cosenos de los ángulos entre variables iguales a los coeficientes de correlación entre ellas.

La inercia en este espacio es:

$$Inercia(N_p) = \sum_{j=1}^{p} d^2(\mathbf{Y_{Cj}}, \mathbf{0}) = \sum_{j=1}^{p} Var(Y_j) \quad (3.17)$$

La contribución de una variable a la inercia es su varianza. En el caso del ACP normado cada variable contribuye con 1 a la inercia y la inercia total es igual al número de variables.

3.3.6. Búsqueda de los nuevos ejes

La proyección de una variable $\mathbf{X}_j$ sobre un eje $\mathbf{v}$ en $\mathbb{R}^n$ es:

$$\mathbf{X'}_j \left(\frac{1}{n}\mathbf{I_n}\right)\mathbf{v} = \frac{1}{n}\mathbf{X'}_j\mathbf{v}$$

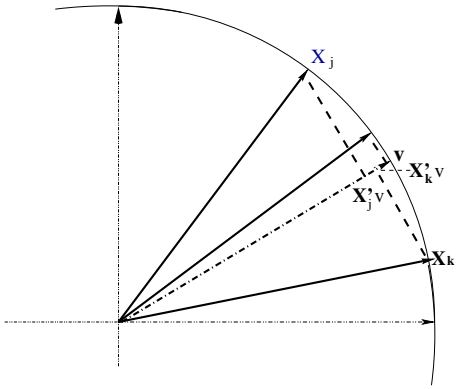


Figura 3.11. Proyección de variables sobre el eje generado por **v**. La dirección de **v** es la que maximiza la suma de cuadrados de las proyecciones de los vectores variables sobre **v**

La inercia de las p variables proyectadas sobre el eje **v** es:

$$\sum_{j=1}^{p}\left(\frac{1}{n}\mathbf{X}'_j\mathbf{v}\right)^2 = \frac{1}{n^2}\mathbf{v}'\mathbf{X}\mathbf{X}'\mathbf{v} = \frac{1}{n}\mathbf{v}'\frac{1}{n}\mathbf{X}\mathbf{X}'\mathbf{v} \tag{3.18}$$

El eje de mayor inercia proyectada se encuentra maximizando (3.18) sujeto a la restricción $\mathbf{v}'\left(\frac{1}{n}\mathbf{I_n}\right)\mathbf{v} = 1$:

$$f(\mathbf{v}) = \frac{1}{n}\mathbf{v}'\frac{1}{n}\mathbf{X}\mathbf{X}'\mathbf{v} - \mu\left(\frac{1}{n}\mathbf{v}'\mathbf{v} - 1\right)$$

Derivando con respecto al vector **v** e igualando a 0:

$$f'(\mathbf{v}) = \frac{2}{n^2}\mathbf{X}\mathbf{X}'\mathbf{v} - \frac{2\mu}{n}\mathbf{v} = 0$$

Se obtiene:

$$\frac{1}{n}\mathbf{X}\mathbf{X}'\mathbf{v} = \mu\mathbf{v} \tag{3.19}$$

La ecuación (3.19) corresponde a la expresión de valores y vectores propios de la matriz $\frac{1}{n}\mathbf{X}\mathbf{X}'$ y, por lo tanto, la solución está dada por uno de los dos vectores **v** asociados al valor propio más grande μ de la matriz $\frac{1}{n}\mathbf{X}\mathbf{X}'$, que se notan $\mathbf{v}_1$ y μ_1, respectivamente. Sin embargo, los valores propios de $\frac{1}{n}\mathbf{X}\mathbf{X}'$ que son mayores que cero, son iguales a los de $\frac{1}{n}\mathbf{X}'\mathbf{X}$, es decir, $\mu_1 = \lambda_1$.

Se buscan los ejes sucesivos ortogonales entre sí y corresponden a los vectores propios, $\frac{1}{n}\mathbf{I_n}$ unitarios, asociados a los valores propios, ordenados de mayor a menor, de la matriz $\frac{1}{n}\mathbf{XX'}$.

3.3.7. Círculo de correlaciones y ayudas a la interpretación

Un plano factorial de las variables estandarizadas se denomina *círculo de correlaciones* (figura 3.12) ya que es la proyección de la *hiperesfera* de correlaciones, flechas que parten del origen y tienen longitud 1. La longitud de la proyección, sin error, de un vector variable es 1. La calidad de la representación en el plano se observa al dibujar un círculo de radio uno en el plano factorial.

El vector de coordenadas sobre un eje s, $\mathbf{G_s}$, se obtiene mediante $\frac{1}{n}\mathbf{X_j'v_s}$ y coincide con la correlación entre la variable j y el eje s. Los valores de la variable sintética representada por el eje s están en el vector F_s, que contiene las coordenadas de los individuos sobre el eje factorial s. La contribución de cada variable a un eje s sirve para seleccionar las variables que dan más significado al eje. En la figura 3.12 se muestran la esfera, el círculo, la matriz de correlaciones, las coordenadas y las ayudas para la interpretación de las variables en el ejemplo "Café".

El primer eje es un "factor tamaño" porque está correlacionado positivamente con las tres variables. Los cafés con valores altos de sus coordenadas sobre el eje tienen valores altos en las tres variables. La correlación negativa de la *Nota* de apreciación de los catadores con el primer eje significa que los mejores cafés están al lado negativo del primer eje, es decir, que valores mayores de las tres variables dañan la calidad apreciada del café.

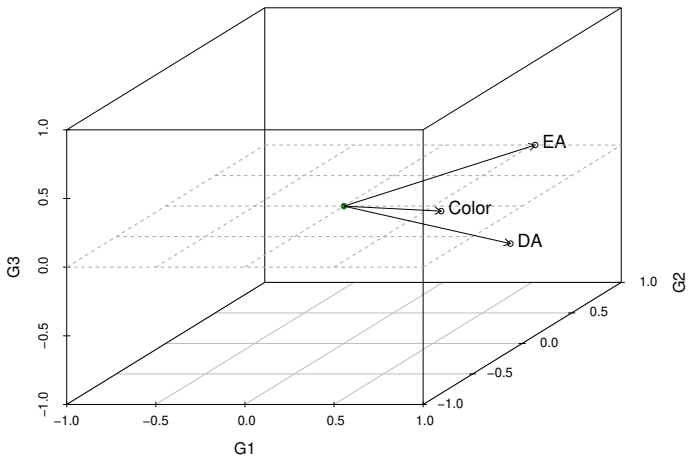
El segundo eje muestra correlación positiva con EA y negativa con *Color*, los cafés con coordenadas positivas tienen mayores valores de extracto acuoso y los de coordenadas negativas, mayores valores de color.

3.4. Relación entre los espacios de individuos y variables

Las propiedades que se presentan a continuación son fáciles de demostrar. También se pueden ver, por ejemplo, en Lebart *et al.* (2006).

1. La matriz $\frac{1}{n}\mathbf{XX'}$ tiene p valores propios, que son iguales a los valores propios de $\frac{1}{n}\mathbf{X'X}$ y los restantes $n - p$ valores propios son 0.

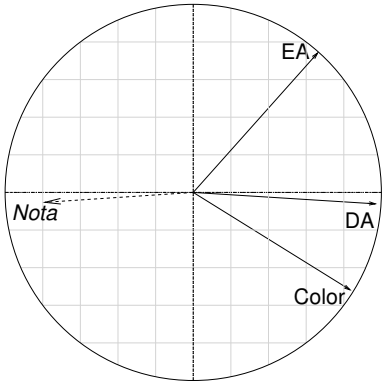
Esfera de correlaciones



Correlaciones

	Color	DA	EA
Color	1.00	0.79	0.19
DA	0.79	1.00	0.57
EA	0.19	0.57	1.00

Círculo de correlaciones



Coordenadas y ayudas a la interpretación

	Coordenadas		Contribución		Cosenos^2		
Variable	G1	G2	G1	G2	G1	G2	Plano
Color	0.83	−0.52	33.58	32.65	69.41	26.83	96.24
DA	0.97	−0.06	45.28	0.45	93.59	0.37	93.96
EA	0.66	0.74	21.14	66.90	43.70	54.97	98.67

Figura 3.12. Esfera y círculo de correlaciones del ejemplo "Café", mostrando la variable *Nota* como ilustrativa. El círculo es una imagen de la matriz de correlaciones y permite la interpretación de los ejes factoriales de los cafés

Código para obtener la esfera y el círculo de correlaciones (figura 3.12), calculando las coordenadas con la relación de transición $\mathbf{G}_s = \sqrt{\lambda_s}\mathbf{u}_s$

```
G<-U%*%diag(sqrt(lambda)); G # G <- cor(Y,F)
colnames(G)<-c("G1","G2","G3");
G3D <- scatterplot3d(G,main="G",xlim=c(-1,1),ylim=c(-1,1),
      zlim=c(-1,1))
coord <- G3D$xyz.convert(G)
text(coord,labels=rownames(G), cex=0.8,col="black",pos=4)
G3D$plane(0,0,0,col="darkgrey")
G3D$points3d(t(c(0,0,0)),pch=19,col="black")
cero <- G3D$xyz.convert(0, 0, 0)
for (eje in 1:3) {
 arrows(cero$x, cero$y, coord$x[eje], coord$y[eje], lwd = 2,
   length = 0.1)
}
#dev.print(device=xfig,file="cafeEspera.fig")#grafica en xfig
s.corcircle(G,clabel=2);
# proyeccion de nota como variable ilustrativa
Nota <- cafe[1:10,16]; Nota;
Fnota <- cor(Nota,F);Fnota;
arrows(0,0,Fnota[1],Fnota[2],col="black",angle=10,lty=2);
text(Fnota,"Nota",col="black",pos=1,cex=2,font=3);
#dev.print(device = xfig,file="cafeCirculo")
# grabar circulo en xfig
```

2. El vector coordenadas de los n individuos sobre el eje s $\mathbf{F}_s$, es un vector propio de $\frac{1}{n}\mathbf{XX}'$.

3. La varianza de $\mathbf{F}_s$ es λ_s, y por lo tanto, el vector propio $\mathbf{v}_s$ se puede calcular mediante: $\mathbf{v}_s = \frac{1}{\sqrt{\lambda_s}}\mathbf{F}_s$.

4. $\mathbf{G}_s$, vector de coordenadas de las p variables sobre el eje s, es un vector propio de $\frac{1}{n}\mathbf{X}'\mathbf{X}$.

5. La varianza de $\mathbf{G}_s$ es λ_s, y por lo tanto, se puede obtener mediante: $\mathbf{G}_s = \sqrt{\lambda_s}\mathbf{u}_s$.

6. En el ACP normado, las coordenadas de $\mathbf{G}_s$ son las correlaciones entre las variables y el eje s: $cor(Y_j, F_s)$.

3.4.1. Variables continuas como ilustrativas

Sobre el círculo de correlaciones de un ACP normado se pueden proyectar variables que no participaron en el análisis. Por ejemplo, cuando se tiene un

puntaje global como suma o promedio de varios puntajes, conviene proyectar como ilustrativo el puntaje global sobre el ACP de los otros puntajes.

En el ejemplo "Café" se proyecta la nota de impresión global dada por un panel de catadores para explorar su relación con las tres variables físicas en conjunto. La correlación con el primer eje es alta y en sentido opuesto (-0.79), lo que significa que valores altos en estas tres propiedades físicas indican detrimento de la calidad apreciada de las tazas de café. Esto explica la ubicación de los cafés excelso del lado negativo del eje 1 (figura 3.9).

3.5. ACP con los paquetes *ade4* y *FactoClass*

El ACP se puede llevar a cabo en casi todos los programas de estadística, tanto comerciales como libres. En R existen varias funciones para realizarlo. En este texto se utilizan funciones de los paquetes `ade4` y `FactoClass`. En la tabla 3.4 se presentan las principales funciones para realizar un ACP en el orden que suele utilizarse en un análisis; algunas son de las librerías básicas del R. A continuación se incluye el código para llevar a cabo el ACP normado del ejemplo "Café". El lector puede ver los resultados en la consola de R y comparar con los presentados en este capítulo.

Código para realizar el ACP normado de las variables físicas del ejemplo "Café"

```
library(FactoClass); # carga de paquetes
data(cafe);          #hacer los datos disponibles
# ACP normado con variables fisicas y reteniendo dos ejes
acp<-dudi.pca(cafe[1:10,1:3],scannf=FALSE);
acp$cent            # medias de las variables:
round(acp$norm,2);   # desviacion estandar de las variables
inertia(acp);        # valores propios y porcentajes
barplot(acp$eig);    # histograma de valores propios
round(acp$c1,3);     # vectores propios
```

Círculo de correlaciones

```
s.corcircle(acp$co);
# proyeccion de la variable Impresion como ilustrativa
(cor(cafe[1:10,16],acp$li)->coimpre);
s.arrow(coimpre,label="Impresion",add.plot = TRUE,boxes=FALSE)
# coordenadas de las variables = correlaciones con los ejes
round(acp$co,3);
# ayudas para la interpretacion de las variables
inertia(acp,,TRUE);
```

Tabla 3.4. Funciones para realizar un ACP con *ade4* y *FactoClass*

Función	Librería	Descripción
`dudi.pca`	`ade4`	ACP, por defecto normado.
`inertia.dudi`	`ade4`	Ayudas para la interpretación, por defecto solo valores propios.
`s.corcircle`	`ade4`	Círculos de correlaciones.
`plotcc`	`FactoClass`	Círculos de correlaciones con `ggplot2`.
`s.arrow`	`ade4`	Planos factoriales de las variables, en un ACP no normado.
`s.label`	`ade4`	Planos factoriales de los individuos.
`plot.dudi`	`FactoClass`	Planos factoriales de los individuos, usando el parámetro `Tcol=FALSE`.
`suprow`	`ade4`	Coordenadas de individuos suplementarios.
`points`	`graphics`	Proyectar puntos en un gráfico.
`text`	`graphics`	Poner etiquetas asociadas a los puntos en un gráfico.
`cor(acp$li,Xsup)`	`stats`	Correlaciones entre variables suplementarias y ejes (coordenadas en un ACP normado. Si `ACP` es el objeto de salida del ACP y `Xsup` es la tabla de variables suplementarias.
`arrows`	`graphics`	Proyectar variables ilustrativas sobre un círculo de correlaciones.
`supqual`	`FactoClass`	Coordenadas y ayudas a la interpretación de variables cualitativas ilustrativas.
`s.class`	`ade4`	Planos factoriales de los individuos destacando las clases de una partición definida por una variable cualitativa.

Primer plano factorial de los cafés

```
plot(acp,Tcol=FALSE,gg=TRUE);
round(acp$li,2);   # coordenadas de los cafes
inertia(acp,TRUE); # ayudas para la interpretacion de los cafes
```

Proyectar los cafés comerciales y la variable *contaminante* como ilustrativa

```
suprow(acp,cafe[11:12,1:3])$lisup -> lcom;
plot(acp,Tcol=FALSE); points(lcom,col="darkgreen");
text(lcom,rownames(lcom),col="darkgreen",pos=2);
# proyeccion variables conta (contaminacion) como ilustrativa
conta<-factor(c("exce","maiz","ceba","maiz","ceba","exce",
    "maiz","ceba","maiz","ceba"));
supqual(acp,conta)->supconta;
points(supconta$coor,col="red",pch=20);
text(supconta$coor,rownames(supconta$coor),col="red",pos=2);
round(supconta$tv,2); # valores test de categorias de conta
```

Código para destacar las clases de la variable *contaminante* en el primer plano factorial

```
plot(acp,Tcol=FALSE)
s.class(acp$li,conta,col=c("orange","darkgreen","red"),
  add.plot = TRUE,cellipse = 0,clabel=0.5)
```

3.6. Ejemplo de aplicación de ACP: resultados del examen de admisión a las carreras de la Facultad de Ciencias

Se presenta un ejemplo sencillo para mostrar el procedimiento del ACP, la interpretación de sus resultados, y el uso de paquetes de R para llevarlo a cabo.

Para el primer semestre de 2013 fueron admitidos 445 estudiantes a las carreras de la Facultad de Ciencias. El examen de admisión tuvo cinco componentes temáticos: matemático, científico, social, textual e imagen.

3.6.1. Objetivos del análisis

Se realiza un ACP con el objeto de: 1) validar el puntaje total que es el resumen de los cinco puntajes, 2) explorar relaciones entre los resultados y las carreras a las que fueron admitidos los estudiantes, y 3) explorar relaciones entre algunas variables sociodemográficas y los resultados del examen.

3.6.2. Resultados de análisis

Para cumplir con esos objetivos se realiza un ACP normado utilizando como variables activas los puntajes obtenidos por los admitidos en los cinco componentes del examen y el puntaje total como variable ilustrativa.

Como variables cualitativas suplementarias se proyectan las carreras y las características sociodemográficas presentes en los datos.

3.6.2.1. Número de ejes a analizar

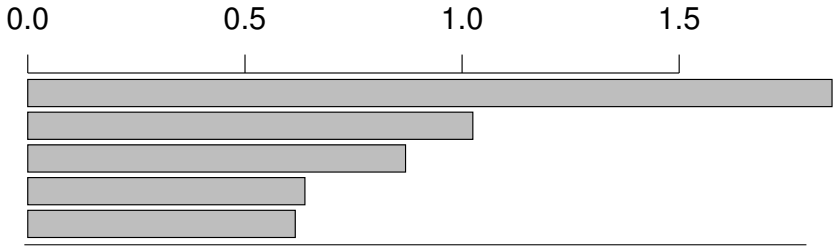
El número de ejes a retener para el análisis es la primera decisión en un ACP. Se toma con varios criterios orientadores: el primero es la forma del "histograma de valores propios"; el segundo, los ejes que corresponden a

valores propios mayores que 1, en el caso del ACP normado, y finalmente, la selección de un eje adicional si se considera que suministra información importante que no se ha visto en los anteriores.

En la figura 3.13 se muestra el histograma de valores propios con los valores numéricos en la parte inferior. Tres valores propios se destacan, pero solo dos son mayores que 1. Seguramente dos son suficientes, pero se debe verificar si el tercer eje permite alguna descripción adicional al primer plano factorial. El primer plano retiene el 57.5 % de la inercia y los tres primeros ejes, el 74.9 %.

Código para realizar ACP normado de las notas del examen y obtener la figura 3.13

```
library(FactoClass);
data(admi);names(admi);
Y<-admi[,2:6];names(Y);
acp<-dudi.pca(Y,scannf=FALSE,nf=3);
barplot(acp$eig); # histograma de valores propios
#dev.print(device = xfig,file="acpExaAdmi.fig");
valp<-t(inertia(acp)$tot.inertia); # valores propios
xtable(valp,digits=rep(3,6));
```



	1	2	3	4	5
Valor propio	1.852	1.025	0.870	0.638	0.616
Inercia acumulada	1.852	2.877	3.746	4.384	5.000
Proporción acum.	0.370	0.575	0.749	0.877	1.000

Figura 3.13. Valores propios del ACP de los resultados del examen de los admitidos: histograma y valores. El primer valor propio se destaca sobre los demás y retiene el 37.0 % de la inercia. El segundo valor propio es también mayor que 1 y el primer plano retiene el 57.5 % de la inercia. La forma del histograma sugiere retener uno o tres ejes para el análisis

3.6.2.2. Círculo de correlaciones

En la figura 3.14 se observa un primer eje de tamaño, que muestra alta correlación con el puntaje total y con los mejores puntajes de coordenadas negativas (lado izquierdo del eje).

El factor tamaño se presenta cuando todas las correlaciones entre las variables activas son positivas y se observan porque tienen coordenadas con el eje del mismo lado. El primer eje se muestra como otra manera de obtener un puntaje global ya que su correlación con el resultado del examen es de -0.985. Para tener la coordenada en el mismo sentido basta cambiarles el signo a todas las coordenadas sobre el primer eje.

El segundo eje contrapone los resultados en los componentes de imagen, matemático y científico versus social y textual. El tercer eje es inferior a uno, pero su valor es cercano al segundo y destaca la oposición entre los resultados en las pruebas de imagen (positivo) y científica (negativo).

3.6.2.3. Primer plano factorial del ACP de los admitidos

Los admitidos son anónimos en este ACP, pero las variables cualitativas permiten observar grupos de ellos, según las categorías que asuman.

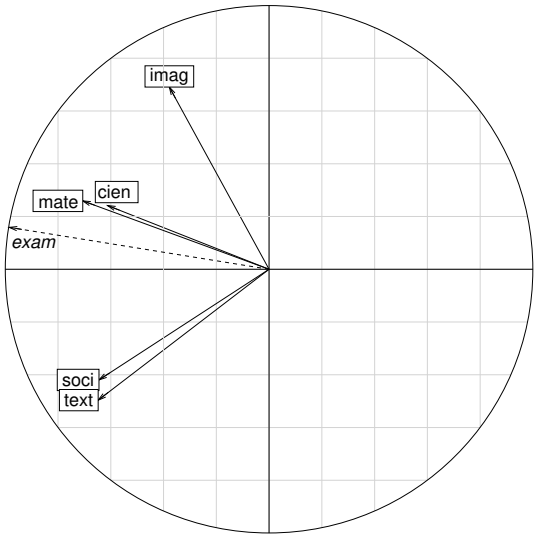
La figura 3.15 muestra el primer plano factorial de los individuos con las categorías de las variables cualitativas como ilustrativas. La estructura del plano está dada por los resultados del examen, de modo que cualquier ordenamiento de las categorías es indicio de alguna relación con esos resultados. En la tabla 3.5 se encuentran las ayudas para la interpretación de este plano y del tercer eje factorial. Obsérvese que las categorías de la variable estrato están ordenadas en el primer eje –bajo, medio y alto– lo cual indica que los admitidos de estratos más altos tienden a obtener mejores resultados en el examen. En cambio, la edad no está ordenada, y son los de diecisiete años quienes tienden a obtener mejores resultados.

En cuanto a género, los hombres obtienen en promedio mejores resultados que las mujeres. Según la variable origen de los admitidos, los de Bogotá en promedio obtienen los mejores resultados; y los de otros lugares los peores.

Las carreras con mejores resultados son Matemáticas, Física y Geología, en oposición a Estadística, Farmacia y Química. Los admitidos a Matemáticas en promedio tienen mejores resultados en las pruebas de imagen, matemática y científica.

Código para obtener el círculo de correlaciones de la figura 3.14

```
s.corcircle(acp$co);
# exam como ilustrativa
Gexam<-cor(admi$exam,acp$li);
rownames(Gexam)<-"exam";Gexam;
s.arrow(Gexam,add.plot=TRUE,boxes=FALSE);
#dev.print(device = xfig,
   file="acpExaAdmiCirculo.fig");
#coordenadas
xtable(acp$co,digits=rep(3,4));
xtable(Gexam,digits=rep(3,4));
```



Coordenadas de las variables

	Comp1	Comp2	Comp3
mate	-0.706	0.259	0.204
cien	-0.612	0.242	0.602
soci	-0.645	-0.420	-0.351
text	-0.648	-0.496	-0.105
imag	-0.378	0.691	-0.576
exam	-0.985	0.159	0.023

Figura 3.14. Círculo de correlaciones del ACP normado de los componentes del examen de los admitidos, con el puntaje total como ilustrativo. Se muestran las coordenadas de las variables sobre los ejes factoriales, que son también las correlaciones variable-componente principal

Son pocos los admitidos que tienen que nivelar lectoescritura (siLE), y por eso se ubican más lejos, al lado derecho arriba, donde se sitúan los de peores resultados en el examen de admisión.

Los que no tienen que nivelar de matemáticas están al lado izquierdo arriba, que son los admitidos con mejores resultados en el examen.

En el tercer eje (ver valores test en la tabla 3.5) se observa que en promedio los que tienen edades de dieciséis años o menos, por un lado, y los que vienen de otra región, por otro, tienen resultados inferiores en la componente de imagen.

Para obtener la figura 3.15

```
Ysupcat<-admi[,c(1,8:13)];
sup<-supqual(acp,Ysupcat);
plot(acp,Tcol=FALSE,ucal=100,cex.row=0.2,
  xlim=c(-1,1.5),ylim=c(-0.5,0.5));
points(sup$coor,col="black");
text(sup$coor,labels=rownames(sup$coor),
   col="black",pos=1,font=2);
#dev.print(device = xfig,
   file="acpExaAdmiCatSup.fig");
```

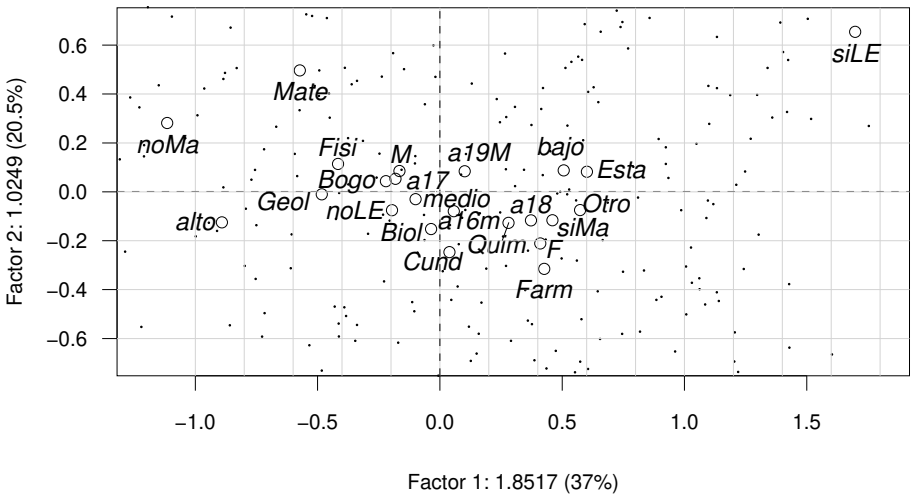


Figura 3.15. Primer plano factorial de los admitidos mostrando las variables cualitativas ilustrativas. Los puntos sin etiqueta corresponden a los admitidos

Tabla 3.5. Pesos, distancias^2 y coordenadas de las categorías suplementarias

Categoría	Peso	dis^2	Coordenadas		
			Eje1	Eje2	Eje3
Biol	0.142	0.117	-0.036	-0.152	-0.080
Esta	0.148	0.403	0.601	0.082	-0.109
Farm	0.164	0.346	0.427	-0.315	0.068
Fisi	0.184	0.214	-0.416	0.114	0.108
Geol	0.101	0.244	-0.483	-0.010	-0.071
Mate	0.119	0.751	-0.573	0.496	-0.014
Quim	0.142	0.106	0.280	-0.127	0.036
F	0.288	0.229	0.410	-0.211	0.117
M	0.712	0.037	-0.166	0.085	-0.047
bajo	0.402	0.267	0.506	0.088	0.035
medio	0.416	0.021	-0.099	-0.030	-0.068
alto	0.182	0.878	-0.891	-0.125	0.078
Bogo	0.699	0.040	-0.182	0.053	-0.062
Cund	0.085	0.080	0.039	-0.246	-0.025
Otro	0.216	0.382	0.573	-0.075	0.210
a16m	0.265	0.052	0.057	-0.079	0.195
a17	0.384	0.079	-0.220	0.044	-0.118
a18	0.126	0.162	0.372	-0.117	-0.068
a19M	0.225	0.059	0.102	0.084	0.011
siLE	0.103	3.636	1.698	0.654	0.388
noLE	0.897	0.048	-0.196	-0.075	-0.045
siMa	0.708	0.309	0.460	-0.116	-0.119
noMa	0.292	1.814	-1.115	0.281	0.287

```
xtable(cbind(wcat=sup$wcat,d2=sup$dis2),digits = rep(3,6))
```

Tabla 3.6. Contribuciones y cosenos^2 de las categorías suplementarias

Categoría	Valores test			Cosenos cuadrados		
	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3
Biol	-0.228	-1.287	-0.730	0.011	0.198	0.054
Esta	3.885	0.711	-1.024	0.897	0.017	0.029
Farm	2.928	-2.908	0.684	0.526	0.287	0.013
Fisi	-3.064	1.127	1.160	0.812	0.061	0.055
Geol	-2.507	-0.071	-0.540	0.956	0.000	0.021
Mate	-3.261	3.799	-0.114	0.437	0.328	0.000
Quim	1.763	-1.071	0.331	0.745	0.152	0.012
F	4.036	-2.793	1.675	0.733	0.194	0.059
M	-4.036	2.793	-1.675	0.733	0.194	0.059
bajo	6.430	1.496	0.642	0.961	0.029	0.005
medio	-1.300	-0.530	-1.292	0.468	0.043	0.217
alto	-6.512	-1.225	0.834	0.905	0.018	0.007
Bogo	-4.286	1.685	-2.129	0.828	0.071	0.096
Cund	0.186	-1.567	-0.174	0.019	0.755	0.008
Otro	4.654	-0.814	2.493	0.860	0.015	0.116
a16m	0.526	-0.989	2.649	0.062	0.121	0.738
a17	-2.697	0.717	-2.115	0.616	0.024	0.178
a18	2.188	-0.921	-0.585	0.855	0.084	0.029
a19M	0.847	0.942	0.128	0.176	0.120	0.002
siLE	8.927	4.624	2.973	0.793	0.118	0.041
noLE	-8.927	-4.624	-2.973	0.793	0.118	0.041
siMa	11.093	-3.763	-4.171	0.685	0.044	0.046
noMa	-11.093	3.763	4.171	0.685	0.044	0.046

```
xtable(cbind(sup$tv,sup$cos2),digits = rep(3,7))
```

3.6.3. Conclusiones del análisis

El primer eje del ACP realizado es un indicador que resume el rendimiento de los estudiantes admitidos a la Facultad de Ciencias en el primer semestre del 2013. El componente de matemáticas es el que más pesa en dicho indicador y el de imagen, el que menos. La correlación entre el indicador y la nota global del examen, suministrada por la Universidad, es de 0.985.

Por lo tanto, los ordenamientos obtenidos con el primer componente del ACP normado y la nota global son similares (figura 3.14).

Según las carreras, los admitidos a Matemáticas, Geología y Física son en promedio los de mejores resultados, mientras que los admitidos a Estadística, Farmacia y Química tienen en promedio resultados más bajos.

Los admitidos a Matemáticas tienen mejor promedio en la componente de imagen (figura 3.15, tablas 3.5 y 3.6).

Los estudiantes de estratos altos, diecisiete años, hombres y de origen bogotano son los que tienen en promedio mejores resultados en los componentes del examen, comparados con las otras categorías de sus respectivas variables: estrato, edad, género y origen.

3.7. Ejercicios

Algunos de los ejercicios están propuestos para los estudiantes que toman regularmente el curso de la Carrera de Estadística. Un lector diferente puede omitir, si lo desea, los ejercicios de demostraciones y utilizar, para llevar a cabo el ACP, programas de computador diferentes a los propuestos.

1. Para buscar un subespacio H de dimensión 1 que maximice la suma de cuadrados de las distancias entre las proyecciones sobre H de todas las parejas de puntos (i, l) –cada punto está dotado de una masa p_i– demuestre que:

$$Max_{(H)}\left\{\sum_i \sum_l d_H^2(i,l)\right\} = Max_{(H)}\left\{\sum_i d_H^2(i,\mathbf{g})\right\}$$

   donde $\mathbf{g}$ es el vector centro de gravedad de todos los puntos (Lebart *et al.*, 2006, pág. 64).

2. Muestre que el multiplicador de Lagrange μ en la fórmula 3.11 es 0.

3. Muestre que el rango máximo de la matriz $\frac{1}{n}\mathbf{X}'\mathbf{X}$ es p, cuando $p < n$.

4. Muestre que dos vectores propios $\mathbf{u}_s$ y $\mathbf{u}_t$, $s \neq t$, son ortogonales.

5. Muestre que las coordenadas sobre el eje factorial s (vector $\mathbf{F}_s$) están centradas.

6. Muestre que la varianza de una componente principal es igual al valor propio asociado λ_s.

7. Escriba la matriz de varianzas y covarianzas de las componentes principales F_1, F_2 y F_3, obtenidas con el ACP normado del ejemplo "Café" y describa sus propiedades.

8. Muestre que la norma del vector $\mathbf{1}_n$ en $\mathbb{R}^n$ con la métrica $\mathbf{M} = \frac{1}{n}\mathbf{I}_n$ es igual a 1.

9. Muestre que la media de la variable Y_j es la coordenada de la proyección del vector $\mathbf{Y}_j$ sobre la primera bisectriz, es decir, el subespacio generado por el vector $\mathbf{1}_n$.

10. Muestre que en el espacio de las variables $\mathbb{R}^n$, centrar una variable Y_j es proyectarla sobre el subespacio ortogonal a la primera bisectriz, con la métrica $\frac{1}{n}\mathbf{I}_n$.

11. Muestre que la operación de reducido en $\mathbb{R}^n$ consiste en multiplicar el vector variable centrado $\mathbf{Y}_{jC}$ por $\frac{1}{\sigma_j}$.

12. Describa el lugar geométrico de las variables centradas y reducidas en $\mathbb{R}^{n-1}$.

13. Demuestre que los valores propios mayores que cero, de los espacios de individuos y variables son iguales.

14. Demuestre que el vector F_s de todas las coordenadas de los n individuos sobre el eje s es un vector propio de la matriz $\frac{1}{n}\mathbf{X}\mathbf{X}'$.

15. Muestre que $G_s(j) = \sqrt{\lambda_s}\mathbf{u}_s$.

16. Muestre que la coordenada de una variable sobre un eje factorial, en el ACP normado, es igual al coeficiente de correlación entre la variable y el primer componente principal.

17. Demuestre que $\mathbf{X} = \sum_{s=1}^{p} \sqrt{\lambda_s}\, \mathbf{v}_s\mathbf{u}'_s$.

18. Sea $\mathbf{X}^*$ la mejor aproximación de $\mathbf{X}$ en el subespacio de dimensión S. Demuestre que la calidad de la aproximación τ_S es: $\tau_S = \frac{\sum_{s=1}^{S} \lambda_s}{\sum_{s=1}^{p} \lambda_s}$

19. Demuestre las relaciones de transición entre los dos espacios de representación: individuos y variables.

20. Muestre claramente el significado geométrico de las estadísticas (media, covarianza, desviación estándar, coeficiente de correlación) en el espacio de las variables de un ACP.

21. Muestre que en el espacio de las variables en el ACP normado la distancia entre dos variables está entre 0 y 2.

22. Obtenga la expresión del valor test de una categoría (3.12) a partir de la fórmula 2.2 de la página 25.

23. Para el ACP normado del ejemplo "Café" dibuje las cuatro versiones posibles del primer plano factorial cambiando el sentido de los ejes F_1 y F_2.

24. Calcule la distancia al cuadrado del café excelso claro al centro de la representación en $\mathbb{R}^3$.

25. Calcule las contribuciones a la inercia en $\mathbb{R}^3$ de cada uno de los diez cafés, en valor y porcentaje.

26. Calcule la contribución a la inercia del primer eje y el coseno cuadrado sobre el primer eje del café excelso claro (ver tabla de la figura 3.7).

27. En R cree un factor con el porcentaje de agregado de granos, cuyos niveles 0, 20 y 40; y proyecte los tres centros de gravedad sobre el primer plano factorial de los cafés.

28. Obtenga el valor test de cero agregado de granos a partir de (3.12).

29. Realice el ACP normado de las variables físicas de los cafés utilizando el paquete `ade4`. Incluya la proyección de la variable *Nota de impresión global* como cuantitativa ilustrativa, los cafés comerciales como individuos ilustrativos y el tipo de contaminación como variable cualitativa ilustrativa.

30. Realice el ACP normado del punto anterior utilizando el programa DtmVic, disponible en `http://www.dtmvic.com/`.

3.8. Talleres

Se recomienda el taller ACP gráfico para entender bien el significado geométrico del ACP. El taller "Whisky" es un ejemplo sencillo para consolidar el aprendizaje del ACP. En el taller "Lactantes" se proponen una serie de preguntas sencillas de resolver empleando R.

3.8.1. Análisis en componentes principales gráfico

Sea la matriz de datos (nótese que está transpuesta):

	1	2	3	4	5	6	7	8	9	10
Y_1	9	7	8	3	1	3	4	7	2	6
Y_2	9	13	6	1	5	11	4	3	8	10

Realice geométricamente sobre papel cuadriculado el ACP de $\mathbf{Y}$ (sin dividir por $n = 10$) ejecutando los pasos siguientes:

1. Diagrama de dispersión de $\mathbf{Y}$.
2. Calcule el centro de gravedad y la matriz de datos centrados $\mathbf{X}$.
3. Grafique la nube de puntos centrados.
4. Obtenga gráficamente los nuevos ejes F_1 y F_2, sobre la nube de individuos del numeral anterior. F_1 corresponde a la recta que pasa por el origen y está en la dirección más alargada de la nube de puntos; F_2 es la recta que pasa por el origen y es perpendicular a F_1.
5. Escriba la matriz con las nuevas coordenadas leyéndolas en la gráfica. Son las proyecciones de los puntos sobre F_1 y F_2 y se leen con una regla o escuadra (no realice cálculos).
6. Dibuje el plano factorial de los individuos. Es el plano con F_1 como eje horizontal y F_2 como vertical.
7. Obtenga las coordenadas de un vector $\mathbf{u}_1$ unitario (de norma 1) que esté sobre la recta F_1 en la gráfica del numeral 4. Se obtiene tomando cualquier vector sobre la recta, encontrando su norma, y entonces $\mathbf{u}_1$ es ese vector multiplicado por 1/norma. Obtenga visualmente $\mathbf{u}_2$ unitario que esté sobre la recta F_2 a partir de $\mathbf{u}_1$ (no realice cálculos).
8. Calcule la suma de cuadrados de las coordenadas sobre F_1 y sobre F_2 (obtenidas en el numeral 5). Se notan λ_1 y λ_2, respectivamente.

9. Obtenga los dos vectores de coordenadas de las variables sobre los dos primeros ejes factoriales, los cuales se notan G_1 y G_2, respectivamente, y se calculan así:

$$G_1 = \sqrt{\lambda_1}\,\mathbf{u}_1\ ;\quad G_2 = \sqrt{\lambda_2}\,\mathbf{u}_2$$

   La primera coordenada de cada uno de los vectores G_1 y G_2 corresponde a la variable X_1 y la segunda, a la variable X_2.

10. Dibuje el primer plano factorial de las variables, con G_1 como eje horizontal y G_2 como eje vertical.

3.8.2. ACP de "Whisky"

Objetivo

El objetivo es estudiar la relación calidad-precio de 35 marcas de whisky utilizando las variables precio (francos franceses), proporción de malta (%), vejez (añejamiento en años) y apreciación (nota promedio de un panel de catadores redondeada a entera). Se dispone además de una variable nominal *categoría*, que clasifica las marcas según su contenido de malta (1 = Bajo, 2 = Estándar, 3 = Puro malta) (Fine, 1996).

Preguntas

Realice primero un ACP no normado y luego un ACP normado y responda a las preguntas.

1. En el ACP no normado, analice la contribución de las variables a la inercia. ¿Realmente se puede considerar un análisis de las cuatro variables?

2. Analice la matriz de varianzas y covarianzas con la ayuda del primer plano factorial de las variables. Haga un resumen (interpretación del primer plano factorial de las variables).

3. Realice el ACP normado y justifique por qué es el que conviene para los objetivos de este taller.

4. ¿Cuántos ejes retiene para el análisis? ¿Por qué?

5. ¿Cuál es la variable que más contribuye al primer eje? ¿Cuál es la que menos? (indique los porcentajes).

6. Según el círculo de correlaciones, ¿cuáles son las variables más correlacionadas? ¿Cuánto es el valor de la correlación? ¿Sí corresponden a lo que se observa en la matriz de correlaciones?

7. ¿Cuál es la variable mejor representada en el primer plano factorial? ¿Cuál la peor? (Escriba los porcentajes).

8. ¿Qué representa el primer eje? ¿Qué nombre le asignaría? ¿Qué representa el segundo eje?

9. ¿Cuál es el individuo mejor representado en el primer plano factorial? Ubique sobre el gráfico de individuos al peor representado sobre el primer plano factorial (indique los porcentajes).

10. Supongamos que usted tiene una gráfica de individuos, donde no se muestran los antiguos ejes de las variables. ¿Cómo dibuja los ejes de apreciación y de precio? (Responda concretamente, es decir, con números).

11. ¿Qué características tienen las marcas de whisky según sus ubicaciones en el plano (a la derecha, a la izquierda, arriba, abajo)?

12. ¿Qué significa el círculo del primer plano factorial de variables? ¿Cómo lo dibujaría en una gráfica impresa donde no está? (Suponga que las escalas de los dos ejes son iguales).

13. A partir de la posición en el plano deduzca las características de las tres categorías de whisky (bajo, estándar y pura malta).

14. Supongamos que usted desea comprar una botella de whisky con buena apreciación y que no sea tan cara. Dé dos números de marcas que compraría. ¿Por qué? ¿Cuáles son las características de las dos marcas?

15. Seleccione dos marcas que definitivamente no compraría. ¿Por qué? ¿Qué características tienen?

16. Realice un resumen práctico del análisis suponiendo que lo va a entregar a una compañía que contrató el estudio. Debe dar respuesta al objetivo y apoyarse en las tablas y gráficas que crea necesarias.

3.8.3. ACP "Lactantes"

De Dalgaard (2008) se tomó el ejemplo *kfm-Breast-feeding data*, cuyos datos están en el objeto `kfm{ISwR}` y son una tabla de 50 filas (bebés de aproximadamente 2 meses) y 6 columnas (Dalgaard, 2020).

Las variables continuas son: leche = leche materna consumida por el niño: dl/24 horas; peso = peso del niño, kg; tetero = alimentación suplementaria, ml/24 horas; peso.madre, kg; talla.madre, cm. Se dispone de la variable categórica sexo (masculino, femenino). Se plantea realizar un ACP que responda a los objetivos siguientes:

1. Descripción de los bebés según su peso, consumo de leche materna y tetero, y su relación con el peso y talla de las madres.
2. ¿Está relacionada la alimentación suplementaria (tetero) con las demás variables?
3. ¿Hay diferencias entre niños y niñas?

Conteste a las siguientes preguntas:

1. ¿Por qué con el ACP se cumplen los objetivos planteados?
2. ¿Realiza un ACP normado o no normado? ¿Por qué?
3. Describa el bebé promedio según las cinco variables.
4. ¿Cuántos ejes retiene para el análisis? ¿Por qué?
5. ¿Qué variables se puede decir que están más altamente correlacionadas con el primer factor? ¿Puede darle algún significado a este primer factor?
6. ¿Puede identificar subconjuntos de variables altamente correlacionas entre sí? ¿Existe algún subconjunto de variables que se pueda decir que no está correlacionado con otro subconjunto de variables?
7. ¿Qué características tienen los lactantes según su posición en el primer plano factorial?
8. ¿Los análisis anteriores sugieren que pueden constituirse grupos de bebés? ¿Podría sugerir algunos?
9. ¿Se puede decir que hay diferencia entre niños y niñas en este análisis? ¿Cuáles son son esas diferencias?
10. Escriba un resumen práctico del análisis que satisfaga los objetivos planteados.

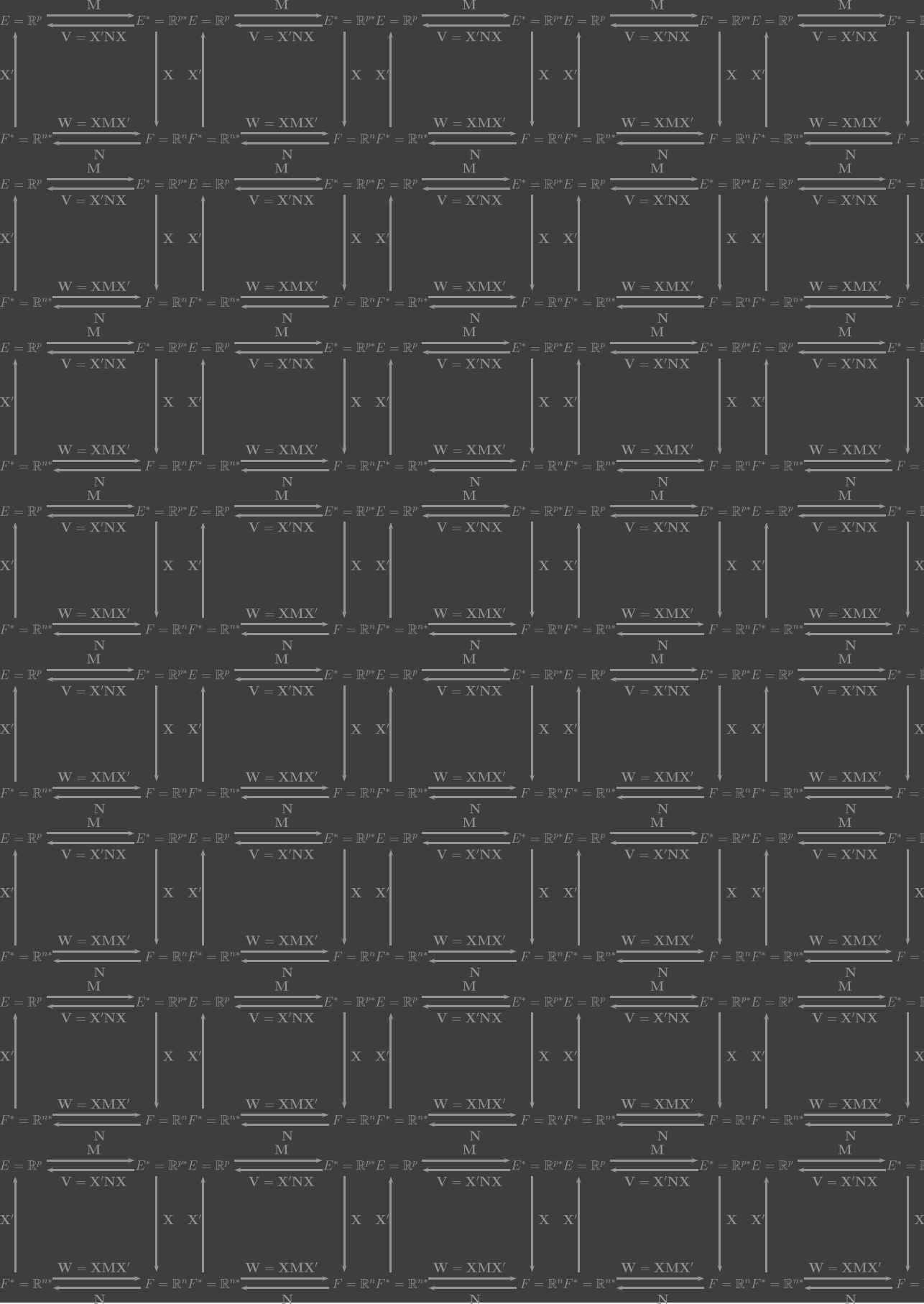
    Las preguntas que siguen son sobre lectura y algunos cálculos en el ejemplo "Lactantes". Responda a los ¿por qué? mencionando la manera cómo dedujo la respuesta: ayuda que utilizó, la gráfica que leyó, etc.

11. La inercia de las nubes de puntos asociadas al ACP es: _____
12. Primer valor propio: _____
13. Primer vector propio: _____
14. Porcentajes de la inercia en: primer eje _____, segundo eje _____ y primer plano factorial _____
15. Correlación entre *tetero* y primer factor: _____
16. Variable que más contribuye al primer eje: _____
    ¿Por qué? _____
17. ¿Las dos variables menos correlacionadas con tetero son: _____
    ¿Por qué? _____
18. Variable mejor representada en el primer plano factorial: _____
    ¿Por qué? _____
19. Coordenadas del bebé promedio sobre el primer plano factorial:
    _____
20. Los dos bebés que más tetero consumen son: _____
21. Para el bebé situado en el extremo superior del primer plano factorial escriba las coordenadas sobre los dos primeros ejes factoriales:
    _____
22. Calcule la contribución del bebé anterior a la inercia del segundo eje factorialy la calidad de representación sobre el mismo eje.
23. Escriba las coordenadas de los antiguos ejes unitarios de las variables *leche* y *tetero* sobre el primer plano factorial.
24. Dibuje los antiguos ejes de *leche* y *tetero* sobre el primer plano factorial, indicando los lados positivos y negativos.

Capítulo

cuatro

Análisis en componentes principales generalizado


M
V = X′NX
X′
X
W = XMX′
N

Cada uno de los métodos en ejes principales se puede ver como un ACP de una matriz $\mathbf{X}$ que contiene los datos a analizar transformados de acuerdo al respectivo método. En cada caso se definien las matrices de métrica y pesos en los dos espacios: de filas y de columnas.

Un producto interno está determinado por una matriz cuadrada, simétrica, definida positiva. En la geometría euclidiana canónica en $\mathbb{R}^p$, la matriz que define el producto interno es la identidad de dimensión p, notada $\mathbf{I}_p$. En este texto se utilizan los términos *matriz de métrica*, o simplemente *métrica*, para referirse en cada caso a una matriz que define un producto interno.

Sea $\mathbf{M}$ una matriz que define un producto interno en un espacio vectorial en los reales, E. A partir de esta matriz se dota al espacio de una geometría euclidiana. Sean $\mathbf{x}$ y $\mathbf{y}$ dos vectores en E, entonces:

- $\mathbf{M}$-producto punto: $< \mathbf{x}, \mathbf{y} >_{\mathbf{M}} = \mathbf{x}'\mathbf{M}\mathbf{y} = \mathbf{y}'\mathbf{M}\mathbf{x}$
- $\mathbf{M}$-norma: $\|\mathbf{x}\|_{\mathbf{M}} = \sqrt{< \mathbf{x}, \mathbf{x} >_{\mathbf{M}}}$
- $\mathbf{M}$-distancia: $d_{\mathbf{M}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_{\mathbf{M}}$
- Vector $\mathbf{M}$-unitario: $\|\mathbf{u}\|_{\mathbf{M}} = 1 \implies \mathbf{u}'\mathbf{M}\mathbf{u} = 1$
- $\mathbf{M}$-proyección sobre $\mathbf{u}$: $< \mathbf{x}, \mathbf{u} >_{\mathbf{M}} = \mathbf{x}'\mathbf{M}\mathbf{u}$
- $\mathbf{M}$-coseno: $Cos_{\mathbf{M}}(\mathbf{x}, \mathbf{y}) = \dfrac{< \mathbf{x}, \mathbf{y} >_{\mathbf{M}}}{\|\mathbf{x}\|_{\mathbf{M}}\|\mathbf{y}\|_{\mathbf{M}}}$
- $\mathbf{M}$-ortogonalidad: $< \mathbf{x}, \mathbf{y} >_{\mathbf{M}} = \mathbf{x}'\mathbf{M}\mathbf{y} = 0$

En los métodos en ejes principales de este texto la métrica se generaliza a matrices diagonales. Se nota $\mathbf{M}$ a la matriz diagonal de métrica en el espacio de las filas y de pesos en el de las columnas, y $\mathbf{N}$ a la matriz diagonal de pesos de las filas y de métrica en el espacio de las columnas. La matriz $\mathbf{X}$ está centrada con los pesos dados en $\mathbf{N}$, es decir, $\mathbf{g} = \mathbf{X}'\mathbf{N}\mathbf{1}_n = \mathbf{0}$, siendo $\mathbf{1}_n$ un vector columna de n unos. Con la definición de la tripleta $(\mathbf{X}, \mathbf{M}, \mathbf{N})$ el análisis en ejes principales queda completamente determinado y se nota ACP$(\mathbf{X}, \mathbf{M}, \mathbf{N})$.

El ACP normado del capítulo anterior se puede expresar como los ACP generalizados siguientes:

- ACP$(\mathbf{Y}_c, diag(\frac{1}{\sigma_j^2}), \frac{1}{n}\mathbf{I}_n)$. La matriz a analizar es la de datos centrados, la métrica $diag(\frac{1}{\sigma_j^2})$ es la matriz diagonal de las inversas de las varianzas y $\frac{1}{n}\mathbf{I}_n$ es la matriz de pesos, donde $\mathbf{I_n}$ es la matriz identidad de dimensión n.

- ACP$\left(\mathbf{X}, \mathbf{I}_p, \frac{1}{n}\mathbf{I}_n\right)$ si $\mathbf{X}$ es la matriz de datos estandarizados.
- ACP$(\mathbf{Z}, \mathbf{I}_p, \mathbf{I}_n)$, donde $\mathbf{Z}= \frac{1}{\sqrt{n}}\mathbf{X}$.

4.1. Análisis en $\mathbb{R}^p$: espacio de las filas

En el espacio de las filas los p ejes están asociados a las columnas, los pesos están definidos en la matriz $\mathbf{N}$ y la matriz de métrica es $\mathbf{M}$. Los pesos intervienen en los cálculos del centro de gravedad y de la inercia, y $\mathbf{M}$ en los cálculos de distancias, ángulos y proyecciones.

4.1.1. Coordenadas y pesos de filas

En $\mathbb{R}^p$ las filas de $\mathbf{X}$ se representan como puntos $\mathbf{x}_i= [x_{i1}\ x_{i2}\ \cdots\ x_{ip}]'$, cuyo conjunto se denomina *nube de puntos fila* y se denota N_n. A cada punto fila se le asocia el peso p_i, que es el término $p_i = n_{ii}$ de la matriz $\mathbf{N}$. La suma de los pesos es 1: $\sum_{i=1}^{n} p_i = 1$.

4.1.2. Distancias entre filas

Con una métrica diagonal $\mathbf{M}$, la distancia entre dos filas i y l es:

$$d^2(i,l) = \sum_{j=1}^{p} m_j (x_{ij} - x_{lj})^2 \tag{4.1}$$

donde $m_j = m_{jj}$.

4.1.3. Inercia de la nube N_n

La inercia total es la suma ponderada de las distancias al cuadrado de los puntos-fila al centro de gravedad de la nube:

$$Inercia = \sum_{i=1}^{n} p_i d^2(\mathbf{i}, \mathbf{0}) = \sum_{i=1}^{n} p_i \sum_{j=1}^{p} m_j x_{ij}^2 = \sum_{i,j} p_i m_j x_{ij}^2 \tag{4.2}$$

Cada punto-fila contribuye a la inercia con el producto de su peso por el cuadrado de la distancia al centro de gravedad, el cual coincide con el origen de la representación.

4.1.4. Descomposición de la inercia en ejes principales

Lo que busca el ACP generalizado es encontrar un sistema de ejes $\mathbf{u}$, $\mathbf{M}$–ortonormales ($\mathbf{u'}_s\mathbf{M}\mathbf{u}_s = 1$ y $\mathbf{u'}_s\mathbf{M}\mathbf{u}_t = 0, s \neq t$) de $\mathbf{M}$-inercia máxima.

Sea $F = \mathbf{XMu}$ el vector de las coordenadas sobre el eje definido por $\mathbf{u}$. Entonces la inercia proyectada sobre el eje es $\sum_{i=1}^{n} p_i F^2 = \mathbf{u'MX'NXMu}$. Maximizar esta cantidad es encontrar la dirección de mayor inercia proyectada.

La solución es un vector propio $\mathbf{u_1}$, $\mathbf{M}$-normado correspondiente al mayor valor propio λ_1 de la matriz $\mathbf{X'NXM}$. Nótese que la inercia (4.2) es también igual a la traza de la matriz de inercia $\mathbf{X'NXM}$.

Se procede luego a encontrar la segunda dirección $\mathbf{M}$-ortogonal a la primera, que maximiza la inercia proyectada sobre ese eje; luego una tercera, $\mathbf{M}$-ortogonal a las dos primeras, y así sucesivamente. Se encuentra entonces un nuevo sistema de ejes $\mathbf{u_s}$, que son vectores propios $\mathbf{M}$-unitarios, asociados a los valores propios de la matriz $\mathbf{X'NXM}$ y ordenados de mayor a menor.

Proyectar sobre el subespacio de dimensión S ($S < p$) con la inercia máxima proyectada sobre él, es seleccionar como ejes los generados por los S primeros vectores propios. La inercia proyectada sobre el subespacio de dimensión S es la suma de las inercias proyectadas sobre los ejes ortogonales que lo conforman.

4.1.5. Coordenadas sobre un eje factorial s

Un eje factorial es la recta generada por $\mathbf{u}_s$, uno de los dos vectores propios $\mathbf{M}$ unitarios asociados al valor propio λ_s. Las coordenadas del vector de proyecciones de todas las filas sobre el eje s son $F_s = \mathbf{XMu_s}$.

Para un individuo i la coordenada de la proyección es:

$$F_s(i) = \sum_{j=1}^{p} m_j x_{ij} u_s(j)$$

La inercia proyectada sobre el eje s es: $\sum_{i=1}^{n} p_i F_s(i)^2 = \mathbf{u'_s MX'NXMu_s} = \lambda_s$

4.2. Análisis en $\mathbb{R}^n$: espacio de las columnas

En la nube de los p puntos columna en $\mathbb{R}^n$, los ejes están asociados a las filas, las distancias están definidas por la matriz de métrica $\mathbf{N}$ y los pesos están en la matriz $\mathbf{M}$.

4.2.1. Coordenadas y pesos

Las coordenadas de un punto j son los n valores de la columna

$$\mathbf{X}_j = [X_{1j}\ X_{2j}\ \cdots\ X_{nj}]'$$

Su peso es m_j y todos los puntos conforman la nube de p columnas, N_p.

4.2.2. Distancias entre columnas

Con una métrica diagonal $\mathbf{N}$, la distancia entre dos columnas j y k es:

$$d^2(j,k) = \sum_{i=1}^{n} p_i (x_{ij} - x_{ik})^2 \tag{4.3}$$

4.2.3. Inercia de la nube N_p

La inercia total es la suma ponderada de las distancias al cuadrado de los puntos-columna al origen:

$$Inercia(N_p) = \sum_{j=1}^{p} m_j d^2(\mathbf{j}, \mathbf{0}) = \sum_{j=1}^{p} m_j \sum_{i=1}^{n} p_i x_{ij}^2 = \sum_{i,j} m_j p_i x_{ij}^2 \tag{4.4}$$

Nótese que las inercias de las nubes de los espacios de filas y de columnas son iguales.

4.2.4. Descomposición de la inercia en ejes principales

Se busca la dirección $\mathbf{v}$ sobre la cual la $\sum_{k=1}^{p} m_k G_j^2$ sea máxima, donde $\mathbf{G_j}$ es la proyección de la columna j sobre la dirección $\mathbf{v}$. El vector de todas las proyecciones sobre $\mathbf{v}$ es $\mathbf{X'Nv}$ y la cantidad a maximizar es $\mathbf{v'NXMX'Nv}$ sujeta

a la restricción $\mathbf{v}'\mathbf{N}\mathbf{v} = 1$. Sin embargo, no es necesario realizar la diagonalización de la matriz $\mathbf{XMX}'\mathbf{N}$ ya que los valores propios de los dos espacios son iguales y sus vectores propios y los ejes factoriales están relacionados.

4.3. Dualidad entre los espacios de filas y columnas

Los ejes y planos factoriales de los espacios de filas y columnas provienen de espacios vectoriales diferentes pero relacionados. Suponiendo que el número de filas n es superior al número de columnas p, el rango máximo de las matrices de inercia de los dos espacios es p. En el espacio de las filas la matriz de inercia es de orden p y en el espacio de las columnas es de orden n. En los cálculos se buscan los valores y vectores propios de la matriz de inercia $\mathbf{X}'\mathbf{NXM}$. Para un eje s las proyecciones se encuentran mediante $\mathbf{XMu_s}$. No se calculan los valores y vectores propios de la matriz $\mathbf{XMX}'\mathbf{N}$, sino que se utilizan algunas de las relaciones entre dos espacios para obtenerlos. Las demostraciones son sencillas y se pueden ver, por ejemplo, en Lebart *et al.* (1995).

Notando un valor propio de la matriz $\mathbf{XMX}'\mathbf{N}$ como μ_s, las relaciones entre los dos espacios son:

- Los valores propios diferentes de cero son iguales en los dos espacios: $\mu_s = \lambda_s$.
- El vector de coordenadas F_s sobre $\mathbf{u_s}$ es un vector propio asociado a μ_s.
- La $\mathbf{N}$-norma al cuadrado del vector F_s es λ_s
- El vector $\mathbf{v_s}$ es igual a $\frac{1}{\sqrt{\lambda_s}}F_s$.
- El vector de coordenadas G_s es igual a $\sqrt{\lambda_s}\mathbf{u_s}$.

4.3.1. Fórmula de reconstitución de los datos

La representación de las nubes de puntos sobre todos los ejes factoriales es un cambio de base. Entonces es posible expresar la matriz $\mathbf{X}$ en función de las coordenadas sobre los ejes factoriales. Escofier & Pagès (1992) obtienen el término general de $\mathbf{X}$ de la siguiente manera:

- Un vector fila $\mathbf{x}'_i$ de $\mathbf{X}$ en función de la base generada por los vectores propios $\mathbf{u}_s$ es: $\mathbf{x}'_i = \sum_s F_s(i)\mathbf{u}_s$.
- Una componente x_{ij} sobre la base canónica es: $x_{ij} = \sum_s F_s(i)\mathbf{u}_s(j)$.
- Como $\mathbf{u}_s = \frac{1}{\sqrt{\lambda_s}} G_s$ entonces:

$$x_{ij} = \sum_s \frac{F_s(i)G_s(j)}{\sqrt{\lambda_s}} \tag{4.5}$$

- En forma matricial la fórmula de reconstitución es una suma de matrices de rango 1:

$$\mathbf{X} = \sum_s \frac{1}{\sqrt{\lambda_s}} F_s G'_s = \sum_s \sqrt{\lambda_s}\mathbf{v}_s\mathbf{u}'_s \tag{4.6}$$

Retener los primeros S ejes equivale a tener una aproximación de la matriz $\mathbf{X}$, denotada por Lebart *et al.* (2006) $\tilde{\mathbf{X}}$:

$$\tilde{\mathbf{X}} = \sum_{s=1}^{S} \frac{1}{\sqrt{\lambda_s}} F_s G'_s = \sum_{s=1}^{S} \sqrt{\lambda_s}\mathbf{v}_s\mathbf{u}'_s \tag{4.7}$$

Y la calidad de la aproximación, denotada τ, es:

$$\tau = \frac{\sum_{s=1}^{S} \lambda_s}{\sum_{s=1}^{p} \lambda_s} \tag{4.8}$$

Un diagrama de barras (*barplot*) de los valores propios es una guía para seleccionar el número de ejes a retener, S, complementado con los valores sucesivos de τ, que informan de la calidad de la representación a medida que se incrementa S.

4.3.2. Fórmulas del ACP generalizado

Un método específico, en ejes principales, queda completamente determinado definiendo las matrices: $\mathbf{X}$, que se obtiene de los datos mediante la transformación adecuada; $\mathbf{M}$, métrica en el espacio de las filas y pesos en el espacio de las columnas, y $\mathbf{N}$, pesos en el espacio de las filas y métrica en

el espacio de las columnas. Entonces las fórmulas del método específico se obtienen reemplazando en las fórmulas ACP($\mathbf{X}$, $\mathbf{M}$, $\mathbf{N}$) (tabla 4.1).

4.3.3. Diagrama de dualidad

Los espacios vectoriales de filas ($E = \mathbb{R}^p$) y columnas ($F = \mathbb{R}^n$) están conectados mediante transformaciones lineales definidas por las matrices $\mathbf{X}$, $\mathbf{M}$ y $\mathbf{N}$, composiciones de estas y las inversas $\mathbf{M}^{-1}$ y $\mathbf{N}^{-1}$.

Se denomina *diagrama de dualidad* al esquema que muestra los espacios vectoriales y las transformaciones lineales que permiten pasar de un espacio a otro. La figura 4.1 muestra el diagrama de dualidad asociado al ACP($\mathbf{X}$, $\mathbf{M}$, $\mathbf{N}$), que tiene además utilidad pnemotécnica.

Algunos ejemplos de transformaciones lineales son:

- Para encontrar el vector de las proyecciones de todas las filas sobre un eje $\mathbf{u}_s$: el espacio de origen es $E = \mathbb{R}^p$ y el de llegada es $F = \mathbb{R}^n$, pasando por E^* la transformación es: $F_s = \mathbf{XMu}_s$.
- La matriz $\mathbf{V}$, transformación lineal de E^* a E, es equivalente a la transformación $\mathbf{X'NX}$ pasando por los espacios F y F^*.
- La matriz de inercia en E se obtiene con la transformación compuesta de las cuatro transformaciones lineales que permiten dar la vuelta al diagrama $\mathbf{X'NXM}$.

En el artículo de Tenenhaus & Young (1985, p. 105) se puede estudiar el diagrama de dualidad con más detalle, y en Holmes (2008) se encuentra un buen resumen.

4.4. Ayudas para la interpretación de las gráficas

Los elementos que aparecen en los planos factoriales tienen errores de proyección y es importante tener indicadores que eviten lecturas erróneas. La inercia de la nube de puntos que se está analizando se descompone en ejes factoriales. Entonces es útil disponer de indicadores que ayuden a detectar los elementos que más contribuyen, a la inercia de espacios y subespacios, en particular a las varianzas de los ejes. La inercia sobre un eje, un plano o sub-espacios de mayor dimensión se puede descomponer en la contribución de elementos o en subconjuntos de una partición de sus elementos.

Tabla 4.1. Fórmulas del ACP(**X**, **M**, **N**)

Espacio	$\mathbb{R}^p$	$\mathbb{R}^n$
Nube	N_n	N_p
Coordenadas	Filas de **X**: $\mathbf{x}'_i$	Columnas de **X**: $\mathbf{X}_j$
Pesos	Diagonal de **N**: p_i	Diagonal de **M**: m_j
Métrica	**M**	**N**
Distancias al cuadrado	$d^2(i,l) = \sum_{j=1}^{p} m_j(x_{ij} - x_{lj})^2$	$d^2(j,k) = \sum_{i=1}^{n} p_i(x_{ij} - x_{ik})^2$
Inercia	$Traza(\mathbf{X'NXM})$	$Traza(\mathbf{XMX'N})$
Valor propio	λ_s	λ_s
Vector propio	$\mathbf{u_s}$	$\mathbf{v_s}$
Fórmula valor-vector propio	$\mathbf{X'NXMu_s} = \lambda_s\mathbf{u_s}$	$\mathbf{XMX'Nv_s} = \lambda_s\mathbf{v_s}$
Coordenadas factoriales	$F_s(i) = \mathbf{x'_i Mu_s}$ $G_s = \mathbf{X'Nv_s} = \sqrt{\lambda_s}\mathbf{u_s}$	$G_s(j) = \mathbf{X'_j Nv_s}$ $F_s = \mathbf{XMu_s} = \sqrt{\lambda_s}\mathbf{v_s}$
Fórmulas de transición	$G_s = \frac{1}{\sqrt{\lambda_s}}\mathbf{X'N}F_s$ $F_s(i) = \frac{1}{\sqrt{\lambda_s}}\sum_{j=1}^{p} x_{ij}m_jG_s(j)$	$F_s = \frac{1}{\sqrt{\lambda_s}}\mathbf{XM}G_s$ $G_s(j) = \frac{1}{\sqrt{\lambda_s}}\sum_{i=1}^{n} x_{ij}p_iF_s(i)$
Fórmula de reconstitución	$x_{ij} = \sum_s \frac{F_s(i)G_s(j)}{\sqrt{\lambda_s}} = \sum_s \sqrt{\lambda_s}\mathbf{u}_s(i)\mathbf{v}_s(j); \quad \mathbf{X} = \sum_s \sqrt{\lambda_s}\mathbf{v}_s\mathbf{u}'_s$	

Fuente: Escofier & Pagès (1992, Cap. 4)

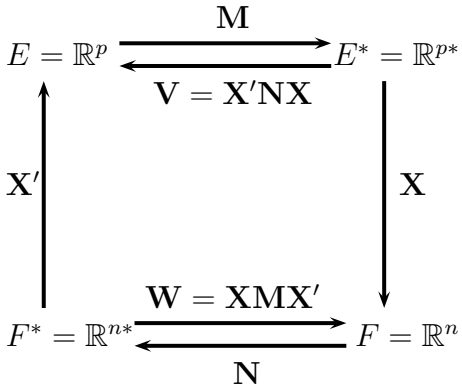


Figura 4.1. Diagrama de dualidad del $ACP(\mathbf{X}, \mathbf{M}, \mathbf{N})$. E es el espacio de las filas y F el de las columnas

Los indicadores para ayudar a la interpretación se pueden definir mediante cocientes de inercias (Escofier & Pagès, 1992).

Para simplificar las fórmulas, en esta sección, no se incluyen las matrices de las métricas involucradas en las distancias e inercias, **M** y **N** para los espacios de las filas y las columnas, respectivamente.

La inercia de la nube de filas N_n en el espacio completo es:

$$I(N_n) = \sum_{i=1}^{n} p_i d^2(i, \mathbf{g}) = \sum_{i=1}^{n} p_i ||\mathbf{x}_i||^2 = \sum_{s} \lambda_s \tag{4.9}$$

La contribución de una fila i a la inercia total de una nube de puntos es $p_i||\mathbf{x}_i||^2$. Por lo tanto, la proporción de inercia con la que contribuye una fila i a la inercia de la nube de puntos es:

$$\frac{p_i||\mathbf{x}_i||^2}{I(N_n)} = \frac{p_i||\mathbf{x}_i||^2}{\sum\limits_{s} \lambda_s} \tag{4.10}$$

La inercia de la nube de puntos, proyectada sobre un eje s es:

$$I_s(N_n) = \sum_{i=1}^{n} p_i F_s^2(i) = \lambda_s \tag{4.11}$$

Y un sumando es la contribución de una fila i a la inercia proyectada sobre un eje s: $p_i F_s^2(i)$.

4.4.1. Calidad de la representación, coseno cuadrado o contribución relativa

Para una fila i, la *contribución relativa* sobre un eje s es la relación entre las contribuciones de i a la inercia: proyectada sobre el eje y en el espacio completo:

$$Cr_s(i) = \frac{p_i F_s^2(i)}{p_i ||\mathbf{x}_i||^2} = \frac{F_s^2(i)}{||\mathbf{x}_i||^2} = Cos_s^2(i) \tag{4.12}$$

Este indicador, como coseno cuadrado, tiene una interpretación geométrica: es el cuadrado del cociente entre la longitud del vector proyectado sobre el eje s y su longitud en el espacio completo. Un valor de 1 indica que el punto está sobre el eje s y un valor de 0, que es perpendicular. El analista de datos debe abstenerse de interpretar los puntos que tienen calidad de representación cercana a cero. Por la ortogonalidad de los ejes factoriales, el coseno cuadrado sobre un plano es la suma de los cosenos cuadrados de los ejes factoriales que lo generan. Esto se generaliza a un subespacio de dimensión S. Para las columnas, la calidad de la representación se define de la misma forma.

4.4.2. Contribución absoluta

La contribución absoluta de una fila i a la inercia de un eje s, expresada en proporción, es:

$$Ca_s(i) = \frac{p_i F_s^2(i)}{\lambda_s} \tag{4.13}$$

Estas contribuciones ayudan a interpretar el eje, lo que se traduce muchas veces en darle un significado dentro del contexto del análisis. También permite detectar uno o unos pocos elementos responsables de casi toda la inercia del eje, es decir, puntos influyentes.

Para las columnas la contribución absoluta se define de la misma forma.

4.4.3. Calidad de la representación sobre un subespacio

La inercia proyectada sobre un eje s es igual al valor propio λ_s y la de un subespacio S es la suma de los valores propios que lo generan.

Las calidades de la representación de la nube de puntos en un subespacio se definen como el cociente de inercia de la nube de puntos: en el subespacio

y en el espacio completo.

$$\frac{\sum_{s=1}^{S} \lambda_s}{\sum_{s} \lambda_s} \tag{4.14}$$

Entonces las calidades de representación de las nubes proyectadas son:

- Sobre un eje s, $\frac{\lambda_s}{\sum_s \lambda_s}$.
- Sobre un plano conformado por los ejes s y t, $\frac{\lambda_s + \lambda_t}{\sum_s \lambda_s}$.
- En particular, sobre el primer plano factorial: $\frac{\lambda_1 + \lambda_2}{\sum_s \lambda_s}$.

La presentación ordenada de las calidades de representación sobre: primer eje, primer plano, subespacio 3D (ejes 1, 2 y 3), y así sucesivamente hasta el espacio completo, es una de las ayudas para decidir el número de ejes a retener en el análisis.

4.5. Elementos suplementarios o ilustrativos

Sobre los subespacios factoriales se pueden proyectar elementos que no participaron en el análisis, ya sean filas o columnas. También es posible proyectar filas *artificiales* –por ejemplo, las filas promedio de grupos construidos a partir de variables cualitativas–. Las fórmulas de proyección son las mismas de los elementos activos. Es válido calcular la calidad de la representación de los elementos suplementarios. La contribución a la formación de los ejes es obviamente nula.

4.6. Imagen euclidiana de matrices de varianzas-covarianzas y correlaciones

Es posible que se desee obtener una imagen geométrica de matrices de varianzas-covarianzas o de correlaciones, cuando no se dispone de los datos originales con los cuales se calcularon. Este problema se puede denominar ACP *a partir de las matrices de covarianzas o de correlaciones*.

En tales casos no se dispone de los datos de los "individuos" y el diagrama de dualidad no se puede completar. Solo se tiene la parte del diagrama

$$E = \mathbb{R}^p \underset{\mathbf{V}}{\overset{\mathbf{M}}{\rightleftarrows}} E^* = \mathbb{R}^{p*}$$

Figura 4.2. Diagrama cuando solo se conoce la matriz de varianzas o de correlaciones. Parte superior del diagrama de dualidad de la figura 4.1

que se muestra en la figura 4.2. Para una matriz de covarianzas o de correlaciones de orden p la métrica $\mathbf{M}$ es $\mathbf{I}_p$.

En el espacio E se encuentran los vectores propios unitarios de $\mathbf{V}$, $\mathbf{u}_s$, asociados a los valores propios λ_s:

- Valores propios: $\lambda_1 \geq \cdots \geq \lambda_s \geq \cdots \geq \lambda_p$.
- Vectores propios: $\mathbf{u}_1, \cdots \mathbf{u}_s, \cdots, \mathbf{u}_p$.
- Coordenadas de las variables: $G_1, \cdots G_s, \cdots G_p$. De las fórmulas de la tabla 4.1:

$$\mathbf{G}_s = \sqrt{\lambda_s}\mathbf{u}_s \tag{4.15}$$

Si la matriz $\mathbf{V}$ es la de correlaciones, los planos factoriales que se obtienen se denominan *círculos de correlaciones*. Se pueden obtener y dibujar con los siguientes comandos de R. Un ejemplo se muestra en el taller 4.9.

Código para obtener las coordenadas y el círculo de correlaciones

```
V # matriz de correlaciones
eigV <- eigen(V)
Lambda<-diag(eigV$values)
U<-eigV$vectors
G<-U%*%sqrt(Lambda)
library(ade4)
s.corcircle(G)
```

4.7. Análisis en coordenadas principales

Se *denomina análisis en coordenadas principales* a la obtención de imágenes euclidianas de matrices de distancias entre individuos. Primero se obtiene la matriz de productos internos $\mathbf{W}$, a partir de la matriz de distancias $\mathbf{D}$. La

matriz $\mathbf{W}$ aparece en el diagrama de la figura 4.3, donde $\mathbf{N}$ es la matriz de pesos de los individuos, $\mathbf{N} = diag(p_i)$. Las distancia entre dos individuos i y l se nota d_{il}. Una celda i, l de $\mathbf{W}$ se obtiene mediante (Escofier & Pagès, 1992, p. 84):

$$w_{il} = \frac{1}{2}(d_{i\cdot}^2 + d_{l\cdot}^2 - d_{il}^2 - d_{\cdot\cdot}^2) \tag{4.16}$$

donde:

$$d_{i\cdot}^2 = \sum_{l=1}^{n} p_l d_{il}^2 \quad \text{y} \quad d_{\cdot\cdot}^2 = \sum_{i,l=1}^{n} p_i p_l d_{il}^2$$

$$F^* = \mathbb{R}^{n*} \underset{\mathbf{N}}{\overset{\mathbf{W}}{\rightleftarrows}} F = \mathbb{R}^n$$

Figura 4.3. Diagrama cuando solo se conoce la matriz de productos internos $\mathbf{W}$. Parte inferior del diagrama de dualidad de la figura 4.1

En el espacio F se encuentran los vectores propios $\mathbf{N}$ unitarios, $\mathbf{v}_s$ de la matriz $\mathbf{WN}$, asociados a sus valores propios λ_s. Sea r el rango de $\mathbf{WN}$:

- Valores propios: $\lambda_1 \geq \cdots \geq \lambda_s \geq \cdots \geq \lambda_r$.
- Vectores propios: $\mathbf{v}_1, \cdots, \mathbf{v}_s, \cdots, \mathbf{v}_r$.
- Coordenadas de los individuos: $F_1, \cdots F_s, \cdots F_r$. De las fórmulas de la tabla 4.1:

$$\mathbf{F}_s = \sqrt{\lambda_s}\mathbf{v}_s \tag{4.17}$$

El análisis en coordenadas principales se encuentra implementado, entre otros, en el `ade4` en la función `dudi.pco`. En el taller 4.9.2 se puede ver un ejemplo de aplicación.

Para disimilitudes no euclidianas también se puede obtener este tipo de gráficas, metodología que se conoce con el nombre de *escalamiento multidimensional.*

4.8. Ejercicios

1. Escriba el diagrama de dualidad y las principales fórmulas del *ACP* $(\mathbf{Y}_c, \mathbf{I}_2, \mathbf{I}_{10})$ del taller ACP geométrico sección 3.8.1 (página 81). Encuentre analíticamente los valores y vectores de la matriz de inercia del taller.

2. Muestre que las distancias entre individuos en el $ACP(\mathbf{Y}_c, diag(\frac{1}{\sigma_j^2}), \frac{1}{n}\mathbf{I}_n)$ son iguales a las del ACP($\mathbf{X}$, $\mathbf{I}_p$, $\frac{1}{n}\mathbf{I}_n$).

3. Demuestre que el primer eje factorial del ACP generalizado es el generado por uno de los dos vectores propios **M**-unitarios asociados al mayor valor propio de la matriz **X′NXM**.

4. Muestre que las coordenadas factoriales de las filas sobre un eje son **N**-centradas.

5. Para el ACP generalizado demuestre que:

   a) Los valores propios diferentes de cero son iguales en los dos espacios: $\mu_s = \lambda_s$.

   b) El vector de coordenadas F_s sobre $\mathbf{u_s}$ es un vector propio asociado a μ_s.

   c) La **N**-norma al cuadrado del vector F_s es λ_s.

   d) El vector $\mathbf{v_s}$ es igual a $\frac{1}{\sqrt{\lambda_s}}F_s$.

   e) El vector de coordenadas G_s es igual a $\sqrt{\lambda_s}\mathbf{u_s}$.

6. Dibuje el diagrama de dualidad para el ACP canónico: ACP($\mathbf{X}$, $\mathbf{I}_p$, $\frac{1}{n}\mathbf{I}_n$).

7. Utilizando el diagrama de dualidad muestre que la matriz a diagonalizar en $\mathbb{R}^p$ se encuentra con la composición de las aplicaciones lineales sobre un vector de E, de modo que le da la vuelta al diagrama hasta llegar a E de nuevo.

8. Utilice el diagrama de dualidad para mostrar las fórmulas que permiten obtener las coordenadas de las variables a partir de una matriz de covarianzas o correlaciones.

9. Demuestre que, en el análisis en coordenadas principales,

$$w_{il} = \frac{1}{2}(d_{i\cdot}^2 + d_{l\cdot}^2 - d_{il}^2 - d_{\cdot\cdot}^2) \text{ donde: } d_{i\cdot}^2 = \sum_{l=1}^{n} p_l d_{il}^2 \text{ y } d_{\cdot\cdot}^2 = \sum_{i,l=1}^{n} p_i p_l d_{il}^2$$

4.9. Talleres

En esta sección se realizan dos talleres: en el primero se busca obtener el círculo de correlaciones cuando no se conocen los datos que dieron origen

a una matriz de correlaciones; en el segundo, obtener la imagen geométrica de una matriz de distancias euclidianas, cuando no se tienen las variables de donde se calcularon.

4.9.1. Imagen euclidiana de matrices de varianzas-covarianzas y de correlaciones

En el artículo de Correa, De Rosa & Lesino (2006) se realiza un análisis del clima en la ciudad de Mendoza (Argentina). El artículo no presenta los datos originales pero tiene dos matrices de correlaciones. El ejercicio de este taller consiste en construir los círculos de correlaciones a partir de las matrices y comparar con los resultados del artículo. Para el caso de la matriz de correlaciones que corresponde a un día de primavera por la mañana (tabla 4.2), responda las siguientes preguntas:

1. Objetivo de análisis.
2. Descripción de las variables.
3. ¿Cuáles son las unidades estadísticas ("individuos") en el análisis?
4. ¿Se pueden obtener coordenadas y ayudas para la interpretación de los "individuos" cuando solo se tienen las matrices de correlaciones? ¿Por qué?
5. ¿Cuántos ejes selecciona para analizar? ¿Por qué?
6. ¿Qué significado le puede dar cada uno de los ejes que va a analizar?
7. Grafique y analice los planos factoriales que estime conveniente.
8. Resuma el análisis respondiendo a los objetivos.

Tabla 4.2. Matriz de correlaciones entre variables de clima en la ciudad de Mendoza

	temp	vevi	alti	fvc	emis	anca	orie	nubo
vevi	0.113							
alti	0.411	0.040						
fvc	0.174	0.291	0.102					
emis	0.067	0.009	0.406	-0.118				
anca	0.134	0.009	-0.293	0.165	-0.361			
orie	0.101	0.018	-0.018	-0.200	-0.104	0.029		
nubo	0.836	0.123	0.157	0.176	-0.007	0.176	0.103	
iner	-0.164	-0.050	-0.443	0.119	-0.966	0.290	0.035	-0.053

Fuente: Correa *et al.* (2006).

4.9.2. Análisis en coordenadas principales

En Hidalgo *et al.* (2007) se construye una distancia cultural entre algunos países latinoamericanos. En la tabla 4.3 se presentan las distancias, que corresponden a la raíz cuadrada de las presentadas en el artículo.

Tabla 4.3. Distancias culturales entre países de Latinoamérica

	ARG	BOL	BRA	COL	CRI	ECU	SLV	GTM	MEX
BOL	1.534								
BRA	1.314	1.351							
COL	0.903	1.559	1.325						
CRI	1.527	1.108	1.205	1.665					
ECU	1.554	1.694	1.554	1.524	1.426				
SLV	1.221	1.748	0.928	1.202	1.833	1.586			
GTM	0.909	1.586	1.268	1.149	1.693	1.345	1.199		
MEX	1.692	1.549	1.412	1.357	1.249	1.358	1.516	1.900	
VEN	1.103	1.204	1.477	1.117	1.798	1.615	1.154	1.318	1.597

Nota: el encabezado de las columnas corresponde al código de tres letras de la norma ISO 3166-1: CRI Costa Rica, SLV El Salvador, GTM Gatemala.

Realice el análisis en coordenadas principales (ACO) sobre la matriz de distancias, utilizando las funciones `dudi.pco` e `inertia.dudi` de `ade4` y responda a las siguientes preguntas:

1. ¿Cuál es la dimensión del espacio de representación?
2. ¿Cuántos ejes selecciona para el análisis? ¿Por qué?
3. ¿Se pueden obtener ayudas para la interpretación de las variables?

4. ¿Algunos países tienen una calidad de representación en el primer plano factorial inferior al 10 %? ¿Cuáles?

5. ¿Qué países tienen una contribución al primer eje por encima del promedio?

6. Analice los planos factoriales. Puede utilizar `plot.dudi{FactoClass}` para graficar los planos factoriales que requiera.

7. A partir de los planos factoriales establezca una partición de los países. Describa comparativamente los grupos de países formados.

8. Compare los resultados con los del artículo.

9. Haga un resumen práctico del análisis.

Capítulo

cinco

Análisis de correspondencias simples

El análisis de correspondencias simples (ACS) se utiliza para describir tablas de contingencia (TC) mediante la representación geométrica de las tablas de condicionales fila y columna (perfiles) derivadas de aquellas. El objetivo del ACS es describir las asociaciones entre las variables fila y columna, a través de sus perfiles:

1. Comparar los perfiles fila.
2. Comparar los perfiles columna.
3. Estudiar las correspondencias entre perfiles fila y columna.

El ACS se puede ver técnicamente como dos ACP o como un ACP. La primera visión conviene para la interpretación de los resultados y la segunda, para los cálculos. En este capítulo se muestran las dos visiones, pero conviene complementar con la lectura del capítulo correspondiente en Lebart, Morineau & Piron (1995) o Lebart, Piron & Morineau (2006).

5.1. Pequeño ejemplo y notación

Se utiliza como ejemplo, la tabla de contingencia (TC) que clasifica a los 445 estudiantes admitidos a las carreras de la Facultad de Ciencias 2013-I, según la carrera y el estrato socioeconómico (tabla 5.1). La tabla se obtiene a partir de los datos `admi{FactoClass}`.

5.1.1. Tabla de contingencia

Siguiendo la misma notación de Lebart *et al.* (1995), $\mathbf{K}$ es la tabla de contingencia; k_{ij}, su término general; $k_{i\cdot}$, la suma de su fila i; $k_{\cdot j}$, la suma de su columna j, y $k = k_{\cdot\cdot}$, su total. Por ejemplo: $k_{11} = 23$ admitidos a Biología que son de estrato bajo; $k_{1\cdot} = 63$ admitidos a Biología; $k_{\cdot 1} = 179$ de los admitidos son de estrato bajo; el total de la tabla es $k = 445$.

5.1.2. Tabla de frecuencias relativas

La tabla de frecuencias relativas se nota $\mathbf{F}$ de término general $f_{ij} = \dfrac{k_{ij}}{k}$.

El término general de su marginal fila se nota $f_{i\cdot}$ y el de su marginal columna $f_{\cdot j}$.

En la tabla 5.1 está la tabla **F** y sus sumas, que representan la distribución de probabilidad conjunta y las distribuciones marginales, respectivamente, expresadas en porcentaje.

Los 23 admitidos a Biología que son de estrato bajo representan el $f_{11} = 5.2\,\%$ de los admitidos; el $f_{1\cdot} = 14.2\,\%$ entran a Biología y el $f_{\cdot 1} = 40.24\,\%$ de los admitidos son de estrato bajo.

Código para obtener **F**, $\mathbf{D}_n$, $\mathbf{D}_p$ y las tabla 5.1

```
library(FactoClass);
data(admi);
K<-unclass(table(admi$carr,admi$estr));
F<-K/sum(K)*100; # o  F<-prop.table(K)*100, en porcentaje
Dn<-diag(rowSums(F));
Dp<-diag(colSums(F));
tabs<-plotct(K,tables = TRUE); #tabla con plotct{FactoClass}
xtable(tabs$ctm,digits = rep(0,5));
xtable(tabs$ctm*100/sum(K),digits = rep(1,5));
```

Tabla 5.1. Tablas de contingencia y de frecuencias relativas de los admitidos a Ciencias, según carreras y estratos

Tabla de contingencia **K**

	Ebajo	Emedio	Ealto	Suma
Biología	23	26	14	63
Estadística	29	29	8	66
Farmacia	30	36	7	73
Física	27	36	19	82
Geología	18	9	18	45
Matemáticas	21	25	7	53
Química	31	24	8	63
Suma	179	185	81	445

Tabla de frecuencias relativas **F**

	Ebajo	Emedio	Ealto	Suma
Biología	5.2	5.8	3.1	14.2
Estadística	6.5	6.5	1.8	14.8
Farmacia	6.7	8.1	1.6	16.4
Física	6.1	8.1	4.3	18.4
Geología	4.0	2.0	4.0	10.1
Matemáticas	4.7	5.6	1.6	11.9
Química	7.0	5.4	1.8	14.2
Suma	40.2	41.6	18.2	100.0

La marginal fila representa la distribución de frecuencias relativas de los admitidos según carreras y la marginal columna es la distribución de los admitidos según estratos.

Con estas marginales se definen las matrices diagonales: $\mathbf{D}_n = diag(f_{i\cdot})$ y $\mathbf{D}_p = diag(f_{\cdot j})$.

5.1.3. Tabla de perfiles fila

Para cada carrera se tiene una distribución de frecuencias relativas entre los tres estratos, que se denomina *distribución condicional* o *perfil fila*. Se obtiene al dividir cada celda de la respectiva fila por la suma de la fila, en la TC o en la tabla de frecuencias relativas **F**. La marginal columna de la tabla **F** se constituye en la distribución promedio de los perfiles fila y es la distribución de todos los 445 admitidos en los tres estratos, sin importar la carrera. Un perfil fila i se nota:

$$\left\{\frac{f_{ij}}{f_{i\cdot}}; j = 1, \cdots, p\right\}$$

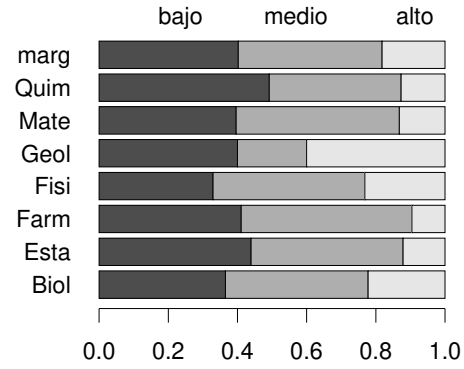
El conjunto de los perfiles fila se notan y calculan mediante $\mathbf{D}_n^{-1}\mathbf{F}$. En R: `solve(Dn)%*%F` (tabla 5.2). En tablas pequeñas, como la de este ejemplo, los perfiles se pueden mostrar gráficamente como barras que representan el 100 %, con colores o *achurados*, que muestran el porcentaje de cada categoría en el perfil (ver figura en la tabla 5.2).

Código para obtener tablas y gráficas de perfiles (tablas 5.2 y 5.3)

```
plotct(K,"row"); # plotct es una funcion de FactoClass
tabs<-plotct(K,"col",tables=TRUE);
# para exportar la grafica a xfig, para su edicion
#dev.print(device = xfig,file="perfilEstratos.fig")
# tablas de perfiles en formato tabular para LaTeX
xtable(cbind(tabs$perR,suma=rowSums(tabs$perR)),
   digits=rep(1,5));
xtable(rbind(tabs$perC,suma=colSums(tabs$perC)),
   digits=rep(1,5));
```

Tabla 5.2. Perfiles fila de la tabla carreras × estratos

	Ebajo	Emedio	Ealto	suma
Biología	36.5	41.3	22.2	100.0
Estadística	43.9	43.9	12.1	100.0
Farmacia	41.1	49.3	9.6	100.0
Física	32.9	43.9	23.2	100.0
Geología	40.0	20.0	40.0	100.0
Matemáticas	39.6	47.2	13.2	100.0
Química	49.2	38.1	12.7	100.0
Marginal C	40.2	41.6	18.2	100.0



Se puede observar que el perfil de Geología es el que más difiere de los demás porque tiene más porcentaje de estrato alto, en detrimento del porcentaje de estrato medio. Física y Biología también tienen más porcentaje de estrato alto, que el promedio. El perfil de Química es el que más porcentaje de estrato bajo tiene, seguido por Estadística.

5.1.4. Tabla de perfiles columna

Cada estrato tiene su distribución según las siete carreras (condicionales o perfiles columna). La distribución marginal fila es la distribución de todos los admitidos, en las siete carreras, sin importar el estrato.

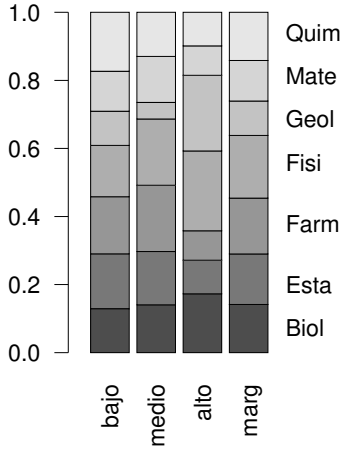
Un perfil columna j se nota:

$$\left\{\frac{f_{ij}}{f_{\cdot j}}; i = 1, \cdots, n\right\}$$

El conjunto de los perfiles columna se calcula mediante $\mathbf{FD}_p^{-1}$. En R: `F %* %solve(Dp)` (tabla 5.3).

Tabla 5.3. Perfiles columna de la tabla carreras × estratos

	Ebajo	Emedio	Ealto	Marginal F
Biología	12.8	14.1	17.3	14.2
Estadística	16.2	15.7	9.9	14.8
Farmacia	16.8	19.5	8.6	16.4
Física	15.1	19.5	23.5	18.4
Geología	10.1	4.9	22.2	10.1
Matemáticas	11.7	13.5	8.6	11.9
Química	17.3	13.0	9.9	14.2
suma	100.0	100.0	100.0	100.0



El perfil de estrato alto se diferencia más del promedio, tiene mayor porcentaje de estudiantes admitidos a Geología y Física y menos de Farmacia, Estadística y Matemáticas (tabla 5.3).

5.1.5. El modelo de independencia

Si se supone que no hay asociación –es decir, que hay independencia estadística entre las variables fila y columna– el modelo es $a_{ij} = f_{i\cdot} f_{\cdot j}$, término general de la tabla de independencia **A** (tabla 5.4).

Las distribuciones condicionales fila (columna) de la tabla **A** son todas iguales a la marginal de las columnas (filas) de la tabla **F**. Las desviaciones del modelo de independencia son **F-A** (ver tabla 5.4).

Tabla 5.4. Frecuencias relativas, independencia y diferencia

	F observada			**A** independencia		
	bajo	medio	alto	bajo	medio	alto
Biol	5.2	5.8	3.1	5.7	5.9	2.6
Esta	6.5	6.5	1.8	6.0	6.2	2.7
Farm	6.7	8.1	1.6	6.6	6.8	3.0
Fisi	6.1	8.1	4.3	7.4	7.7	3.4
Geol	4.0	2.0	4.0	4.1	4.2	1.8
Mate	4.7	5.6	1.6	4.8	5.0	2.2
Quim	7.0	5.4	1.8	5.7	5.9	2.6

	F – **A** diferencia		
	bajo	medio	alto
Biol	-0.5	-0.0	0.6
Esta	0.6	0.4	-0.9
Farm	0.1	1.3	-1.4
Fisi	-1.3	0.4	0.9
Geol	-0.0	-2.2	2.2
Mate	-0.1	0.7	-0.6
Quim	1.3	-0.5	-0.8

5.2. El ACS como dos ACP

En el ACS se describen simultáneamente los perfiles fila y columna. Para cada tabla de perfiles se realiza un ACP($\mathbf{X}$, $\mathbf{M}$, $\mathbf{N}$), pero los dos ACP están relacionados, lo que permite representaciones simultáneas de los planos factoriales.

5.2.1. ACP de los perfiles fila

La tabla que se analiza es la de perfiles fila, así que el histograma que representa a un perfil se ve como un punto en $\mathbb{R}^p$. La diferencia entre dos histogramas se traduce en una distancia entre los puntos que los representan.

Los pesos de los puntos fila son la distribución marginal y la suma de las filas de $\mathbf{F}$, que forman la matriz diagonal $\mathbf{D_n} = diag(f_{i\cdot})$.

Las distancias entre distribuciones condicionales se definen a partir del producto punto dado por la matriz $\mathbf{D_p^{-1}}$, donde $\mathbf{D}_p = diag(f_{\cdot j})$.

Las imágenes, para comparar a los perfiles fila, son los planos factoriales derivados del ACP($\mathbf{D_n^{-1}F}$, $\mathbf{D_p^{-1}}$, $\mathbf{D_n}$).

La matriz $\mathbf{D_n^{-1}F}$ no está centrada, pero el valor propio más grande de la matriz a diagonalizar ($\mathbf{F'D_n^{-1}FD_p^{-1}}$, tabla 4.1) es 1 y el vector propio asociado es el centro de gravedad de la nube. De modo que lo que se hace, en lugar de centrar, es eliminar este valor propio y su vector propio. Partir del segundo vector propio de esta matriz es equivalente a centrar y así se logra una simplificación de las formulas del ACS. A continuación se describen los elementos del ACS, con más detalle.

5.2.1.1. Coordenadas, pesos

Las coordenadas de los perfiles fila son: $\mathbf{x}_i; i = 1, 2, \cdots, n$. Con:

$$\mathbf{x}_i(j) = \frac{f_{ij}}{f_{i\cdot}};\ j = 1, 2, \cdots, p$$

En el ejemplo están en la tabla 5.2. Cuando $i = 3$ se tiene el perfil de Farmacia con coordenadas $\mathbf{x}_3 = [0.411, 0.493, 0.096]'$.

Este es el punto que en $\mathbb{R}^3$ representa la distribución de los admitidos a Farmacia según los 3 estratos.

Los pesos están en la diagonal de $\mathbf{D}_n$, que reune las marginales fila de $\mathbf{F}$ (tabla 5.1). Para Farmacia es: 0.164, así que el 16.4 % de los admitidos a Ciencias son de esta carrera.

Una categoría fila i define un grupo de individuos, que tiene un peso $f_{i\cdot}$, definido por la proporción con respecto al total de los individuos. Las coordenadas del perfil representan la proporción del grupo de individuos i que hay en cada categoría columna.

5.2.1.2. Centro de gravedad

El centro de gravedad se calcula con los pesos de los n perfiles:

$$\mathbf{g}_p = \sum_{i=1}^{n} f_{i\cdot}\mathbf{x}_i$$

La coordenada j, notada $\mathbf{g}_p(j)$, del centro de gravedad es:

$$\mathbf{g}_p(j) = \sum_{i=1}^{n} f_{i\cdot}\frac{f_{ij}}{f_{i\cdot}} = \sum_{i=1}^{n} f_{ij} = f_{\cdot j}$$

Es decir, el centro de gravedad es la marginal columna de la tabla **F**. En el ejemplo es $\mathbf{g}_p = [0.402, 0.416, 0.182]'$, que corresponde a la distribución de los 445 admitidos entre los 3 estratos y es el valor típico para comparar los perfiles de las 7 carreras.

Por ejemplo, en Farmacia hay un poco más de estratos bajo y medio y menos de alto, con respecto al promedio.

El centro de gravedad se sitúa en el origen de la representación. Para simplificar las fórmulas, el ACP se hace sin centrar y luego se eliminan el primer valor propio (que da 1) y el primer vector propio que es el centro de gravedad, esta operación la podemos llamar *centrado a posteriori*.

5.2.1.3. Distancia entre perfiles fila

En este análisis la matriz de producto interno que genera la métrica es $\mathbf{D}_p^{-1}$, cuyo elemento diagonal es $\dfrac{1}{f_{\cdot j}}$.

Dada esta matriz la distancia al cuadrado entre dos perfiles fila, i y l, es:

$$d^2(i,l) = \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} (x_{ij} - x_{lj})^2 = \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{lj}}{f_{l\cdot}} \right)^2 \tag{5.1}$$

La distancia (5.1), denominada distancia *ji cuadrado o de Benzècri*, amplifica más las diferencias al cuadrado entre coordenadas cuando se deben a columnas de baja frecuencia marginal. La distancia *ji* cuadrado le confiere al ACS dos propiedades: la equivalencia distribucional (sección 5.3.1) y las relaciones cuasi-baricéntricas (sección 5.3.2).

5.2.1.4. Inercia de la nube de perfiles fila

La inercia de la nube N_n, de los n puntos en $\mathbb{R}^p$ es:

$$Inercia(N_n) = \sum_{i=1}^{n} f_{i\cdot} d^2(i, \mathbf{g}_p) = \sum_{i=1}^{n} f_{i\cdot} \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j} \right)^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} \tag{5.2}$$

La inercia (5.2) es el coeficiente ϕ^2 (2.4), una medida de asociación entre las dos variables cualitativas. En las tablas de contingencia se suele probar independencia entre las dos variables cualitativas.

La hipótesis nula que se plantea es (Canavos, 1988, p.372):

$$H_0 : f_{ij} = f_{i\cdot} f_{\cdot j}; i = 1, 2, \cdots, n; j = 1, 2, \cdots, p$$

Bajo H_0 la estadística:

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{\left(k_{ij} - k f_{i\cdot} f_{\cdot j}\right)^2}{k f_{i\cdot} f_{\cdot j}} = k\, Inercia(N_n)$$

tiende a una distribución χ^2 con $(n-1)(p-1)$ grados de libertad. En el ejemplo, a partir de $\mathbf{F} - \mathbf{A}$ y $\mathbf{A}$ (tabla 5.4) se puede calcular la inercia.

En R: `sum((F-A)^2/A)`, su valor es 0.0656, de modo que la χ^2 calculada es $\chi^2_c = 445 \times 0.0656 = 29.19$.

El valor p se encuentra con la distribución χ^2 de parámetro 12 grados de libertad: $(7-1)(3-1)$. En R, este valor p se obtiene con el comando: `pchisq(29.19,12,lower.tail =FALSE)`, cuyo resultado es 0.0037. Entonces, la decisión estadística es rechazar H_0.

Recordemos que la media y la varianza de la distribución χ^2 son los grados de libertad y 2 veces los grados de libertad, respectivamente. Entonces la distribución χ^2_{12} se puede aproximar a una normal con media 12 y varianza 24 ($\sigma = 4.9$), $\mu + 3\sigma = 26.7$, otra forma de ver que χ^2_c se encuentra en la zona de rechazo porque $29.19 > 26.7$. El lector puede verificar lo afirmado en este párrafo con el código de R, que se muestra a continuación.

Código para comparar las dos distribuciones y realizar la prueba de independencia para K

```
curve(dchisq(x,12),xlim=c(0,30),las=1)
curve(dnorm(x,12,4.9),col="blue",add=TRUE)
abline(v=c(26.7,29.19),col="orange")
chisq.test(K)  # prueba de independencia
```

5.2.1.5. Búsqueda de los nuevos ejes en el espacio de las filas

La matriz de inercia $\mathbf{X'DXM}$ es (5.3) y su término general es 5.4.

$$\mathbf{F'D}_n^{-1}\mathbf{D}_n\mathbf{D}_n^{-1}\mathbf{FD}_p^{-1} = \mathbf{F'D}_n^{-1}\mathbf{FD}_p^{-1} \tag{5.3}$$

$$\sum_{i=1}^{n} \frac{f_{ij}}{f_{i\cdot}} \frac{f_{ij'}}{f_{\cdot j'}} \tag{5.4}$$

Para mostrar que (5.3) tiene el valor propio 1 asociado al centro de gravedad $\mathbf{g}_p = [f_{\cdot 1} \cdots f_{\cdot j} \cdots f_{\cdot p}]'$ se debe cumplir:

$$\mathbf{F}'\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_p^{-1}\mathbf{g}_p = \mathbf{g}_p$$

Lo que se puede ver con el término general:

$$\sum_{j'=1}^{p} \sum_{i=1}^{n} \frac{f_{ij}}{f_{i\cdot}} \frac{f_{ij'}}{f_{\cdot j'}} f_{\cdot j'} = f_{\cdot j}$$

Como $\mathbf{g}_p$ es un vector propio, $\mathbf{D}_p^{-1}$ ortogonal a los demás vectores propios, se puede retirar y las coordenadas de los perfiles fila sobre los ejes factoriales no cambian. Quitar el vector propio $\mathbf{g}_p$ de la nueva base equivale a centrar la nube de puntos. Con esto se pierde una dimensión y la nube queda soportada en el subespacio de dimensión $min(n, p) - 1$.

En el ejemplo la nube de perfiles fila está soportada en $\mathbb{R}^2$, así que en el primer plano factorial se comparan los perfiles sin perder información.

El valor propio 1 no entra en el análisis, pero corresponde a la norma al cuadrado de $\mathbf{g}_p$, y los demás valores propios son menores que 1, demostración que se puede ver en Lebart, Piron & Morineau (2006, p. 149).

5.2.1.6. Ejes y subespacios vectoriales

La nube de perfiles fila se observa mediante las proyecciones sobre ejes y planos factoriales. Algunos utilizan representaciones en 3D ($\mathbb{R}^3$). En el ejemplo toda la información está en el primer plano factorial (figura 5.1).

5.2.2. ACP de los perfiles-columna

Los histogramas de las distribuciones condicionales columna se representan como puntos en $\mathbb{R}^n$. A cada punto j se le asigna el peso $f_{\cdot j}$.

El análisis de los perfiles columna es el ACP($\mathbf{D}_p^{-1}\mathbf{F}'$, $\mathbf{D}_n^{-1}$, $\mathbf{D}_p$). Este análisis es simétrico al de perfiles fila, y es un buen ejercicio para el lector, verificar cada una de las secciones cambiando los subíndices.

Código para ejecutar ACS y obtener la figura 5.1

```
acs<-dudi.coa(K,scannf=FALSE)
plot(acs,Tcol=FALSE,xlim=c(-0.7,0.3),cframe=1)
# si se desea grabar la grafica en formato xfig para editarla
#dev.print(device = xfig,file="cienciasACScarreras12.fig")
```

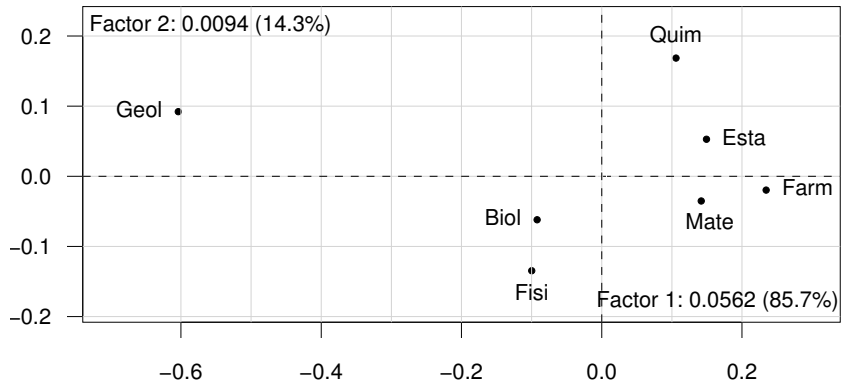


Figura 5.1. Primer plano factorial de los perfiles de carreras según estratos. Geología tiene el perfil más diferente del promedio y de las demás carreras; Biología se parece más al perfil promedio, Matemáticas, Estadística y Farmacia tienen perfiles parecidos. El primer eje retiene el 85.7 % de la inercia

En la figura 5.2 se muestra el primer plano factorial de los estratos, que son las categorías columna en el ejemplo.

5.2.3. Representación simultánea

Los dos ACP están relacionados debido a que los perfiles fila y columna se derivan de la misma matriz **F**, y la inversa de la matriz de pesos en un espacio es la métrica en el otro.

Las relaciones de transición entre los dos espacios (sección 5.3.2) permiten la representación simultánea de los mapas factoriales. La deducción de las relaciones de transición es más fácil cuando se ve el ACS como un ACP, como se muestra en la sección siguiente.

En cada ACP del ACS los mapas factoriales y sus ayudas a la interpretación son análogos a los de los individuos en el ACP clásico.

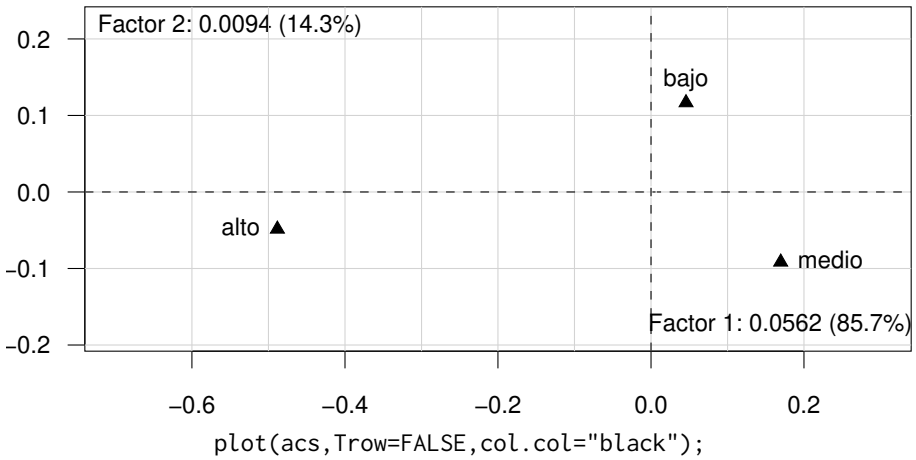


Figura 5.2. Primer plano factorial de los perfiles de estratos según carreras. El primer eje opone el estrato alto con el medio y el segundo eje, sobre todo, el bajo con el medio. El estrato alto es el que más se diferencia del promedio

5.3. El ACS como un ACP(X,M,N)

El ACS de la tabla **F** también se obtiene mediante el ACP de la tabla **X**, cuyo término general está dado por (5.5), usando $\mathbf{N} = \mathbf{D}_n = diag(f_{i\cdot})$ como pesos de las filas y matriz de métrica en el espacio de las columnas y $\mathbf{M} = \mathbf{D}_p = diag(f_{\cdot j})$ como pesos de las columnas y matriz de métrica en el espacio de las filas.

$$x_{ij} = \frac{f_{ij} - f_{i\cdot} f_{\cdot j}}{f_{i\cdot} f_{\cdot j}} \tag{5.5}$$

Todas las fórmulas del ACS se pueden derivar de las fórmulas correspondientes al ACP generalizado (tabla 4.1). La **M**-distancia al cuadrado entre dos filas i y l y la **D**-distancia al cuadrado entre las columnas j y k de **X** son:

$$d^2(i,l) = \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{lj}}{f_{l\cdot}} \right)^2; \quad d^2(j,k) = \sum_{i=1}^{n} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ik}}{f_{\cdot k}} \right)^2 \tag{5.6}$$

La expresiones de (5.6) son las mismas distancias ji-cuadrado entre los perfiles fila (5.1) y columna, derivados de sus respectivos ACP. Estas distancias tienen dos propiedades muy importantes para la interpretación de las salidas del ACS: equivalencia distribucional y relaciones cuasibaricéntricas.

5.3.1. Equivalencia distribucional

El ACS no se modifica si se unen dos puntos que tienen el mismo perfil. El peso del punto colapsado es la suma de los pesos de los puntos que se unen. Esto permite unir filas o columnas con perfiles parecidos, para simplificar las tablas originales. Por ejemplo, las carreras de Estadística, Matemáticas y Farmacia se pueden colapsar en un punto, y las carreras Biología y Física en otro (figura 5.3).

Esta propiedad hace que el ACS sea robusto ante la "arbitrariedad" en la conformación de las categorías de una variable en un estudio. En Lebart *et al.* (2006, p. 145) se puede ver una demostración formal de esta propiedad.

5.3.2. Relaciones cuasibaricéntricas

Nótese, por ejemplo, que los estudiantes 25 y 125 tienen distancia cero, es decir que, asumen las mismas categorías para las 4 variables, lo que se puede verificar en la tabla 6.1. Lo mismo sucede para las parejas 75 y 275, 250 y 325, 350 y 375. En el ACS la relación de transición que expresa la coordenada de un perfil fila i en función de las coordenadas de los perfiles columna es:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{p} \frac{f_{ij}}{f_{i\cdot}} G_s(j) \tag{5.7}$$

En (5.7) un sumando j es $\frac{f_{ij}}{f_{i\cdot}} G_s(j)$, donde $\frac{f_{ij}}{f_{i\cdot}}$ es la coordenada j del perfil de la fila i, es decir la altura de la barra j del histograma.

Como $\sum_{j=1}^{p} \frac{f_{ij}}{f_{i\cdot}} = 1$, la sumatoria de (5.7) es un promedio ponderado de las coordenadas de las columnas, es decir, un centro de gravedad o baricentro.

La multiplicación por $\frac{1}{\sqrt{\lambda_s}}$. dilata, es decir, aleja la coordenada del perfil fila del baricentro, razón por la que se denomina *relación o fórmula cuasibaricéntrica* a ecuación (5.7).

Las proyecciones de los perfiles cambian porque las ponderaciones son diferentes, están dadas por las proporciones de cada perfil.

De forma simétrica la sumatoria de la fórmula (5.8) para una columna j es el promedio de las coordenadas de todos los puntos fila sobre un eje s,

ponderado por los valores del perfil j:

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{n} \frac{f_{ij}}{f_{\cdot j}} F_s(i) \tag{5.8}$$

En el primer plano, por ejemplo, el punto promedio se ubica dentro del polígono que une a los puntos columna, pero por las dilataciones $\frac{1}{\sqrt{\lambda_s}}$, donde $s = 1, 2$; el punto puede ubicarse afuera.

Las dilataciones de las fórmulas cuasibaricéntricas son las que hacen posible la representación simultánea de los dos espacios sobre los ejes y planos factoriales. Además, permiten la interpretación de las posiciones de puntos fila y columna como una doble atracción o jalonamiento. Por ejemplo, un punto fila se ubica más cerca de los puntos de las columnas que más contribuyen a su perfil.

La dilatación hace que la asociación más destacada sea también la más alejada. Por ejemplo, en la figura 5.3 Geología es atraída por el estrato alto y viceversa. Geología está más alejada del origen que el estrato alto: el porcentaje de estrato alto en el perfil de Geología es de 40 % (marginal 18.2 %), mientras que el porcentaje de Geología en el estrato alto es de 22.2 %, (marginal 10.1 %) (tabla 5.2).

Para entender mejor las relaciones cuasibaricéntricas, calculemos la coordenada del perfil de Geología sobre el primer eje (-0.6):

- El perfil de Geología es [0.4 0.2 0.4] (tabla 5.2). Las coordenadas de los estratos sobre el primer eje son [0.0458 0.1695 -0.4884] (tabla de la figura 5.3).
- El primer valor propio es 0.0562 (figura 5.3).

  La fórmula (5.7) queda

$$\frac{1}{\sqrt{0.0562}} * (0.4 * 0.0458 + 0.2 * 0.1695 - 0.4 * 0.4884)$$

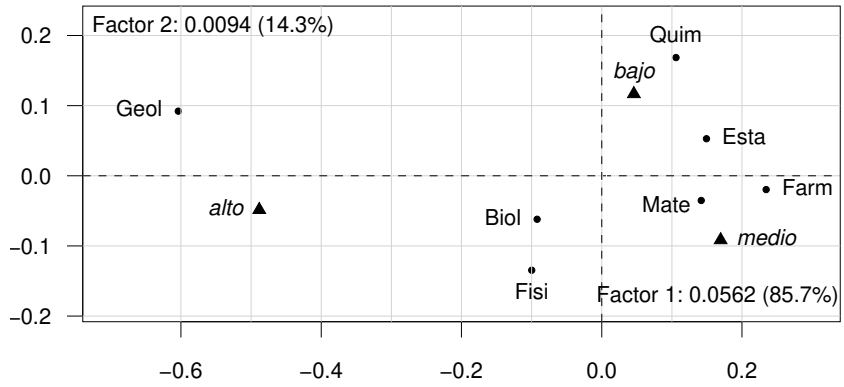
$$= 4.2182 * (0.0118 + 0.0339 - 0.1954) = 4.2182 * (-0.1497) = -0.6315$$

El promedio ponderado por el perfil de Geología es -0.1497. Se aleja del centro debido a la la dilatación por 4.2182. Nótese que la coordenada de estrato alto es la que más suma por dos efectos: la ponderación (0.4) y su alejadamiento del origen (-0.4884). La diferencia de la coordenada calculada

-0.6315 con el valor -0.6037 del programa (figura 5.3) se debe a los errores de redondeo.

Cálculo en R con más cifras significativas

```
1/sqrt(acs$eig[1])*sum(c(0.4,0.2,0.4)*acs$co[,1])
[1] -0.603727
```




```
plot(acs,xlim=c(-0.7,0.3),cframe=1,col.col="black");
```

Coordenadas y ayudas para la interpretación

Carrera	Coordenadas		Contribuciones		Cosenos^2		Contribución
	Eje 1	Eje 2	Eje1	Eje 2	Eje1	Eje 2	Inercia en $\mathbb{R}^2$
Biología	-0.0922	-0.0620	2.1	5.8	68.9	31.1	2.7
Estadística	0.1494	0.0528	5.9	4.4	88.9	11.1	5.7
Farmacia	0.2345	-0.0196	16.1	0.7	99.3	0.7	13.8
Física	-0.0998	-0.1347	3.3	35.6	35.4	64.6	7.9
Geología	-0.6037	0.0922	65.6	9.2	97.7	2.3	57.5
Matemáticas	0.1417	-0.0352	4.3	1.6	94.2	5.8	3.9
Química	0.1059	0.1685	2.8	42.8	28.3	71.7	8.6
Estrato							
bajo2	0.0458	0.1167	1.5	58.3	13.3	86.7	9.6
medio3	0.1695	-0.0916	21.3	37.2	77.4	22.6	23.5
alto4	-0.4884	-0.0485	77.2	4.6	99.0	1.0	66.8

Figura 5.3. Primer plano factorial del ACS carreras × estratos y ayudas para la interpretación. La posición de Geología se debe a que tiene, con relación al promedio, mayor porcentaje de estrato alto; Química tiene mayor de estrato bajo; y Farmacia, mayor de estrato medio. Biología es la carrera con perfil más parecido al promedio

5.3.3. Ayudas para la interpretación

Las ayudas para la interpretación de los individuos están disponibles para el ACS. Los perfiles fila son análogos a individuos en el primer ACP y los perfiles columnas a individuos en el segundo ACP.

5.3.3.1. Contribución absoluta

La contribución de un perfil a la varianza del eje (inercia proyectada), depende del peso y de la coordenada al cuadrado:

$$Ca_s(i) = \frac{f_{i\cdot}(F_s(i))^2}{\lambda_s} \tag{5.9}$$

En el ejemplo (figura 5.3), en la nube de carreras, la dirección del primer eje se debe sobre todo a Geología y Farmacia (81.7 % de contribución) y la del segundo eje, a Química y Física (78.4 % de contribución).

5.3.3.2. Coseno cuadrado, calidad de la representación o contribución relativa

Es el cociente de los cuadrados para cada punto perfil de la longitud proyectada sobre un eje s y la del vector perfil en $\mathbb{R}^p$:

$$Cos_s^2(i) = \frac{F_s^2(i)}{d^2(i, \mathbf{g})} \tag{5.10}$$

Es decir, en la coordenada al cuadrado de cada punto sobre la longitud al cuadrado en el espacio completo, que es la norma al cuadrado del vector centrado, o la distancia al cuadrado entre el punto y el centro de gravedad. El nombre de *contribución relativa* se da porque el coseno cuadrado es también un cociente de contribuciones a la inercia. Multiplicando numerador y denominador de (5.10) por $f_{i\cdot}$:

$$Cos_s^2(i) = \frac{f_{i\cdot}F_s^2(i)}{f_{i\cdot}d^2(i, \mathbf{g})}$$

que es la contribución a la inercia de la categoría i en el eje s sobre la su contribución en el espacio completo $\mathbb{R}^p$.

La fórmulas para las columnas se obtienen de forma simétrica, es decir cambiando símbolos y subíndices.

5.4. Ejemplo de aplicación de ACS

Resultados de los exámenes de Estado de la educación básica en Colombia según departamentos. Con los resultados del examen de Estado realizado por el Icfes en Colombia durante 2008 se construyó una TC a partir de la clasificación de los colegios según los resultados de sus estudiantes. La tabla se encuentra en Pardo, Bécue-Bertaut & Ortiz (2013) y está disponible en `icfes08{FactoClass}`. La tabla tiene 29 departamentos e incluye una fila que agrupa los departamentos de menos de 100 mil habitantes P01: SAP, AMA,VID,VAU,GUA. Los departamentos se estructuran en 4 grupos según su población, en millones de habitantes: P5: más de 2, P4: entre 1 y 2, P3: entre 0.5 y 1 y P2: menor de 0.5.

La TC tiene doce columnas, que es la clasificación combinada de las jornadas del colegio: completa, mañana y tarde, y sus categorías de rendimiento: inferior, medio, bajo y alto.

5.4.1. Objetivos del análisis

El objetivo principal del análisis es comparar los perfiles departamentales según la calidad educativa de los colegios. También se desea explorar la influencia de la jornada sobre el ordenamiento de los departamentos y la influencia de su tamaño, en términos de población, en ese ordenamiento.

5.4.2. Perfiles de los departamentos

El ACS ordena a los departamentos según sus perfiles, construidos con las variables *jornada* y *rendimiento*, de los colegios. Esos perfiles se muestran en la figura 5.4. La longitud de cada barra del histograma de la fila, es proporcional a la frecuencia relativa de la categoría jornada × rendimiento y está diferenciada de las demás por su escala de cuatro grises, que se repite en cada jornada. La escala de grises permite diferenciar también las categorías asociadas.

Nótese que Chocó se diferencia bastante de los demás departamentos: tiene las barras de *rendimiento inferior* de mayor longitud en las tres jornadas.

Esto se evidencia en el ACS de la tabla, pues muestra al Choco alejado de los demás departamentos. Entonces se toma al Chocó como ilustrativo para poder ver mejor las diferencias de los demás departamentos.

5.4.3. Resultados del ACS

Número de ejes a analizar

La inercia total asociada al ACS es 0.265. Los tres primeros ejes retienen el 84.2 % (0.151, 59.9 %; 0.049, 18.7 %, y 0.023, 8.6 %). Los dos primeros retienen una inercia superior a la inercia promedio (0.024 = 0.265/11) (figura 5.5, arriba). Los dos primeros ejes proveen una buena síntesis para analizar las asociaciones entre departamentos y jornadas × rendimiento. Sin embargo, se incluyen las coordenadas y ayudas para la interpretación del tercer eje con el objeto afinar un poco el análisis.

Primer plano factorial y tercer eje

El primer eje ordena las categorías de rendimiento de menor (izquierda) a mayor (derecha), y por lo tanto, Bogotá es el departamento de mejor rendimiento y Magdalena, el de peor rendimiento aparte de Chocó, que ya se había detectado con un perfil atípico por su rendimiento muy bajo.

El segundo eje muestra arriba las categorías inferior, baja y media de la jornada completa, lo que se debe a una atracción de los departamentos que tienen más porcentaje de colegios con esa jornada en su perfil, que en el promedio.

En las tablas 5.5 y 5.6 se muestran las ayudas para la interpretación. Las contribuciones a la inercia de los ejes sirven para detectar los puntos que son más relevantes en cada eje. El coseno cuadrado se utiliza para ver las calidades de las proyecciones. El coseno cuadrado sobre un plano es la suma de los cosenos cuadrados de los ejes.

El primer plano factorial presenta el *efecto Guttman*, que es una forma de parábola de las categorías de una variable ordinal. Nótese que por cada una de las tres jornadas, se ven las categorías de rendimiento como parábolas invertidas. En este efecto el primer eje opone los rendimientos extremos (inferior vs. alto) y el segundo eje, los medios a los extremos (bajo y medio vs. inferior y alto).

Código para cargar datos y obtener la figura 5.4

```
library(FactoClass)
data(icfes08)
par(mai=c(0,1,0,0),cex=0.6) # margenes de la grafica
plotct(icfes08,"row",col=gray.colors(4,0.5,0.9),legend=FALSE)
# para grabar la grafica en formato xfig y editarla
#dev.print(device = xfig,file="TCicfesPerpilesDeptos.fig")
```

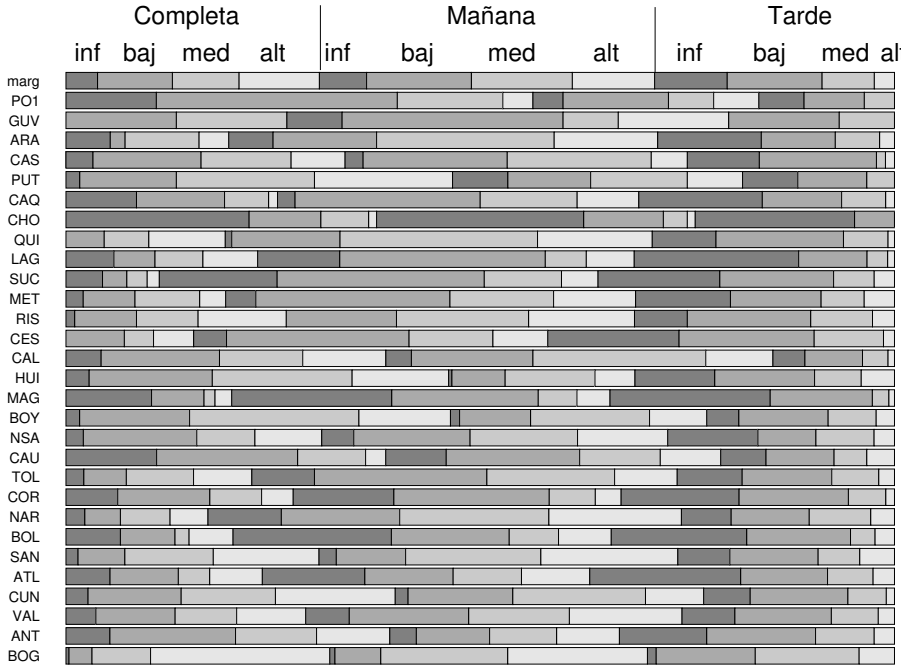
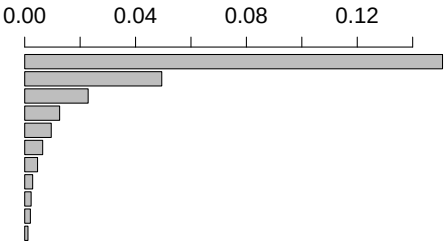


Figura 5.4. Perfiles de los departamentos según *jornadas* × *rendimiento*

En el tercer eje se destacan del lado positivo las categorías baja y media de la jornada de la mañana, asociadas a los departamentos de Meta, Quindío, Tolima y Arauca.

Código para obtener la figura 5.5

```
tab<-icfes08[-23,]; # Choco no activa
acs<-dudi.coa(tab,scannf=FALSE,nf=3);
barplot(acs$eig); #histograma valores propios
#dev.print(device = xfig,file="ACSicfesValP.fig"); # exportar a xfig
valp<-t(inertia.dudi(acs)$TOT); #valores propios
xtable(valp,digits=rep(3,12)); # obtener formato tabular LaTeX
plot(acs,cframe=1,xlim=c(-1.2,0.8)); # primer plano
# proyeccion de Choco como ilustrativa
Fchoco<-suprow(acs,icfes08[23,])$lisup;
points(Fchoco,col="brown");
text(Fchoco,"CHO",col="brown",pos=1,cex=0.8);
#dev.print(device = xfig,file="ACSicfesP12.fig");
plot(acs,2,3,cframe=1.1)#,xlim=c(-1.2,0.8)); # plano 2-3
#dev.print(device = xfig,file="ACSicfesP23.fig");
```



	1	2	3	4	5	6	7	8	9	10	11
valor P	0.151	0.049	0.023	0.013	0.010	0.006	0.005	0.003	0.002	0.002	0.001
iner. acu.	0.151	0.200	0.223	0.236	0.245	0.252	0.256	0.259	0.262	0.264	0.265
prop. acu.	0.569	0.756	0.842	0.890	0.926	0.951	0.968	0.979	0.988	0.996	1.000

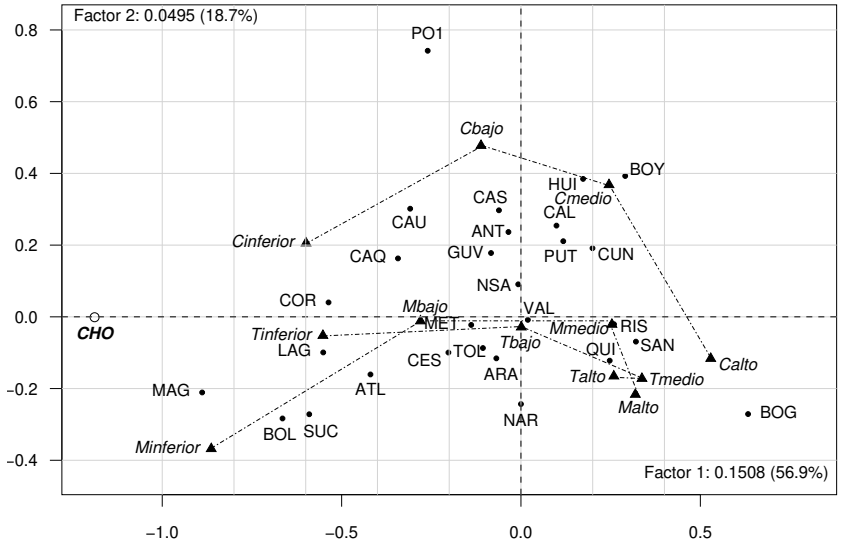


Figura 5.5. Primer plano factorial del ACS de la TC departamentos × categorías de rendimiento. Arriba valores propios con su histograma

Código para obtener las ayudas para la interpretación del ACS de la TC de departamentos × (jornadas, nivel de colegios) (tablas 5.5 a 5.6)

```
ayu<-inertia.dudi(acs,T,T); # ayudas
# Pesos, contribuciones y contribuciones a la inercia
# Departamentos
xtable(data.frame(peso=acs$lw*100,
  coniner=ayu$row.contrib,acs$li),digits=rep(2,6));
# columnas
xtable(data.frame(peso=acs$cw*100,
  coniner=ayu$col.contrib,acs$co),digits=rep(2,6));
# Contribuciones a la inercia y cosenos cuadrados
# Departamentos
xtable(data.frame(con=ayu$row.abs,cos=abs(ayu$row.rel)),
  digits=rep(2,7));
# Columnas
xtable(data.frame(con=ayu$col.abs,cos=abs(ayu$col.rel)),
  digits=rep(2,7))
```

El ordenamiento de las categorías se traslada a los departamentos mostrando una parábola invertida. Siguiéndola desde la izquierda hasta la derecha se observa que los departamentos de la región Atlántico son los de menor rendimiento, siguen los departamentos del sur del país, luego los de la región Andina, y sobresalen Risaralda, Quíndio y Santander, hasta llegar a Bogotá, la región de mayor rendimiento.

Retorno a los perfiles de departamentos

Se utilizan las coordenadas sobre el primer eje para presentar los perfiles ordenando los departamentos, de modo que los parecidos queden vecinos en la figura 5.6. Los perfiles de la parte inferior son los peores y van mejorando a medida que se sube en la gráfica. El perfil superior es el marginal que se incluye como referencia para la comparación ya que se sitúa en el origen de la representación –coordenadas (0,0) en los planos– y la lejanía del centro de un punto indica que el perfil que representa es el más diferente del perfil promedio (marginal).

Tabla 5.5. Pesos, contribuciones a la inercia y coordenadas del ACS de la TC departamentos × (jornadas, nivel de los colegios)

		Contrib.	Coordenadas		
Depto.	Peso	Inercia	Eje 1	Eje 2	Eje 3
BOG	14.21	26.49	0.63	-0.27	-0.12
ANT	11.80	4.57	-0.03	0.24	-0.18
VAL	9.66	0.95	0.02	-0.01	0.07
CUN	7.08	3.12	0.20	0.19	0.07
ATL	5.79	5.60	-0.42	-0.16	-0.17
SAN	4.75	2.81	0.32	-0.07	0.04
BOL	4.68	10.14	-0.66	-0.28	-0.17
NAR	3.48	2.13	-0.00	-0.24	0.19
COR	2.86	3.35	-0.54	0.04	-0.02
TOL	3.68	1.50	-0.11	-0.09	0.26
CAU	3.27	4.14	-0.31	0.30	-0.04
NSA	2.83	0.71	-0.01	0.09	0.05
BOY	3.60	4.25	0.29	0.39	0.02
MAG	2.98	9.87	-0.89	-0.21	-0.13
HUI	2.48	2.42	0.17	0.38	-0.15
CAL	2.58	1.82	0.10	0.25	0.19
CES	2.26	2.31	-0.20	-0.10	0.23
RIS	1.87	0.80	0.26	-0.02	0.14
MET	1.91	1.27	-0.14	-0.02	0.32
SUC	2.03	3.82	-0.59	-0.27	0.21
LAG	1.20	2.04	-0.55	-0.10	0.12
QUI	1.29	1.26	0.25	-0.12	0.31
CAQ	0.93	1.11	-0.34	0.16	0.21
PUT	0.60	0.52	0.12	0.21	-0.07
CAS	0.91	0.65	-0.06	0.30	0.24
ARA	0.56	0.44	-0.07	-0.12	0.21
GUV	0.15	0.28	-0.08	0.18	0.39
PO1	0.55	1.62	-0.26	0.74	-0.18
		Contrib.	Coordenadas		
Columna	Peso	Inercia	Eje 1	Eje 2	Eje 3
Cinferior	3.63	7.97	-0.60	0.21	-0.23
Cbajo	9.04	9.19	-0.11	0.48	-0.10
Cmedio	8.04	7.41	0.25	0.37	0.02
Calto	9.82	13.26	0.53	-0.12	-0.17
Minferior	5.50	20.08	-0.86	-0.37	-0.13
Mbajo	12.67	8.17	-0.28	-0.01	0.28
Mmedio	12.29	5.76	0.25	-0.02	0.19
Malto	10.01	7.30	0.32	-0.22	0.02
Tinferior	8.66	12.60	-0.55	-0.05	-0.07
Tbajo	11.53	1.22	0.00	-0.03	-0.02
Tmedio	6.36	4.95	0.34	-0.17	-0.11
Talto	2.46	2.09	0.26	-0.17	-0.16

Tabla 5.6. Contribuciones absolutas y cosenos cuadrados del ACS de la TC departamentos × (jornadas, nivel de los colegios)

	Contribuciones			Cosenos^2		
Depto.	Eje 1	Eje 2	Eje 3	Eje 1	Eje 2	Eje 3
BOG	37.76	21.14	8.82	81.15	14.90	2.88
ANT	0.10	13.37	15.80	1.19	54.67	29.90
VAL	0.02	0.01	2.10	1.41	0.27	19.02
CUN	1.87	5.24	1.48	34.03	31.36	4.11
ATL	6.76	3.01	7.43	68.70	10.05	11.46
SAN	3.23	0.46	0.29	65.50	3.04	0.88
BOL	13.73	7.61	5.63	77.07	14.01	4.80
NAR	0.00	4.16	5.59	0.00	36.39	22.66
COR	5.47	0.09	0.07	92.81	0.52	0.18
TOL	0.28	0.56	11.19	10.45	7.00	64.49
CAU	2.07	6.02	0.24	28.47	27.17	0.49
NSA	0.00	0.47	0.32	0.09	12.48	3.88
BOY	2.02	11.23	0.04	26.98	49.32	0.08
MAG	15.64	2.68	2.19	90.18	5.08	1.92
HUI	0.49	7.41	2.40	11.58	57.24	8.57
CAL	0.17	3.37	3.97	5.25	34.71	18.92
CES	0.61	0.45	5.35	15.10	3.65	20.02
RIS	0.83	0.02	1.69	59.07	0.37	18.20
MET	0.24	0.02	8.46	10.93	0.29	57.70
SUC	4.69	3.02	3.77	69.85	14.77	8.52
LAG	2.43	0.24	0.70	67.70	2.20	2.99
QUI	0.52	0.39	5.49	23.68	5.78	37.63
CAQ	0.73	0.50	1.72	37.39	8.43	13.37
PUT	0.06	0.54	0.11	6.03	19.29	1.87
CAS	0.02	1.63	2.37	2.02	47.10	31.61
ARA	0.02	0.15	1.08	2.27	6.44	21.42
GUV	0.01	0.10	0.97	1.40	6.36	29.87
PO1	0.25	6.09	0.74	8.60	70.03	3.96

	Contribuciones			Cosenos^2		
Categoría	Eje 1	Eje 2	Eje 3	Eje 1	Eje 2	Eje 3
Cinferior	8.64	3.11	8.48	61.72	7.27	9.20
Cbajo	0.75	41.68	4.07	4.62	84.72	3.83
Cmedio	3.21	22.02	0.08	24.68	55.52	0.10
Calto	18.21	2.69	12.97	78.21	3.79	8.46
Minferior	27.19	15.05	4.14	77.08	13.99	1.78
Mbajo	6.60	0.04	42.40	45.98	0.08	44.87
Mmedio	5.24	0.11	19.26	51.79	0.36	28.91
Malto	6.75	9.48	0.13	52.61	24.23	0.16
Tinferior	17.52	0.48	1.81	79.15	0.70	1.24
Tbajo	0.00	0.19	0.18	0.00	2.92	1.30
Tmedio	4.80	3.80	3.66	55.13	14.34	6.39
Talto	1.09	1.36	2.82	29.69	12.13	11.66

Código para obtener la figura 5.6

```
ordep<-order(acs$li[,1]);
par(mai=c(0,1,0,0));
plotct(tab[ordep,],"row",col=gray.colors(4,0.5,0.9),
  legend=FALSE);
#dev.print(device = xfig,file="ACSicfesPerfilesOrdenados.fig");
```

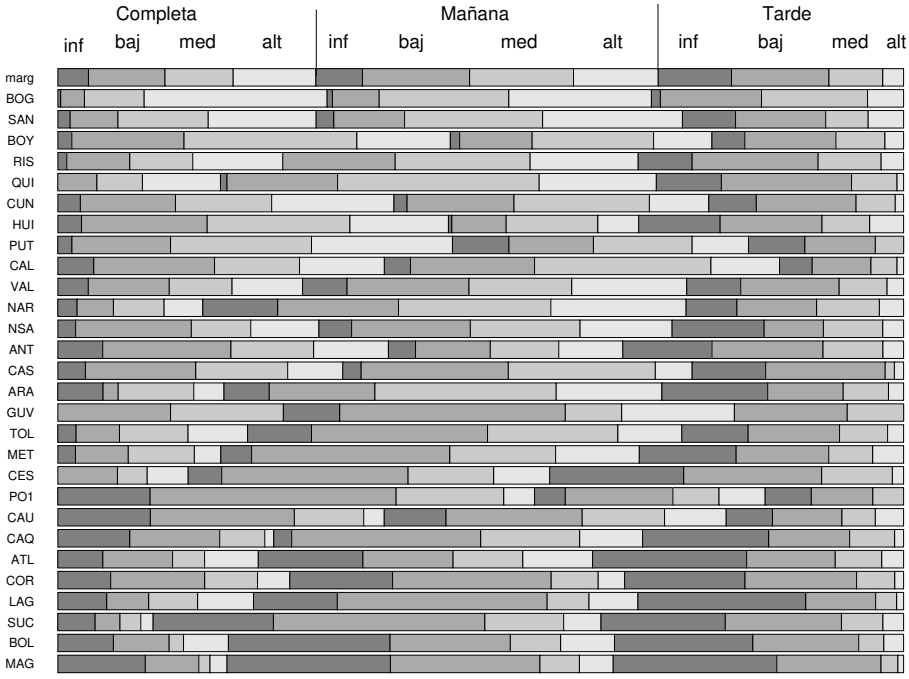


Figura 5.6. Perfiles de los departamentos ordenados por las coordenadas sobre el primer eje del ACS. Los departamentos de la parte superior son los de mayor rendimiento

5.4.4. Conclusiones del análisis

Los departamentos se ordenan en el primer plano factorial según el rendimiento de sus colegios. Bogotá es el de mayor rendimiento; Chocó es de lejos el de menor rendimiento. La figura 5.6 es un buen resumen del ordenamiento de los departamentos según su rendimiento en los resultados del examen realizado por el Icfes.

Cada jornada ordena los departamentos de manera similar, pero la jornada de la tarde es la de menor dispersión.

En Pardo *et al.* (2013) se pueden ver algunas variaciones del análisis de correspondencias para tener en cuenta la estructura inducida por los cuatro tamaños de los departamentos, según su población, y las tres jornadas.

5.5. Ejercicios

1. Demostrar la propiedad de la equivalencia distribucional.
2. Mostrar que las distancias entre dos categorías fila de un AC visto como un ACP es igual a la distancia ji-cuadrado del ACP de los perfiles fila.x
3. Demostrar que los valores propios en un ACS son inferiores a uno.
4. Demostrar que la matriz de inercia del ACP de los perfiles fila tiene un valor propio igual a uno y su vector propio asociado es el centro de gravedad de la nube de perfiles fila.
5. Demostrar que la estadística χ^2 de una tabla de contingencia es igual al total de la TC por la inercia asociada al ACS.
6. En el ACS de la TC *carrera* $\times$ *estrato* encontrar la matriz de inercia en el espacio de las carreras y calcularle los valores y vectores propios en R. Verificar que hay un valor propio igual a uno y que el centro de gravedad es un vector propio asociado a este.

5.6. Talleres de ACS

5.6.1. ACS de la TC manzanas de Bogotá según localidades y estratos

Objetivo

Describir la estratificación de Bogotá a partir de la TC del número de manzanas según localidades × estratos (DAPD, 1997, p.77).

Los datos

La TC que clasifica a las manzanas de Bogotá en localidad × estrato, se encuentra en `Bogota{FactoClass}`. La primera columna de la TC corresponde a manzanas que no están estratificadas, porque no son residenciales (parques, colegios, etc.). Esta columna se proyecta como ilustrativa en el ACS.

Preguntas

Realizar el ACS de la TC utilizando los estratos del uno al seis como columnas (frecuencias) activas y la columna *sin estrato* como ilustrativa. Responder a las preguntas siguientes:

1. Comente la repartición de las manzanas según estratos –histograma de la distribución de las manzanas en los seis estratos (distribución marginal)–.

2. ¿Cómo es la distribución de las manzanas según localidades (distribución marginal)?

3. ¿Utilizaría la columna sin estrato como activa en un análisis de correspondencias?; ¿Porque sí? ¿Por qué no?

4. Compare la estadística χ^2 asociada a la tabla de contingencia con la teórica. ¿Hay asociación entre estratos y localidades?

5. ¿Cuántos ejes retiene para el análisis? ¿Por qué?

6. Identifique en el primer eje las localidades más contributivas y sus oposiciones (localidades con coordenadas de signo negativo sobre el eje vs. las de signo positivo).

7. Identifique los estratos más contributivos al primer eje y sus oposiciones.

8. Repita 6 para el segundo eje.

9. Repita 7 para el segundo eje.

10. Resuma la comparación de los perfiles de las localidades utilizando el primer plano factorial.

11. Resuma la comparación de los perfiles de los estratos utilizando el primer plano factorial.

12. Según el primer plano factorial, ¿cómo es la asociación entre localidades y estratos?

13. ¿Hay efecto Guttman? Explique.

14. ¿Hay contraposiciones en el tercer eje que no se observen en el primer plano factorial? Según lo anterior, ¿vale la pena interpretar el tercer eje?

15. Agregue a los datos una columna de orden de las localidades según el primer plano factorial. Ordene la TC por esa variable y haga una gráfica que muestre los perfiles de las localidades así ordenadas y el perfil promedio. No incluya la columna sin estrato. Resuma la comparación de los perfiles utilizando esta gráfica y el primer plano factorial.

16. Proponga un partición de las localidades en cinco clases y haga una gráfica de perfiles incluyendo el perfil marginal. Como otra síntesis del análisis, comente la gráfica obtenida.

17. Compruebe "a mano" (utilizando R) las relaciones de transición. Por ejemplo: calcule la coordenada sobre el eje 1 de Usme a partir de las coordenadas de los seis estratos. Analice el ejercicio (¿Quién atrae a quién y por qué?).

18. Resumen: describa la distribución geográfica de los habitantes de Bogotá según su nivel socio-económico, utilizando el estrato de la manzana donde vive cada uno como indicador de ese nivel.

5.6.2. ACS adjetivos × colores

Este ejemplo se encuentra en español en Fine (1996) y en inglés en Jambu (1983), de donde se tomó la tabla `ColorAdjective{FactoClass}`.

Objetivo

Una agencia de publicidad encarga un estudio sobre las asociaciones entre colores y adjetivos, para armonizar la publicidad de los productos con las imágenes que los compradores potenciales tienen de los colores.

Los datos

A cada encuestado se le pide que diga, para cada uno de 11 colores propuestos, cuál es el adjetivo que le parece corresponder lo mejor posible. Se conservan solamente los adjetivos que se han mencionado por lo menos tres veces. Las unidades estadísticas en el análisis son las asociaciones color-adjetivo, con las que se construye una tabla de contingencia de 89 filas (adjetivos) por 11 columnas (colores).

Preguntas

1. ¿Es posible determinar el número de personas encuestadas a partir de la tabla de contingencia adjetivos × colores?. En caso afirmativo, ¿cuántas son?

2. ¿Qué significa el total 1097 de la tabla de contingencia?.

3. ¿Cuántos ejes retiene para el análisis? ¿Por qué?

4. Teniendo en cuenta todos los ejes identifique en qué planos están mejor representados cada uno de los once colores.

5. Para cada color o grupo de colores identifique los adjetivos más asociados leyendo en el plano donde estén mejor representados.

6. Para cada color o grupos de colores presente gráficamente su perfil mostrando los adjetivos más asociados y reuniendo los de baja frecuencia en la categoría de *otros*.

7. Para un adjetivo cualquiera compruebe numéricamente la fórmula de transición (cuasibaricentro de las coordenadas de los once colores ponderadas por el perfil del respectivo adjetivo).

8. Escriba la conclusión del análisis (¿qué adjetivos se asocian más a cada color?).

Capítulo

seis

Análisis de correspondencias múltiples

El análisis de correspondencias múltiples (ACM) se utiliza para analizar tablas de individuos descritos por variables cualitativas. Una fuente importante de este tipo de tablas son las encuestas sociales, en donde se obtiene información a través de formularios, con preguntas estructuradas en temas. Dependiendo del objetivo del análisis, las preguntas de uno de los temas juegan el papel de activas. La estructura de asociaciones entre categorías de las preguntas activas, se puede explicar, de manera exploratoria mediante algunas de las preguntas de los otros temas, proyectándolas como ilustrativas.

Las variables cualitativas utilizadas en el ACM son las que generan particiones de los individuos, es decir, las que los dividen en grupos disyuntos. En los cuestionarios de encuestas se traduce en que cada pregunta solo se puede responder con una opción. En este caso, la respuestas de todos los individuos a la pregunta se pueden registrar en una columna. La tabla que resulta, de individuos por variables, se denominada *de código condensado* y no tiene propiedades numéricas. Entonces, desde el punto de vista teórico el ACM parte de la tabla de individuos con las categorías de todas las variables activas, que se denomina *tabla disyuntiva completa*. Las propiedades de estas tablas otorgan al ACM características especiales, que justifican su estudio como un método particular.

Los objetivos del ACM son:

1. Describir las asociaciones entre las categorías de las variables activas.
2. Descubrir patrones de individuos, que pueden dar origen a grupos de ellos.
3. Sintetizar en imágenes geométricas (planos factoriales) las asociaciones entre categorías y los posibles grupos de individuos.
4. Explorar la explicación de la estructura inducida por las categorías activas, con variables que juegan el papel de ilustrativas.

En ocasiones los ejes factoriales pueden originar indicadores sintéticos del tema que se está analizando. El ACM también se puede utilizar como un pretratamiento de los datos para aplicar luego métodos estadísticos multivariados aptos para variables continuas, como regresión y algunos métodos de agrupamiento y de discriminación. En ese caso las coordenadas factoriales de los ejes que se seleccionen juegan el papel de variables continuas.

6.1. Ejemplo: ACM de admitidos

Para ilustrar los conceptos del ACM utilizamos el ejemplo de la descripción de los 445 admitidos a la Facultad de Ciencias, para el semestre de 2013-I, datos disponibles en `admi{FactoClass}`. Se utilizan como variables activas las sociodemográficas disponibles:

1. Género: femenino, masculino.
2. Edad: 16 o menos, 17, 18, 19 o más.
3. Estrato: bajo, medio, alto;
4. Procedencia: Bogotá, Cundinamarca, otro.

Los objetivos de análisis son describir el espacio sociodemográfico de los admitidos, generado por estas cuatro variables, y explorar las diferencias en las condiciones sociodemográficas, entre los grupos de admitidos según las carreras.

6.2. Transformaciones de la tabla de datos

En esencia se utiliza la misma notación de Lebart *et al.* (2006), que se va presentando a medida que aparecen los distintos elementos.

6.2.1. Tabla de código condensado

La tabla de datos se denomina *de código condensado* (denotada por $\mathbf{Y}$) y no tiene significado numérico. Las n filas representan a los individuos y las s columnas, a las variables cualitativas. En el lenguaje del diseño de experimentos las columnas son factores y las categorías, los niveles de los factores. En R estas variables son de tipo factor. En la tabla 6.1 se muestran las filas múltiplos de 25, como un extracto de la tabla $\mathbf{Y}$.

6.2.2. Tabla disyuntiva completa

La tabla disyuntiva completa (TDC), denotada por $\mathbf{Z}$, es una tabla binaria de n individuos por p categorías, indicadora de las s particiones definidas por

las variables cualitativas. Su término general es:

$$z_{ij} = \begin{cases} 1 & \text{si el individuo } i \text{ asume la categoría } j, \\ 0 & \text{si no la asume.} \end{cases}$$

En la tabla 6.1 se muestra un extracto de la TDC para el ejemplo de los admitidos.

Código para obtener **Y**, **Z** y producir la tabla 6.1

```
library(FactoClass)
data(admi)
Y<-admi[,c(8,11,9,10)]
# Para tabla del texto, registros multiplos de 25
Z<-acm.disjonctif(Y); data.frame(Y,Z)[seq(0,nrow(Y),25),]
xtable(data.frame(Y,Z)[seq(0,nrow(Y),25),],digits=rep(0,17))
```

La TDC $\mathbf{Z}$ es una yuxtaposición de s tablas, en la que s es el número de variables:

$$\mathbf{Z} = [\mathbf{Z}_1\ \mathbf{Z}_2\ \cdots \mathbf{Z}_q\ \cdots\ \mathbf{Z}_s]$$

Cada $\mathbf{Z}_q$ es la matriz indicadora de la partición originada por la variable cualitativa q. Para el ejemplo $\mathbf{Z} = [\mathbf{Z}_1\ \mathbf{Z}_2\ \mathbf{Z}_3\ \mathbf{Z}_4]$ (tabla 6.1). Los grupos son disyuntos y su unión es igual al conjunto de los n individuos. Por esta razón en cada fila de $\mathbf{Z}_q$ hay un solo 1 y siempre hay un 1. Por lo tanto, la suma de la fila es 1. Por esta propiedad es que se le da el nombre de *tabla disyuntiva completa* (TDC).

Tabla 6.1. Extracto de las tablas: de código condensado **Y** y disyuntiva-completa **Z**

	Y				**Z**											
Ide	Ge	Ed	Es	Or	Ge $\mathbf{Z_1}$		Edad $\mathbf{Z_2}$				Estrato $\mathbf{Z_3}$			Origen $\mathbf{Z_4}$		
					F	M	16-	17	18	19+	ba	me	al	Bo	Cu	Ot
25	F	17	medio	Otro	1	0	0	1	0	0	0	1	0	0	0	1
50	M	18	bajo	Bogo	0	1	0	0	1	0	1	0	0	1	0	0
75	M	17	bajo	Bogo	0	1	0	1	0	0	1	0	0	1	0	0
100	M	18	medio	Bogo	0	1	0	0	1	0	0	1	0	1	0	0
125	F	17	medio	Otro	1	0	0	1	0	0	0	1	0	0	0	1
150	F	16om	bajo	Bogo	1	0	1	0	0	0	1	0	0	1	0	0
175	M	19oM	alto	Bogo	0	1	0	0	0	1	0	0	1	1	0	0
200	F	17	bajo	Otro	1	0	0	1	0	0	1	0	0	0	0	1
225	M	16om	alto	Otro	0	1	1	0	0	0	0	0	1	0	0	1
250	M	17	alto	Bogo	0	1	0	1	0	0	0	0	1	1	0	0
275	M	17	bajo	Bogo	0	1	0	1	0	0	1	0	0	1	0	0
300	M	19oM	bajo	Otro	0	1	0	0	0	1	1	0	0	0	0	1
325	M	17	alto	Bogo	0	1	0	1	0	0	0	0	1	1	0	0
350	M	19oM	medio	Bogo	0	1	0	0	0	1	0	1	0	1	0	0
375	M	19oM	medio	Bogo	0	1	0	0	0	1	0	1	0	1	0	0
400	F	18	bajo	Bogo	1	0	0	0	1	0	1	0	0	1	0	0
425	M	16om	alto	Bogo	0	1	1	0	0	0	0	0	1	1	0	0

Como hay s submatrices $\mathbf{Z}_q$, la suma de cada fila de $\mathbf{Z}$, es s, es decir, su marginal fila es un vector de n veces s y el total de $\mathbf{Z}$ es ns. En el ejemplo $s = 4$, $n = 445$ y el total de la tabla $4 * 445 = 1780$.

La suma de cada columna de $\mathbf{Z}$ es el número de individuos que asumen la categoría j que se nota n_j: $n_j = z_{\cdot j}$ y $\sum_{j \in Z_q} n_j = n$. Z_q es el conjunto de categorías de la variable q. En el ejemplo los valores de n_j se muestran en la diagonal de la matriz $\mathbf{B}$ (tabla 6.2).

Tabla 6.2. Tabla de Burt del ejemplo "Admitidos a Ciencias"

Categoría	Género		Edad				Estrato			Origen		
	F	M	16-	17	18	19+	ba	me	al	Bo	Cu	Ot
Ge.F	128	0	46	45	18	19	46	59	23	89	9	30
Ge.M	0	317	72	126	38	81	133	126	58	222	29	66
Ed.16menos	46	72	118	0	0	0	44	47	27	70	9	39
Ed.17	45	126	0	171	0	0	58	74	39	116	19	36
Ed.18	18	38	0	0	56	0	22	26	8	47	2	7
Ed.19mas	19	81	0	0	0	100	55	38	7	78	8	14
Es.2bajo	46	133	44	58	22	55	179	0	0	95	22	62
Es.3medio	59	126	47	74	26	38	0	185	0	151	11	23
Es.4alto	23	58	27	39	8	7	0	0	81	65	5	11
Or.BOG	89	222	70	116	47	78	95	151	65	311	0	0
Or.CUN	9	29	9	19	2	8	22	11	5	0	38	0
Or.OTR	30	66	39	36	7	14	62	23	11	0	0	96

```
B<-acm.burt(Y,Y); xtable(B,digits=rep(0,13))
```

6.2.3. Tabla de Burt o de contingencias múltiples

Se denomina *tabla de Burt* o *tabla de contingencias múltiples* a la matriz: $\mathbf{B} = \mathbf{Z}'\mathbf{Z}$, que es cuadrada y simétrica de orden $p \times p$. $\mathbf{B}$ es una yuxtaposición de tablas de contingencia cruzando todas las variables por parejas, $\mathbf{Z}'_q\mathbf{Z}_{q'}$ y en la diagonal tiene matrices diagonales con las frecuencias de las categorías de la respectiva variable q: $\mathbf{Z}'_q\mathbf{Z}_q$.

Se designa $\mathbf{D}_p$ a la matriz diagonal que tiene los mismos valores de la diagonal de $\mathbf{B}$, es decir, el vector marginal columna de $\mathbf{Z}$ (suma de las columnas).

En el ejemplo, la tabla de Burt (6.2) cruza las 12 categorías de las 4 variables de los admitidos a Ciencias. Algunas lecturas son: entre los admitidos hay 128 mujeres y 317 hombres; 46 mujeres tienen 16 años o menos y 81 de los admitidos son de estrato alto.

La matriz diagonal para género es $\mathbf{Z_1'Z_1} = \begin{pmatrix} 128 & 0 \\ 0 & 317 \end{pmatrix}$ y la TC de Género × Estrato es $\mathbf{Z_1'Z_3} = \begin{pmatrix} 46 & 59 & 23 \\ 133 & 126 & 58 \end{pmatrix}$.

6.3. El ACM como un ACS de la TDC

El ACS se puede generalizar a más de dos variables, realizando el análisis de correspondencias (AC) de la tabla disyuntiva completa TDC o el AC de la tabla de Burt. En el segundo caso se pierde la información de los individuos, razón por la que se prefiere ver el ACM como el AC de la TDC $\mathbf{Z}$, el cual se presenta en esta sección.

La tabla de frecuencias relativas, asociada a la tabla $\mathbf{Z}$ es $\mathbf{F} = \frac{1}{ns}\mathbf{Z}$ con marginales fila $f_{i\cdot} = \frac{1}{n}; \forall i$ y marginales columna $f_{\cdot j} = \frac{n_j}{ns}; \forall j$, donde n_j es el número de individuos que asumen la categoría j.

En el ejemplo: $f_{i\cdot} = \frac{1}{445} = 0.22\,\%; \forall i = 1, \cdots, 445$ y

$f_{\cdot j} = \frac{n_j}{445 * 4} = \frac{n_j}{1780}; \forall j = 1, \cdots, 12.$

Por ejemplo $f_{\cdot 1} = \frac{128}{1780} = 7.19\,\%$.

6.3.1. Nube de individuos

Los n individuos conforman la nube N_n en $\mathbb{R}^p$, cuyas propiedades se muestran a continuación.

6.3.1.1. Coordenadas, pesos

Los perfiles de los individuos son las filas de la tabla $\frac{1}{s}\mathbf{Z}$, es decir, son barras de altura $1/s$ cuando el individuo asume la categoría j y 0 cuando no la asume.

El peso, igual para todos los individuos, es $\frac{1}{n}$ y la métrica es $\mathbf{M} = ns\mathbf{D}_p^{-1}$.

En el ejemplo: un perfil fila es $\frac{1}{4}z_{ij}; j = 1, \cdots, 12$, con peso $0.22\,\%$.

El perfil del primer individuo (25) de la tabla 6.1 es:

$$\{0.25\ 0\ 0\ 0.25\ 0\ 0\ 0\ 0.25\ 0\ 0\ 0\ 0.25\}.$$

La métrica en este espacio tiene término general: $m_j = \dfrac{1780}{n_j}$.

6.3.1.2. Centro de gravedad

La coordenada j del centro de gravedad, de las categorías g_p, es:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{s}z_{ij} = \frac{n_j}{ns}$$

Entonces el centro de gravedad es la marginal columna de $\mathbf{F} = \dfrac{1}{ns}\mathbf{Z}$.

Código para obtener el centro de gravedad, expresado en porcentaje

```
g <- colSums(Z)/nrow(Z)/4
xtable(data.frame(t(g)*100),digits=rep(1,13))
```

F	M	E16-	E17	E18	E19+	Ebaj	Emed	Ealt	Bogo	Cund	Otro
7.2	17.8	6.6	9.6	3.1	5.6	10.1	10.4	4.6	17.5	2.1	5.4

6.3.1.3. Distancia entre individuos

La distancia al cuadrado entre dos individuos es:

$$d^2(i,l) = ns\sum_{j=1}^{p}\frac{1}{n_j}\left(\frac{1}{s}\left[z_{ij} - z_{lj}\right]\right)^2 = \frac{n}{s}\sum_{j=1}^{p}\frac{1}{n_j}\left(z_{ij} - z_{lj}\right)^2 \qquad (6.1)$$

Dos individuos se parecen cuando asumen más o menos las mismas categorías. La distancia se amplifica más cuando uno solo de los dos individuos asume una categoría de baja frecuencia.

Como ejemplo, se calcula la distancia al cuadrado entre los individuos 50 y 100 de la tabla 6.1:

$$d^2(i50, i100) = \frac{445}{4}\left(\frac{1}{179} + \frac{1}{185}\right) = 1.22$$

La única diferencia entre ellos es que el primero es de estrato bajo y el segundo es de estrato medio.

La distancia (6.1) se puede expresar como una distancia euclidiana canónica, introduciendo la métrica en las coordenadas:

$$d^2(i,l) = \frac{n}{s}\sum_{j=1}^{p}\frac{1}{n_j}\left(z_{ij}-z_{lj}\right)^2 = \sum_{j=1}^{p}\left(\frac{\sqrt{n}z_{ij}}{\sqrt{sn_j}}-\frac{\sqrt{n}z_{lj}}{\sqrt{sn_j}}\right)^2 \tag{6.2}$$

Nótese, en la tabla 6.3, que los estudiantes 25 y 125 tienen distancia cero, es decir que, asumen las mismas categorías para las 4 variables, lo que se puede verificar en la tabla 6.1. Lo mismo sucede para las parejas 75 y 275, 250 y 325, 350 y 375.

Código para calcular las distancias entre individuos usando la función `dist` y obtener la tabla 6.3

```
n<-nrow(Z); Dp<-diag(colSums(Z)); s<-ncol(Y);
X<-sqrt(n/s)*as.matrix(Z)%*%solve(sqrt(Dp));
selin<-seq(25,445,25);
Dis<-dist(X[selin,]); round(as.dist(Dis),1)
```

Tabla 6.3. Distancia asociada al ACM entre los admitidos que están en la tabla 6.1

```
     25  50  75 100 125 150 175 200 225 250 275 300 325 350 375 400
50   2.6
75   2.0 1.6
100  2.3 1.1 2.0
125  0.0 2.6 2.0 2.3
150  2.1 2.0 1.7 2.3 2.1
175  2.5 2.3 1.9 2.3 2.5 2.3
200  1.1 2.3 1.7 2.6 1.1 1.8 2.5
225  2.2 2.5 2.3 2.5 2.2 2.2 1.9 2.2
250  2.2 2.2 1.4 2.1 2.2 2.2 1.3 2.2 1.8
275  2.0 1.6 0.0 2.0 2.0 1.7 1.9 1.7 2.3 1.4
300  2.1 2.1 1.8 2.4 2.1 2.2 1.9 1.7 2.0 2.3 1.8
325  2.2 2.2 1.4 2.1 2.2 2.2 1.3 2.2 1.8 0.0 1.4 2.3
350  2.1 2.1 1.7 1.8 2.1 2.1 1.4 2.4 2.4 1.9 1.7 1.7 1.9
375  2.1 2.1 1.7 1.8 2.1 2.1 1.4 2.4 2.4 1.9 1.7 1.7 1.9 0.0
400  2.3 1.1 2.0 1.6 2.3 1.7 2.5 2.0 2.8 2.4 2.0 2.4 2.4 2.4 2.4
425  2.5 2.2 1.9 2.2 2.5 1.8 1.4 2.5 1.2 1.3 1.9 2.4 1.3 2.0 2.0 2.5
```

6.3.1.4. Inercia de la nube de perfiles fila

La inercia de la nube de puntos es:

$$\frac{1}{n}\sum_{i=1}^{n} d^2(i, \mathbf{g}_p) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} \frac{ns}{n_j}\left(\frac{z_{ij}}{s} - \frac{n_j}{ns}\right)^2 = \frac{1}{s}(p - 2s + s) = \frac{p}{s} - 1.$$

La inercia de la nube de puntos depende del cociente entre el número de categorías y el número de variables, no de los valores internos de la tabla. Por lo tanto, no tiene significado estadístico. En el ejemplo es $12/4 - 1 = 2$.

6.3.1.5. Ejes y subespacios vectoriales

Las proyecciones de la nube de individuos se consiguen mediante el

$$ACP\left(\frac{1}{s}\mathbf{Z}, ns\mathbf{D}_p^{-1}, \frac{1}{n}\mathbf{I}_n\right).$$

La matriz de inercia es (utilizando las fórmulas del ACM($\mathbf{X}$, $\mathbf{N}$, $\mathbf{N}$) presentadas en la tabla 4.1 de la página 96):

$$\frac{1}{s}\mathbf{Z}'\frac{1}{n}\mathbf{I}_n\frac{1}{s}\mathbf{Z}ns\mathbf{D}_p^{-1} = \frac{1}{s}\mathbf{Z}'\mathbf{Z}\mathbf{D}_p^{-1} = \frac{1}{s}\mathbf{B}\mathbf{D}_p^{-1}.$$

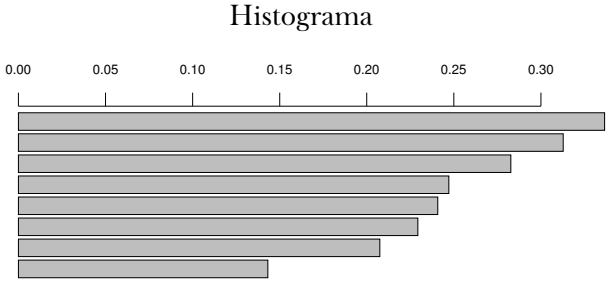
El rango de esta matriz es igual al rango de $\mathbf{Z}$, que es $p - s$, porque por cada variable hay una columna que es linealmente dependiente. Es decir, que una columna se puede obtener como la diferencia entre el vector de n unos y la suma de las demás columnas asociadas a la variable. Entonces, la nube de puntos está soportada en un subespacio de dimensión $p - s$, que es el número de valores propios mayores que cero. En el ejemplo hay $12 - 4 = 8$ valores propios mayores que cero, que se muestran en la figura 6.1.

Código para obtener la figura 6.1

```
acm<-dudi.acm(Y,scannf=FALSE,nf=3)
barplot(acm$eig,cex.axis=0.6)
dev.print(device = xfig,file="ACMadmiValP.fig")
eiglst<-data.frame(vp=acm$eig,porce=acm$eig*100/sum(acm$eig),
                   acupor=cumsum(acm$eig)*100/sum(acm$eig))
xtable(eiglst,digits=c(1,3,1,1)) #tabla en formato LaTeX
```

La decisión del número de ejes a retener se basa sobre todo en la forma del histograma puesto que el ACM tiene ejes "parásitos", es decir, que aparecen y no contienen información.

El porcentaje de inercia no es un criterio apropiado en el caso del ACM. Del histograma de la figura 6.1 se puede concluir que tres ejes son suficientes, a pesar de que retienen menos del 50 % de la inercia (46.6 %).



Valores propios

	Valor P	%	% acum.
1	0.337	16.8	16.8
2	0.313	15.6	32.5
3	0.283	14.1	46.6
4	0.247	12.4	59.0
5	0.241	12.0	71.0
6	0.229	11.5	82.5
7	0.208	10.4	92.8
8	0.143	7.2	100.0

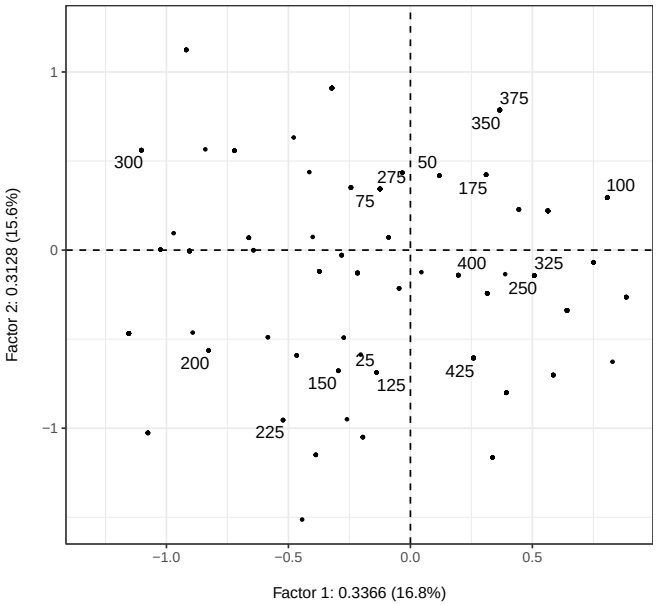
Figura 6.1. Histograma de valores propios del ACM de admitidos. Nótese que, los tres primeros ejes se destacan sobre los demás y que retienen el 46.6 % de la inercia

En la figura 6.2, se identifican los admitidos que están en la tabla 6.1. En el plano factorial no hay 445 puntos porque los admitidos que asuman las mismas categorías en las cuatro variables quedan superpuestos: en la tabla hay cuatro pares de estudiantes superpuestos: 25 y 125; 75 y 275; 250 y 325; 350 y 375.

El plano tiene más densidad de admitidos a la derecha y muestra mayor dispersión a la izquierda en las coordenadas sobre el segundo eje. En este ejemplo los individuos son anónimos, pero se comparan a través de sus categorías, cuyos planos factoriales se presentan en las figuras 6.3 y 6.7.

Código para obtener el primer plano factorial, figura 6.2

```
# con gg=TRUE se usa ggplot2 y ggrepel para la grafica
plot(acm,Tcol=FALSE,roweti=as.character(selin),cframe=1,cex.row=0.6,
     cex.global=0.8,gg=TRUE)
xtable(acm$li[selin,],digits=rep(2,4)) # coordenadas
```



Coordenadas sobre los tres primeros ejes

Admitido	F1	F2	F3
25	-0.14	-0.69	0.08
50	0.12	0.42	0.57
75	-0.12	0.34	-0.38
100	0.81	0.29	0.63
125	-0.14	-0.69	0.08
150	-0.30	-0.68	0.51
175	0.31	0.42	-0.32
200	-0.83	-0.56	0.02
225	-0.52	-0.95	-0.38
250	0.51	-0.14	-0.99
275	-0.12	0.34	-0.38
300	-1.10	0.56	0.35
325	0.51	-0.14	-0.99
350	0.37	0.79	0.35
375	0.37	0.79	0.35
400	0.20	-0.14	0.91
425	0.26	-0.61	-0.43

Figura 6.2. Admitidos sobre el primer plano factorial del ACM. Los admitidos etiquetados son los que están en la tabla 6.1

Para obtener las ayudas para la interpretación, tabla 6.4

```
ineracm<-inertia.dudi(acm,T,T); #ayudas
xtable(data.frame(contr=ineracm$row.abs[selin,],
      cos2=abs(ineracm$row.rel[selin,]),
      Ciner=ineracm$row.contrib[selin]),digits=rep(2,8));
```

Tabla 6.4. Extracto de las ayudas para la interpretación del ACM de Admitidos

Admitido	Contribuciones %			Coseno^2 %			Cinercia
	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3	%
25	0.01	0.34	0.00	0.84	20.70	0.27	0.26
50	0.01	0.13	0.26	0.61	7.54	14.00	0.26
75	0.01	0.08	0.11	1.58	12.02	14.52	0.11
100	0.43	0.06	0.32	28.34	3.78	17.38	0.26
125	0.01	0.34	0.00	0.84	20.70	0.27	0.26
150	0.06	0.33	0.21	4.87	25.64	14.70	0.20
175	0.06	0.13	0.08	4.39	8.16	4.69	0.25
200	0.46	0.23	0.00	29.74	13.81	0.01	0.26
225	0.18	0.65	0.11	9.64	32.23	5.08	0.32
250	0.17	0.01	0.77	14.87	1.18	56.27	0.19
275	0.01	0.08	0.11	1.58	12.02	14.52	0.11
300	0.81	0.23	0.09	54.18	14.03	5.30	0.25
325	0.17	0.01	0.77	14.87	1.18	56.27	0.19
350	0.09	0.44	0.10	9.44	43.48	8.69	0.16
375	0.09	0.44	0.10	9.44	43.48	8.69	0.16
400	0.03	0.01	0.65	1.36	0.70	29.04	0.32
425	0.04	0.26	0.15	3.30	18.12	9.33	0.23

6.3.2. Nube de categorías

En el ACM se pone más atención a las categorías, porque los individuos son anónimos en la mayoría de las aplicaciones. La estructura de la tabla $\mathbf{Z}$ por paquetes de categorías, según las variables y su código disyuntivo completo, otorga al espacio de las categorías propiedades interesantes.

6.3.2.1. Coordenadas

Cada perfil columna j tiene solo dos alturas: cero o $1/n_j$, pero las alturas son, en general, diferentes en cada perfil.

La tabla de perfiles categoría es $\mathbf{ZD_p^{-1}}$, ya que al postmultiplicar por una matriz diagonal, cada columna queda multiplicada por el respectivo valor de la diagonal, en este caso $1/n_j$. El peso de cada categoría es n_j/ns.

6.3.2.2. Centro de gravedad

El centro de gravedad de la nube de categorías es el vector de n valores $\frac{1}{n}$, lo cual se verifica a continuación, para cualquier coordenada $\mathbf{g}(i)$:

$$\mathbf{g}(i) = \sum_{j=1}^{p} \frac{n_j}{ns} \frac{z_{ij}}{n_j} = \sum_{j=1}^{p} \frac{1}{ns} z_{ij} = \frac{1}{ns} s = \frac{1}{n}$$

6.3.2.3. Distancia entre dos categorías

La métrica en el espacio de las categorías es $n\mathbf{I}_n$, donde $\mathbf{I}_n$ es la matriz identidad de dimensión n.

Entonces la distancia al cuadrado entre dos categorías j y k es:

$$d^2(j,k) = \sum_{i=1}^{n} n \left(\frac{z_{ij}}{n_j} - \frac{z_{ik}}{n_k} \right)^2 \tag{6.3}$$

Para calcular la distancia entre categorías utilizando la función `dist{stats}`, hay que introducir n en el paréntesis para tener las coordenadas de una distancia euclidiana canónica.

En la tabla 6.5 se muestran las distancias entre las categorías del ACM del ejemplo "Admitidos".

La interpretación de la distancia entre categorías como está dada en (6.3) no es directa: cuando un individuo asume las dos categorías, el valor del paréntesis no se anula porque los valores son, en general, diferentes. Sin embargo, es fácil derivar una expresión interpretable porque el valor entre paréntesis tiene cuatro posibilidades que se pueden contar:

		Categoría k		Suma
		1	0	
Categoría j	1	a	b	$a+b=n_j$
	0	c	d	$c+d$
Suma		$a+c=n_k$	$b+d$	n

Es decir, a es el número de individuos que asumen simultáneamente las categorías j y k, d el número de los que no asumen ninguna de las dos, b el número de los que asumen j pero no k, y c los que asumen k pero no j.

Código para calcular las distancias entre categorías (tabla 6.5)

```
Xcat<-sqrt(n)*solve(Dp) %*% t(Z) #coord eucli canonicas
rownames(Xcat)<-substr(colnames(Z),6,nchar(colnames(Z)))
Discat<-dist(Xcat) # distancias
as.dist(round(Discat,1))
```

Tabla 6.5. Distancia entre las categorías activas asociadas al ACM del ejemplo "Admitidos"

	F	M	a16m	a17	a18	a19M	bajo	medio	alto	Bogo	Cund
M	2.2										
a16m	2.1	1.9									
a17	2.1	1.4	2.5								
a18	3.0	2.7	3.4	3.2							
a19M	2.6	1.9	2.9	2.7	3.5						
bajo	2.0	1.3	2.1	1.8	2.9	2.0					
medio	1.9	1.4	2.1	1.7	2.8	2.2	2.2				
alto	2.6	2.2	2.6	2.4	3.4	3.0	2.8	2.8			
Bogo	1.7	0.9	1.9	1.4	2.6	1.9	1.5	1.2	2.2		
Cund	3.7	3.3	3.7	3.4	4.3	3.8	3.4	3.6	4.0	3.6	
otro	2.4	2.0	2.3	2.3	3.4	2.8	2.0	2.4	3.0	2.5	4.0

El desarrollo del cuadrado en (6.3) da:

$$d^2(j,k) = \sum_{i=1}^{n} n\left(\frac{z_{ij}^2}{n_j^2} - 2\frac{z_{ij}}{n_j}\frac{z_{ik}}{n_k} + \frac{z_{ik}^2}{n_k^2}\right) = n\left(\frac{1}{n_j} + \frac{1}{n_k} - 2\sum_{i=1}^{n}\frac{z_{ij}}{n_j}\frac{z_{ik}}{n_k}\right) \quad (6.4)$$

El último término del paréntesis en (6.4) solo suma cuando el individuo i asume las dos categorías, es decir, esa sumatoria da a individuos. Entonces:

$$d^2(j,k) = n\left(\frac{n_k + n_j - 2a}{n_j n_k}\right) = n\left(\frac{a + c + a + b - 2a}{n_j n_k}\right) = \frac{n}{n_j n_k}(b + c) \quad (6.5)$$

La fórmula (6.5) muestra que en la distancia de dos categorías solo suman los individuos que asumen una y solo una de las dos. Además las categorías de baja frecuencia se alejan más de las otras.

Se muestran a continuación dos ejemplos de cálculo de distancias entre categorías usando "la calculadora R":

1. Entre femenino y masculino: son las únicas categorías de una misma variable. Entonces no hay coincidencias, es decir, $b + c = n = 445$:

```
c(n,B[1,1],B[2,2]); ## [1] 445 128 317
sqrt(445/(128*317)*445); # distancia entre cat. F y M
## [1] 2.209151
```

2. Entre femenino y 16 años o menos:

```
table(Z[,1],Z[,3]);
##       0   1
##   0 245  72
##   1  82  46
c(B[1,1],B[3,3]); ## [1] 128 118
sqrt(445/(128*118)*(72+82)); ## [1] 2.130072
```

La distancia entre la categoría j y el centro de gravedad $\mathbf{g}_n = \frac{1}{n}\mathbf{1}_n$ (todas las n coordenadas valen $1/n$), es:

$$d^2(j, \mathbf{g}_n) = n \sum_{i=1}^{n} \left(\frac{z_{ij}}{n_j} - \frac{1}{n} \right)^2 = n \sum_{i=1}^{n} \left(\frac{z_{ij}^2}{n_j^2} - 2\frac{z_{ij}}{n_j}\frac{1}{n} + \frac{1}{n^2} \right) = \frac{n}{n_j} - 1 \quad (6.6)$$

Se observa en (6.6) que las categorías de menores frecuencias son las más alejadas del origen.

6.3.2.4. Inercia de la nube de categorías

Es interesante obtener la inercia calculando la contribución de una categoría a esta inercia, luego sumando la inercia de las categorías de una variable y finalmente las de las s variables. Procediendo de ese modo, la fórmula para obtener la inercia de la nube de categorías N_p es:

$$I(N_p) = \sum_{q=1}^{s} \sum_{j \in J_q} \frac{n_j}{ns} d^2(j, \mathbf{g}_n) = \sum_{q=1}^{s} \sum_{j \in J_q} \frac{n_j}{ns} \left(\frac{n}{n_j} - 1 \right) = \sum_{q=1}^{s} \sum_{j \in J_q} \frac{1}{s} \left(1 - \frac{n_j}{n} \right) \quad (6.7)$$

donde J_q es el conjunto de categorías que pertenecen la variable q. A partir de (6.7) se observan o derivan las contribuciones a la inercia de una categoría, una variable y la inercia total:

De una categoría j: $\frac{1}{s}\left(1 - \frac{n_j}{n}\right)$, lo que indica que contribuyen más a la inercia las categorías de baja frecuencia.

De una variable q: $\sum_{j \in J_q} \frac{1}{s}\left(1 - \frac{n_j}{n}\right) = \frac{1}{s}\left(p_q - \frac{n}{n}\right) = \frac{1}{s}\left(p_q - 1\right)$, donde p_q es el número de categorías de la variable q. Se observa que las variables con más número de categorías contribuyen más a la inercia.

Inercia total: $\sum_{q=1}^{s} \frac{1}{s}\left(p_q - 1\right) = \frac{1}{s}(p - s) = \frac{p}{s} - 1$. Igual a la inercia de la nube de individuos $I(N_n)$, que no tiene significado estadístico, porque no depende de los valores de la tabla, sino de la relación entre número de categorías y número de variables.

6.3.2.5. Ejes factoriales

En el espacio de las categorías los valores propios mayores que cero son iguales a los del espacio de los individuos. Los vectores propios y las coordenadas de los ejes se obtienen mediante las relaciones de transición, que se abordan en la sección 6.3.3.

El primer plano factorial de las categorías, obtenido en el ACM del ejemplo "Admitidos", se presenta en la figura 6.3 y las ayudas para la interpretación en la tabla 6.6.

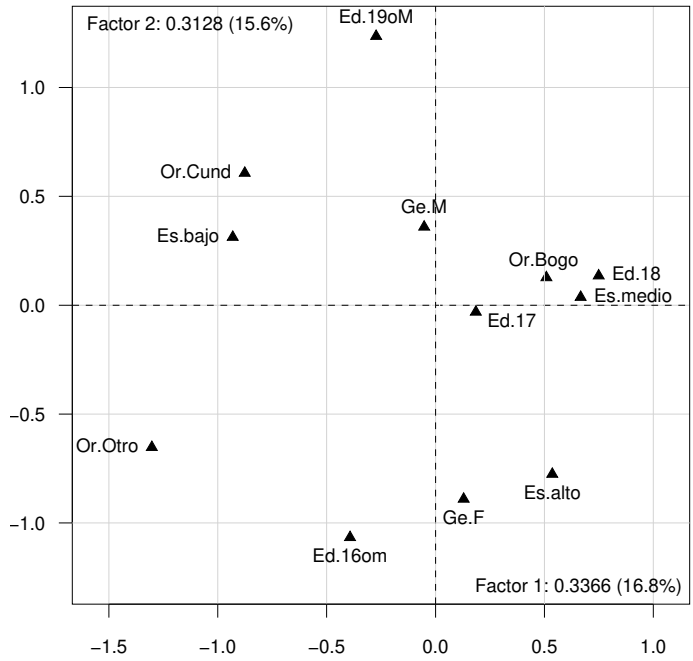
En el primer eje, las categorías con contribuciones absolutas superiores al promedio (100/12 = 8.333 %) y sus oposiciones son: estrato bajo (-) contra estrato alto (+) y origen de otro departamento (-) contra origen Bogotá.

En el segundo, las mujeres se sitúan abajo (-), los admitidos de 16 años o menos (-) se oponen a los de 19 años o más (+), y los de estrato alto se ubican abajo (-).

En el primer plano factorial (figura 6.3) se ven los grupos de categorías. Abajo: 16 años o menos, mujeres y estrato alto; a la derecha: estrato medio y origen Bogotá; arriba: 19 años o más; y arriba a la izquierda origen Cunidinamarca y estrato bajo.

En el tercer eje, se oponen los de 17 años (-) a los de 18 (+), los de estrato alto están al lado negativo, al igual que los que vienen de Cundinamarca. Este eje vale la pena leerlo por la edad y por el departamento de origen.

El primer plano factorial es una buena síntesis del análisis, pero conviene leer también los planos 1-3, que representan mejor las categorías de estrato y origen, y 2-3, para ver mejor las categorías de edad y estrato alto.




```
plot(acm,Trow=FALSE,cex.global=0.8,cframe=1,gg=TRUE,col.col="black")
```

Pesos y coordenadas de las categorías

Cate	Peso	Coordenadas		
	%	G1	G2	G3
gene.F	7.19	0.13	-0.89	0.51
gene.M	17.81	-0.05	0.36	-0.21
edad.a16m	6.63	-0.39	-1.07	0.29
edad.a17	9.61	0.19	-0.03	-0.88
edad.a18	3.15	0.75	0.14	1.13
edad.a19M	5.62	-0.27	1.24	0.53
estr.bajo	10.06	-0.93	0.31	0.18
estr.medio	10.39	0.67	0.04	0.31
estr.alto	4.55	0.54	-0.78	-1.12
orig.Bogo	17.47	0.51	0.13	0.11
orig.Cund	2.13	-0.88	0.61	-1.44
orig.Otro	5.39	-1.30	-0.65	0.23

```
xtable(data.frame(Peso=acm$cw*100,coor=acm$co),digits=rep(2,5))
```

Figura 6.3. Primer plano factorial del ACM de admitidos, mostrando las categorías

Tabla 6.6. Ayudas para la interpretación de las categorías activas

Categoría	Contribuciones %			Cosenos2 %			Cinercia
	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3	%
gene.F	0.36	18.21	6.66	0.67	31.97	10.58	8.90
gene.M	0.14	7.35	2.69	0.67	31.97	10.58	3.60
edad.a16m	3.04	24.09	1.99	5.56	41.02	3.07	9.19
edad.a17	0.98	0.03	26.54	2.14	0.06	48.74	7.70
edad.a18	5.24	0.18	14.21	8.08	0.26	18.38	10.93
edad.a19M	1.24	27.43	5.67	2.16	44.26	8.27	9.69
estr.bajo	25.90	3.15	1.17	58.34	6.58	2.21	7.47
estr.medio	13.71	0.05	3.61	31.59	0.10	6.99	7.30
estr.alto	3.88	8.75	20.06	6.39	13.38	27.73	10.22
orig.Bogo	13.45	0.91	0.70	60.15	3.76	2.65	3.76
orig.Cund	4.87	2.51	15.72	7.16	3.44	19.43	11.43
orig.Otro	27.18	7.34	0.97	46.66	11.71	1.39	9.80

6.3.3. El ACM como un solo ACP

Es conveniente ver el ACM como un solo ACP($\mathbf{X}$, $\mathbf{M}$, $\mathbf{N}$), ya que las relaciones entre los espacios de individuos y categorías se observan más fácilmente y porque esta visión se utiliza en la derivación de otros métodos, como el análisis factorial múltiple para variables cualitativas propuesto por Escofier & Pagès (1992), que no se abordan en este texto.

Haciendo específica la sección 5.3 al caso del ACM, las matrices $\mathbf{X}$, $\mathbf{M}$, $\mathbf{N}$ son:

- $\mathbf{X} = n\mathbf{I}_n \dfrac{1}{ns}\mathbf{Z}ns\mathbf{D} = n\mathbf{Z}\mathbf{D}_p^{-1}$, término general: $x_{ij} = \dfrac{n}{n_j} z_{ij}$.
- $\mathbf{M} = \dfrac{1}{ns}\mathbf{D}_p$, con $m_j = \dfrac{n_j}{ns}$. $\quad \mathbf{N} = \dfrac{1}{n}\mathbf{I}_n$, con $n_i = \dfrac{1}{n}$.

La $\mathbf{M}$-distancia al cuadrado entre dos individuos i y l y la $\mathbf{N}$-distancia al cuadrado entre dos categorías j y k de $\mathbf{X}$ son:

$$d^2(i,l) = \sum_{j=1}^{p} \frac{n_j}{ns}\left(\frac{n}{n_j} z_{ij} - \frac{n}{n_j} z_{lj}\right)^2 = \frac{n}{s}\sum_{j=1}^{p} \frac{1}{n_j}(z_{ij} - z_{lj})^2 \tag{6.8}$$

$$d^2(j,k) = \sum_{i=1}^{n} \frac{1}{n}\left(\frac{n}{n_j} z_{ij} - \frac{n}{n_k} z_{ik}\right)^2 = n\sum_{i=1}^{n} \left(\frac{z_{ij}}{n_j} - \frac{z_{ik}}{n_k}\right)^2 \tag{6.9}$$

Que corresponden a las fórmulas (6.1) y (6.3), respectivamente.

6.3.4. Relaciones cuasibaricéntricas

En el ACM la fórmula general que permite escribir la coordenada de un individuo i en función de las coordenadas de las categorías de las variables es:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{s} \sum_{j=1}^{p} z_{ij} G_s(j) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{s} \sum_{j \in J_i} G_s(j) \qquad (6.10)$$

donde J_i es el conjunto de categorías asumidas por el individuo i. La fórmula muestra que la coordenada sobre un eje s del individuo i se sitúa en el promedio aritmético de las coordenadas de las categorías que asume, dilatadas por el inverso de la raíz cuadrada del valor propio.

El primer individuo de la tabla 6.1 asume las categorías Ge.F (0.1 es la coordenada sobre el primer eje, que se lee en la tabla incluida en la figura 6.3), Ed.17 (0.2), Es.medio (0.7) y Or.Otro (-1.3).

Entonces el promedio aritmético es (0.1+0.2+0.7-1.3)/4 = -0.08, la dilatación es $1/\sqrt{0.337} = 1.72$ y la coordenada es 1.72*(-0.08) = -0.14, que se puede leer con el comando `acm$li[25,]` y observar en la figura 6.2.

La fórmula es análoga para la coordenada de una categoría j en función de las coordenadas de los individuos:

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{n_j} \sum_{i=1}^{n} z_{ij} F_s(i) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{n_j} \sum_{i \in I_j} F_s(i) \qquad (6.11)$$

donde I_j es el conjunto de individuos que asumen la categoría j. Entonces, esta categoría j se sitúa en el promedio aritmético de las coordenadas de los individuos que la asumen, dilatada por el inverso de la raíz cuadrada del valor propio.

Las fórmulas de transición, denominadas también *cuasibaricéntricas*, permiten la representación simultánea y su interpretación. Para el ejemplo "Admitidos a la Facultad de Ciencias", el primer plano de esa representación se muestra en la figura 6.4.

De la misma manera que en el ACS (capítulo 5), las categorías que pertenecen a cada variable están centradas en el origen de la representación. Como consecuencia, las que están más próximas al origen son las que más frecuencia tienen.

Por ejemplo, la variable género, se puede ver como una palanca que está en equilibrio, con apoyo en (0,0). Este equilibrio que implica que hay

bastantes más hombres que mujeres, dentro de los admitidos a las carreras de Ciencias.

Por otro lado, las categorías son cuasibaricentros de las coordenadas de los admitidos que las asumen. Entonces, por ejemplo, a la derecha del plano hay admitidos que asumen simultáneamente las categorías origen Bogotá, 18 años de edad y estrato medio.

El admitido 100 puede ser uno ellos, pero para estar seguros hay que verificarlo en los datos (tabla 6.1, 100: M, 18, medio, Bogo). En cambio el individuo 225 puede ser de 16 años o menos y de otro origen (225: M, 16om, alto, otro).

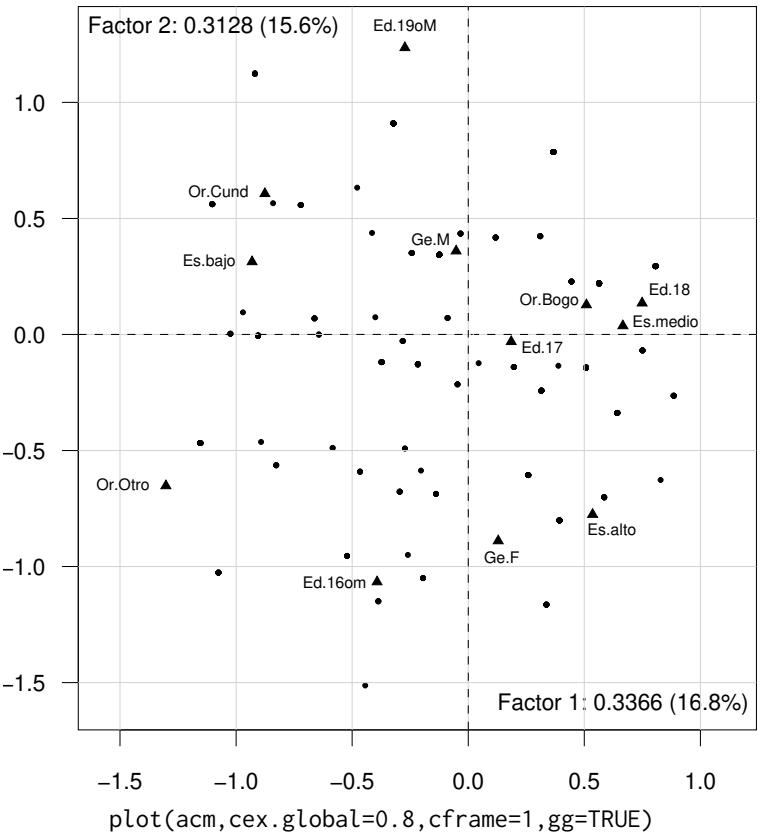


Figura 6.4. Primer plano factorial del ACM de admitidos, mostrando individuos y categorías. Se puede ver, por ejemplo, la asociación entre estrato medio y dieciocho años, lo que se puede verificar en los perfiles de la figura 2.4, donde se muestra la asociación bivariada entre edad y estrato de los admitidos.

Para ilustrar el cuasibaricentro de una categoría, en la figura 6.5 se muestran los grupos de estudiantes según su origen y las categorías del ACM. Cada rayo une el (0,0) con la posición de la coordenada. Un rayo pasa por el centro de gravedad de los individuos que asumen la categoría —etiqueta en cursiva y punto con una cruz—. La posición de la categoría se aleja del centroide, porque las coordenadas de la proyección de aquella están multiplicadas por el inverso de la raíz cuadrada de los valores propios asociados a los ejes 1 y 2, respectivamente. También se muestra que el origen es el centro de gravedad de las tres categorías de la variable origen del admitido. Las cercanías al centro indican la frecuencia: hay más bogotanos, luego cundinamarqueses y menos de otros departamentos.

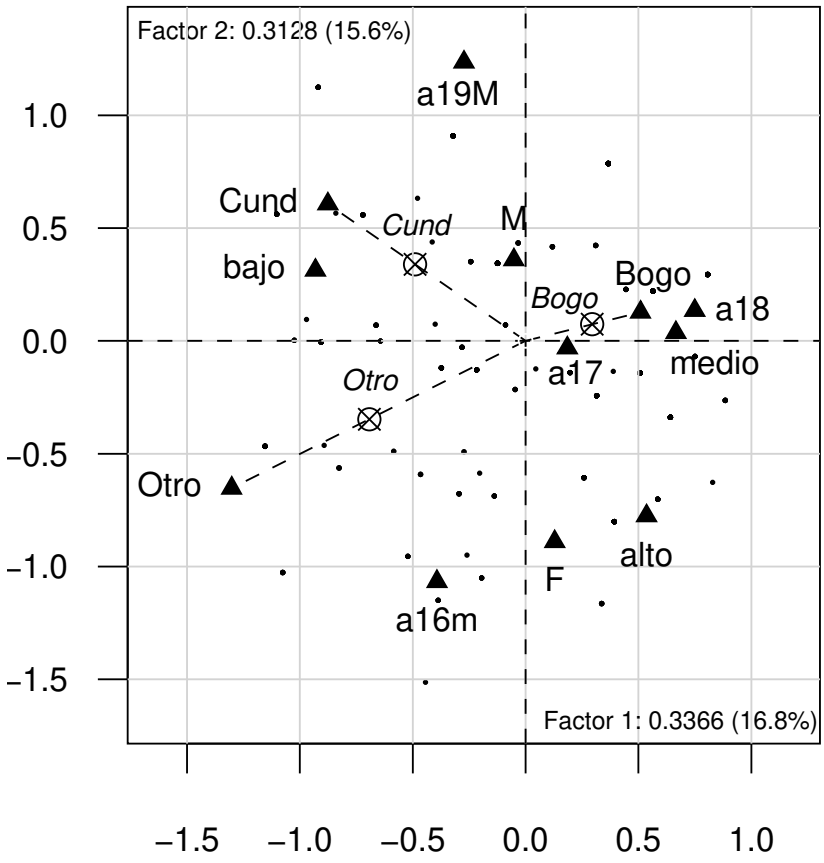


Figura 6.5. Primer plano factorial del ACM de admitidos mostrando los individuos según su origen

6.3.5. Ayudas para la interpretación

Las ayudas para la interpretación de los individuos y de las categorías tienen las mismas expresiones que las del ACP y ACS (sección 5.3.3). Se adiciona la contribución absoluta de las variables, obtenida como la suma de las contribuciones de sus categorías.

6.3.5.1. Razón de correlación

Para una variable cualitativa se puede calcular la razón de correlación con respecto a un eje s, que es una variable continua.

Una variable cualitativa induce una partición de los n individuos y se puede descomponer la inercia (varianza) de los individuos sobre el eje en $varianza\ inter + varianza\ intra$.

La razón de correlación se define como el cociente entre varianza inter y varianza total.

La varianza total de F_s es λ_s; la varianza inter con respecto a una variable q es:

$$\sum_{j \in J_q} \frac{n_j}{n} (\overline{F}_{sj})^2$$

donde $\overline{F}_{sj} = \sum_{i \in I_{j \in J_q}} \frac{1}{n_j} F_s(i)$, es decir, el promedio aritmético de las coordenadas sobre el eje s de los individuos que asumen la categoría j de la variable q.

El promedio de las n coordenadas sobre el eje s es 0, es decir, las coordenadas sobre s están centradas. J_q es el conjunto de categorías de la variable q.

Por las relaciones de transición $\overline{F}_{sj} = \sqrt{\lambda_s} G_s(j)$, entonces:

$$Varianza\ intra(q) = \lambda_s \sum_{i \in J_q} \frac{n_j}{n} G_s^2(j)$$

Por lo tanto, la razón de correlación es:

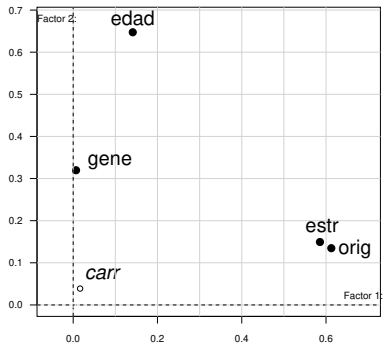
$$\eta_s^2(q) = \sum_{j \in J_q} \frac{n_j}{n} G_s^2(j) \tag{6.12}$$

que se puede expresar como función de la contribución absoluta de las categorías:

$$\eta_s^2(q) = \lambda_s s \sum_{j \in J_q} Ca_s(j) \tag{6.13}$$

$Ca_s(j)$ es la contribución absoluta de la categoría j sobre el eje s. Los valores de las razones de correlación se encuentran en el objeto de salida de la función `dudi.acm{ade4}` en la tabla `$cr`, con estos se pueden obtener los planos factoriales para las variables. La figura 6.6 muestra el primer plano factorial de las variables del ACM de "Admitidos".

Ejemplo de cálculo para la variable *Origen*: la suma de las contribuciones de las tres categorías sobre el primer eje es $13.45 + 4.87 + 27.18 = 45.5\,\%$ (valores tomados de la figura 6.3), el primer valor propio es 0.3366 y el número de variables 4. Entonces la razón de correlación es $\eta_1^2(orig) = 4 * 0.3366 * 0.455 = 0.613$, valor que se puede leer en el eje horizontal de la figura 6.6, de forma aproximada.



`plotfp(acm$cr,gg=TRUE)`

Figura 6.6. Relaciones de correlación de las variables sobre el primer plano factorial del ACM de admitidos

La figura 6.6 muestra que hay una relación alta de las variables estrato y origen con el eje 1 y edad y género con el eje 2. Buena parte de la inercia del primer eje es entre estratos y también, entre origen; y la del segundo eje es entre edades. También aparece proyectada la variable carrera, que es ilustrativa, y no muestra asociación con ninguno de los ejes. La cercanía de las variables estrato y origen indican que las contribuciones de las dos variables a la varianza de los dos ejes son similares (fórmula 6.13).

6.3.6. Elementos suplementarios

De la misma manera que en ACP y ACS, en el ACM se pueden proyectar individuos, variables cualitativas y variables continuas como elementos ilustrativos.

Individuos

Los individuos ilustrativos se pueden proyectar utilizando la fórmula cuasibaricéntrica (6.10). Dicho de otro modo, la coordenada de un individuo suplementario es el promedio de las coordenadas de las categorías que asume, dilatado por el inverso de la raíz cuadrada del valor propio.

Variables cualitativas

Las categorías de una variable ilustrativa se proyectan mediante la fórmula cuasibaricéntrica (6.11), como el promedio de las coordenadas de los individuos que la asumen, dilatado por el inverso de la raíz cuadrada del valor propio. Las categorías suplementarias no contribuyen a la inercia de los ejes, ya que no participan en su determinación, pero se pueden calcular sus cosenos cuadrados sobre los ejes.

Adicionalmente se suelen utilizar los denominados *valores test* para cada categoría, con el fin de indicar si la coordenada de su proyección se puede considerar diferente de cero (sección 2.2.2 y fórmula 2.2).

El valor test es la distancia al origen de la coordenada $G_s(j)$ en términos de desviaciones estándar:

$$t_s(j) = \frac{G_s(j)}{\sigma_j}$$

Para obtener la desviación estándar:

- Una categoría j es asumida por los n_j individuos.
- La varianza del promedio de n_j individuos tomados al azar de los n individuos es:

$$\frac{n - n_j}{n - 1} \frac{\lambda_s}{n_j}$$

- La coordenada $G_s(j)$ es el promedio aritmético de las coordenadas de esos individuos, multiplicada por el inverso de la raíz cuadrada del valor propio λ_s (fórmula 6.11). Entonces la varianza asociada a la

coordenada es:

$$\frac{1}{\lambda_s}\left(\frac{n-n_j}{n-1}\right)\frac{\lambda_s}{n_j}=\frac{n-n_j}{n_j(n-1)}$$

Por lo tanto, el valor test, para la coordenada de la categoría j en el análisis, $G_s(j)$, es:

$$t_s(j)=\sqrt{\frac{n_j(n-1)}{n-n_j}}G_s(j) \tag{6.14}$$

Como ejemplo se muestra el cálculo del valor test para Química sobre el primer eje:

Los admitidos a Química son 63 (ver con `summary(admi$car)`) y la coordenada de Química es -0.259 (tabla de la figura 6.7). Entonces:

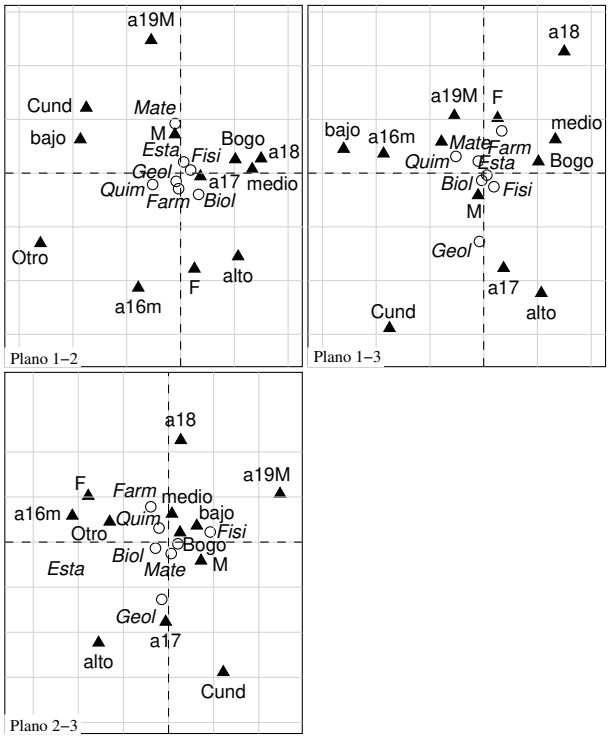
$$t_1(Qui)=\sqrt{\frac{63(445-1)}{445-63}}\ (-0.26)=-2.225$$

La diferencia con el valor de la figura 6.7 se debe al número de cifras significativas. Este valor, menor que -2, indica que es válido leer la posición negativa de Química sobre el primer eje, lo que significa que está asociada con estrato bajo y origen fuera de Bogotá, ya que son categorías activas que tienen mayores coordenadas negativas sobre el primer eje.

Los valores test de la figura 6.7 indican que solo es legítimo interpretar, como diferentes de cero, las coordenadas de: Química (-) sobre el primer eje, Matemáticas (+) sobre el segundo y Geología (-) opuesto a Farmacia (+) sobre el tercero.

Matemáticas tiene, en comparación al promedio, mayor proporción de admitidos de 19 o más años; Geología, de 17 años y estrato alto; Farmacia, de 18 años y género femenino. Para corroborar sobre los datos, se pueden ver los perfiles de las carreras según las cuatro variables en la figura 2.5.

A una variable cualitativa suplementaria también se le puede calcular la relación de correlación con la primera igualdad de la fórmula (6.12) e incluirla en la gráfica de las variables. En el ejemplo se proyecta la variable carrera, que se presenta en la figura 6.6. Su posición, cerca al origen, muestra poca relación con las cuatro variables sociodemográficas.



Número de admitidos, distancias al cuadrado y coordenadas

Carrera	Número admitidos	Distancia 2	Coordenadas		
			Eje1	Eje2	Eje3
Biol	63	6.06	-0.02	-0.15	-0.07
Esta	66	5.74	0.03	0.10	-0.02
Farm	73	5.10	0.17	-0.20	0.39
Fisi	82	4.43	0.09	0.03	-0.13
Geol	45	8.89	-0.04	-0.08	-0.64
Mate	53	7.40	-0.05	0.46	0.11
Quim	63	6.06	-0.26	-0.11	0.16

Valores test y cosenos cuadrados

Carrera	Valores test			Cosenos cuadrados		
	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3
Biol	-0	-1.25	-0.579	0.000	0.003	0.001
Esta	0	0.91	-0.160	0.000	0.002	0.000
Farm	2	-1.85	3.664	0.006	0.008	0.030
Fisi	1	0.29	-1.267	0.002	0.000	0.004
Geol	-0	-0.54	-4.490	0.000	0.001	0.045
Mate	-0	3.58	0.869	0.000	0.029	0.002
Quim	-2	-0.91	1.335	0.011	0.002	0.004

Figura 6.7. Plano factoriales: 1-2, 1-3 y 2-3, mostrando las carreras como categorías suplementarias y ayudas para la interpretación

Código para proyectar la carrera del admitido como variable cualitativa suplementaria (figura 6.7)

```
supcar<-supqual(acm,admi$carr)
par(mfrow=c(2,2))
plot(acm,Trow=FALSE,infaxes = "no",main="Plano␣1-2",
     ylim=c(-1.5,1.3),col.col="black")
points(supcar$coor,col="black")
text(supcar$coor,labels=attributes(admi[,1])$levels,
     col="black",pos=1,cex=0.8,font=3)
xtable(cbind(ncat=supcar$ncat,d2=supcar$dis2,supcar$coor,
       supcar$tv,supcar$cos2),digits = c(0,0,rep(3,10)))
plot(acm,1,3,Trow=FALSE,infaxes="no",main="Plano␣1-3",
     ylim=c(-1.5,1.3),col.col="black")
points(supcar$coor[,c(1,3)],col="black")
text(supcar$coor[,c(1,3)],labels=attributes(admi[,1])$levels,
     col="black",pos=1,cex=0.8,font=3)
plot(acm,2,3,Trow=FALSE,infaxes="no",main="Plano␣2-3",
     ylim=c(-1.5,1.3),col.col="black")
points(supcar$coor[,c(2,3)],col="black")
text(supcar$coor[,c(2,3)],labels=attributes(admi[,1])$levels,
     col="black",pos=1,cex=0.8,font=3)
#dev.print(device = xfig,file="ACMadmiCarreraSup3planos.fig")
```

Variables continuas.

Las variables continuas se pueden proyectar en los planos de individuos y categorías, utilizando como coordenadas los coeficientes de correlación entre la variable y el factor. Se interpretan como en un ACP.

Por ejemplo, en el ACM de los admitidos, las correlaciones entre los tres ejes factoriales y el resultado global del examen son: 26.9 %, 3.5 % y -29.3 %. El valor de 26.9 % de correlación indica que hay alguna relación positiva entre el primer eje y el resultado global del examen. Entonces las categorías que caracterizan el sentido positivo del eje: edad de 18 años, estratos medio y alto, y origen Bogotá, tienen, en promedio, notas más altas.

6.3.7. Retorno a los datos

En la aplicación del ACM y los otros métodos factoriales conviene regresar a los datos para observar lo que los mapas factoriales revelan, y a veces mostrarlos de otra forma.

Con el pequeño ejemplo de los admitidos, el lector puede corroborar la lectura de las gráficas del ACM.

La proyección de la variable carrera como ilustrativa (figura 6.7) muestra asociación de Química con origen Cundinamarca y otros departamentos, y estrato bajo (lado negativo del primer eje); lo que, también se observa en las asociaciones bivariadas de la figura 2.5.

La asociación de Matemáticas con admitidos de diecinueve años y mayores, y de género masculino, también se observa en la descripción bivariada de la misma figura (2.5).

6.4. AC derivados de la misma tabla

Con el análisis de correspondencias de la tabla de Burt o de contingencia múltiple se obtienen las mismas conclusiones que con el ACM (AC de la tabla disyuntiva completa), con respecto a las relaciones entre las categorías, pero no hay información directa sobre los individuos.

Estudiar los diferentes análisis de correspondencias aplicables a dos variables cualitativas, es un ejercicio académico que permite, aclarar el por qué de algunas decisiones que se toman en el ACM.

6.4.1. AC de la tabla de Burt

El AC aplicado a la matriz $\mathbf{B}$ está relacionado con el ACM, que es el AC de $\mathbf{Z}$ (Lebart *et al.*, 1995, p. 126):

- Los valores propios del AC de $\mathbf{B}$, $\lambda_{\mathbf{B}}$, son el cuadrado de los del ACM: $\lambda_{\mathbf{B}} = \lambda_Z^2$. Este resultado implica que los porcentajes de inercia de los primeros ejes son mayores en el AC de la tabla de Burt que en el ACM, y por lo tanto, más "optimistas".

- Las coordenadas factoriales del AC de $\mathbf{B}$, $G_{\mathbf{B}}$, son homotecias de las de las categorías del ACM: $G_{\mathbf{B}} = \sqrt{\lambda_Z} G$; es decir, que todas las categorias están más cerca del origen con respecto a las del ACM; entonces, los planos del AC de $\mathbf{B}$ son imágenes reducidas de los del ACM.

6.4.2. ACS y ACM de dos variables

La asociación entre dos variables cualitativas se describe con el ACS de la TC obtenida de ellas. Es posible realizar el ACM de la tabla de n individuos

descritos por las dos variables. Lebart *et al.* (2006, p. 203) analizan ese caso y aquí se presentan algunos de los resultados.

El ACM, que es el AC de $\mathbf{Z} = [\mathbf{Z_1}\ \mathbf{Z_2}]$, se compara con el ACS de la tabla $\mathbf{K} = \mathbf{Z_1'Z_2}$. La notación que se viene usando en este texto, se deja para el ACS: λ para valores propios, $\mathbf{u}$ y $\mathbf{v}$ para vectores propios, y F y G para los vectores de coordenadas de categorías fila y columna, respectivamente. Para el ACM se agrega el subíndice $\mathbf{Z}$ para los valores propios, los vectores propios y las coordenadas de las categorías. Se omiten los subíndices que hacen referencia al eje, entendiéndose que las relaciones se dan para cualquier eje. En el ACM de $\mathbf{Z}$ con respecto al ACS de $\mathbf{K}$ se obtienen las siguientes relaciones (Lebart *et al.*, 2006):

- Valores propios: $\lambda_{\mathbf{Z}} = \dfrac{1+\sqrt{\lambda}}{2}$.
- Vectores propios, si $p_2 \leq p_1$, en ACM hay $p_2 - 1$ vectores propios asociados a los valores propios $\dfrac{1+\sqrt{\lambda}}{2}$: $\mathbf{v}_Z = \begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}$, donde, $\mathbf{u}$ y $\mathbf{v}$ son vectores propios de los espacios fila y columna, respectivamente, del ACS de $\mathbf{K}$.

Los resultados de los dos análisis son diferentes pero permiten obtener las mismas descripciones. Las principales consecuencias prácticas de la comparación son:

1. Hay más ejes en el ACM, p_1+p_2-2, que en ACS, p_2-1, entonces p_1-1 ejes se pueden considerar "parásitos", en el sentido que no aportan información.

2. Las tasas de inercia retenidas son menores en el ACM con respecto al ACS.

6.4.3. El criterio de Benzécri

Benzécri (1979) propuso considerar solamente los ejes asociados a valores propios superiores al inverso del número de variables $\dfrac{1}{s}$ y recalcular las tasas de inercia mediante la fórmula:

$$\tau(\lambda_Z) = \left(\frac{s}{s-1}\right)^2 \left(\lambda_Z - \frac{1}{s}\right)^2 \quad para \quad \lambda_Z > \frac{1}{s} \tag{6.15}$$

En el caso del ACM con dos variables $\tau(\lambda_Z) = \lambda$, donde λ representa el correspondiente valor propio del AC de la tabla de contingencia que cruza las dos variables. El histograma de las tasas de inercia (6.15) se puede usar, como complemento al de valores propios, para decidir el número de ejes a retener en un ACM.

6.5. Aplicación: ACM de consumo cultural

Se presenta un ejemplo de análisis parcial de la Encuesta de Consumo Cultural del Dane (2014). La encuesta aplica un formulario a una subpoblación de niños de 5 a 11 años, sobre consumo cultural. Se adicionan algunas variables sociodemográficas de los módulos de hogares y viviendas. Para este análisis, se seleccionan los niños que tienen edades entre 8 y 11 años y que saben leer.

6.5.1. Objetivos del análisis

Describir el consumo cultural de niños entre 8 y 11 años, que saben leer y explorar su relación con algunas variables sociodemográficas.

6.5.2. Datos

Para este ejemplo se toman las siguientes preguntas del tema de consumo cultural:

Teat ¿El niño o la niña asistió a teatro, danza u ópera en los últimos doce meses? ¿Con qué frecuencia?

Libr ¿En los últimos doce meses el niño o la niña leyó libros y con qué frecuencia?

Cine ¿En los últimos doce meses el niño o la niña fue a cine y con qué frecuencia?

Vide ¿El niño o la niña vio videos en el último mes y con qué frecuencia?

Radi ¿En la última semana el niño o la niña escuchó radio y con qué frecuencia?

Musi ¿El niño o la niña escuchó música grabada en la última semana y con qué frecuencia?

La tabla activa construida para este análisis tiene 1971 niños de todo el país y seis variables cualitativas sobre consumo cultural. El formulario de la encuesta tiene dos preguntas para cada actividad: la primera, si se ha realizado, y la segunda, su frecuencia. Las respuestas a las dos preguntas en cada actividad, se recodificaron en una sola y se incluyó la categoría *no* como otro ítem de frecuencia. Adicionalmente se unieron algunas categorías de muy baja frecuencia. Las distribuciones de frecuencias de las variables se muestran en forma de tortas en la figura 6.8, donde se pueden ver sus categorías. Se tienen entonces $s = 6$ variables activas con un total de $p = 31$ categorías.

Código para leer datos y definir variables activas e ilustrativas

```
load("ninios8a11.Rda")
ninios8a11$Edad<-factor(ninios8a11$Edad)
# variables activas
Y <- subset(ninios8a11,select=c(Teat,Libr,Cine,Vide,Radi,Musi))
# variables suplementarias
Ys<-ninios8a11[,c(2,29,30,32,35)]
```

Como variables sociodemográficas ilustrativas se seleccionaron las siguientes:

Pare ¿Cuál es el parentesco del niño o la niña con el(la) jefe(a) del hogar?

Sexo Sexo.

Edad ¿Cuántos años cumplidos tiene?

Regi Región a la que corresponde la vivienda.

Estr Estrato para tarifa.

Las variables ilustrativas se muestran en la figura 6.9 junto con las distribuciones de frecuencias. Las variables estrato y parentesco, se recodificaron a menos categorías, dada la presencia de frecuencias muy bajas en algunas de ellas.

6.5.3. Resultados del análisis

Para realizar el ACM se usa la función `dudi.acm{ade4}`; las ayudas para la interpretación se obtienen con `inertia.dudi{ade4}`; las coordenadas y ayudas para la interpretación de las variables cualitativas ilustrativas, con `supqual{FactoClass}`; los planos factoriales, con `plot.dudi{FactoClass}` para las variables activas y `plotfp{FactoClass}` para los planos factoriales cuando se desean solo las variables cualitativas ilustrativas.

6.5.3.1. Número de ejes a interpretar

La primera decisión a tomar es el número de ejes a interpretar. No hay recetas pero sí criterios que ayudan a tomar esta decisión. La guía principal es la forma del histograma de valores propios.

Los ejes que sobresalen claramente, antes de ver una forma de *S* regular del histograma serían los ejes a analizar.

Código para obtener la figura 6.8

```
par(las=1,mfrow=c(2,3),mai=c(0.55,0.5,0.1,0.1))
for(i in 1:6){
  cat<-attributes(Y[,i])$levels;
  per<-tabulate(Y[,i])/nrow(Y)*100;
  pl<-plot(Y[,i],horiz=TRUE,col=gray(seq(1.0,0.9,
           length=length(cat))),ylim=c(0,8),
           xlim=c(0,1400),xlab=colnames(Y)[i]);
  text(800,pl,round(per,1),cex=0.8,pos=4);
}
  dev.print(device = pdf) # grabar la grafica como Rplots.pdf
```

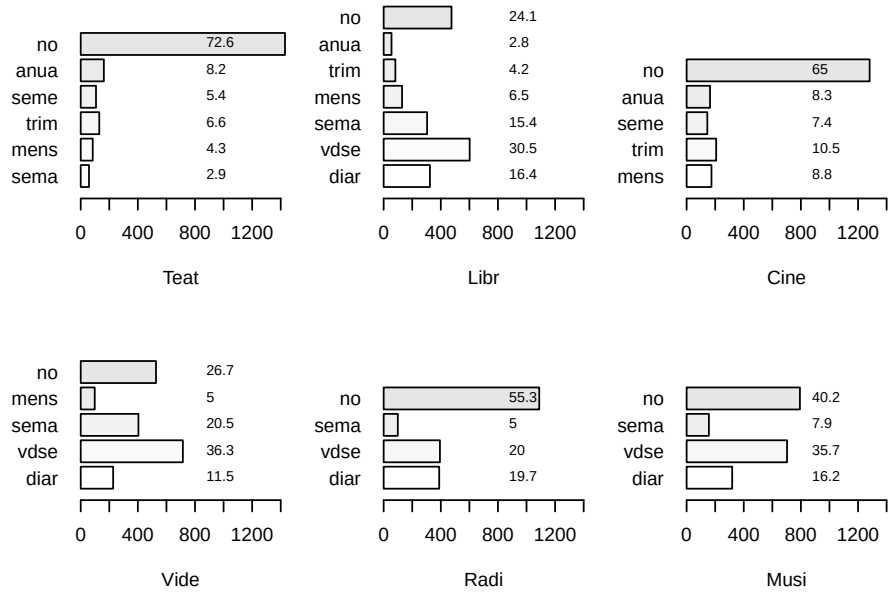


Figura 6.8. Distribuciones de frecuencias de las variables activas del ACM consumo cultural: ir a teatro, lectura de libros, ir a cine, ver videos, escuchar radio, oír música. *vdse* significa "varios días a la semana". Los números son porcentajes. Nótese los bajos consumos de teatro, cine y radio, y que alrededor de la cuarta parte de los niños dicen no leer libros

El criterio de Benzécri, que utiliza un histograma de una transformación de los valores propios, considerando los que son superiores a $1/s$, también ayuda. Finalmente, en concordancia con los objetivos, se debe interpretar un eje adicional si provee información relevante que no se ha obtenido con los ejes anteriores.

Para el ejemplo se muestran en la figura 6.10 los histogramas de valores propios y del criterio de Benzécri, los valores propios y porcentajes de inercia. Los gráficos sugieren analizar tres ejes.

El criterio de Benzécri se construye considerando los 11 valores propios, que tienen un valor superior a $1/6 = 0.1667$. Los tres primeros ejes retienen el 17.8 % de la inercia, pero en el ACM este es un índice pesimista.

Código para obtener la figura 6.9

```
par(las=1,mfrow=c(2,3),mai=c(0.55,0.5,0,0.1))
for(i in 1:5){
  cat<-attributes(Y[,i])$levels;
  per<-tabulate(Ys[,i])/nrow(Y)*100;
  pl<-plot(Ys[,i],horiz=TRUE,col=gray(seq(1.0,0.9,
           length=length(cat))),ylim=c(0,8),
           xlim=c(0,1000),xlab=colnames(Ys)[i]);
  text(100,pl,round(per,1),cex=0.8,pos=4);
}
dev.print(device = pdf) # grabar la grafica como Rplots.pdf
```

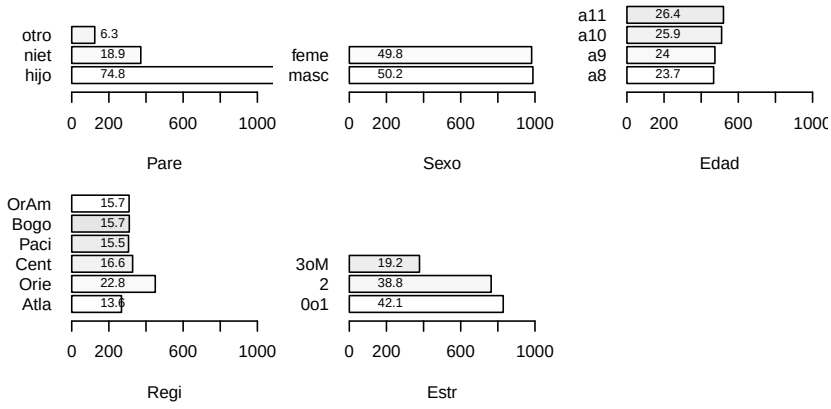


Figura 6.9. Distribuciones de frecuencias de las variables ilustrativas del ACM consumo cultural: parentesco con el jefe del hogar, sexo, edad, región, estrato (0y1 = muy bajo, 2 = bajo, 3oM = medio y alto). Los números son porcentajes

En el criterio de Benzècri se retiene un 81.8 % de la variabilidad ajustada, pues el criterio considera que 25 – 11 = 14 ejes son "parásitos", es decir, no suministran información. El bajo porcentaje de inercia retenida en los ejes, hace que el analista tenga que ser menos exigente en las calidades de las representaciones.

En la figura 6.11 se muestra el primer plano factorial con categorías activas. Las coordenadas y ayudas para la interpretación para los tres primeros ejes están en la tabla 6.7. Una guía para saber en qué categorías poner atención en cada uno de los ejes es ver aquellos que superen el porcentaje promedio de contribución a la inercia del eje (contribución absoluta): en este ejemplo 100/31 = 3.2 (están en negrilla en la tabla).

6.5.3.2. Primer eje factorial

El primer eje factorial contrapone: ir a teatro al menos una vez al año (-) contra no ir (+), leer libros diariamente (-) contra no leer (+), ir a cine mensualmente (-) contra no ir (+), ver videos diariamente (-), contra no ver (+), escuchar radio diariamente (-) contra no escuchar (+) y escuchar música grabada diariamente (-) contra no escuchar (+). Es decir, separa a los niños de mayor consumo cultural, al lado negativo, de los que no tiene consumo cultural, al lado positivo. Entonces el primer eje es un indicador de consumo cultural y ordena a los niños: al lado negativo los que más consumen, al lado positivo los que no consumen.

6.5.3.3. Segundo eje factorial

El segundo eje contrapone valores medios contra valores extremos en algunas variables, lo que indica que hay presencia de efecto Guttman: leer varias veces a la semana contra no leer o hacerlo diariamente, ver videos varias veces a la semana contra no ver o hacerlo diariamente, escuchar radio varios días a la semana contra hacerlo diariamente y escuchar música grabada varios días a la semana contra escucharla diariamente o no hacerlo.

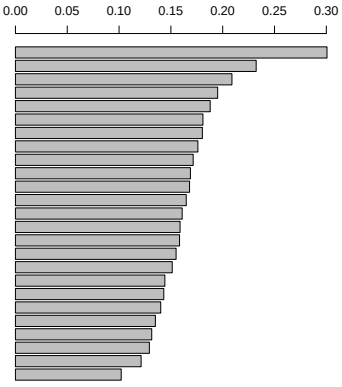
6.5.3.4. Tercer eje factorial

El tercer eje separa algunas categorías intermedias: ver videos varios días a la semana contra hacerlo semanal o mensualmente, escuchar radio varias veces a la semana contra hacerlo semanalmente; y escuchar música grabada varios días a la semana contra semanalmente.

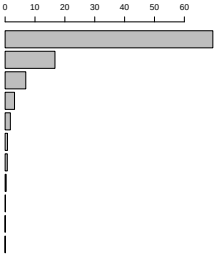
Para obtener las gráficas y la tabla de la figura 6.10

```
library(FactoClass)
acm<-dudi.acm(Y,scannf = FALSE,nf=3)
barplot(acm$eig,las=3)
#dev.print(device = xfig)#,encoding="latin1")
eigtab<-data.frame(valp=acm$eig,porc=acm$eig/sum(acm$eig)*100,
                   pacu=cumsum(acm$eig)/sum(acm$eig)*100)
xtable(cbind(eje=1:8,eigtab[1:8,],eje=9:16,eigtab[9:16,],
       eje=17:24,eigtab[17:24,]),digits=c(0,rep(c(0,3,1,1),3)))
# criterio de Benzecri
s<-6; 1/s
# --> se calcula tau para los primeros 11 ejes
eig11<-acm$eig[1:11]
tau<-(s/(s-1))^2*(eig11-(1/s))^2
ptau<-tau/sum(tau)*100
```

Histograma de valores propios



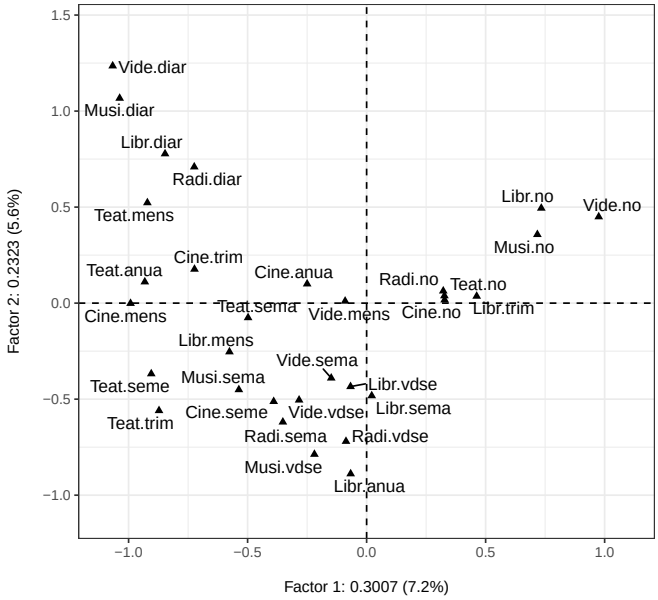
Histograma del criterio de Benzécri



Valores propios y porcentajes de inercia

eje	valp	porc	pacu	eje	valp	porc	pacu	eje	valp	porc	pacu
1	0.301	7.2	7.2	9	0.171	4.1	44.0	17	0.151	3.6	74.9
2	0.232	5.6	12.8	10	0.169	4.1	48.1	18	0.144	3.5	78.3
3	0.209	5.0	17.8	11	0.168	4.0	52.1	19	0.143	3.4	81.8
4	0.195	4.7	22.5	12	0.165	4.0	56.1	20	0.140	3.4	85.1
5	0.188	4.5	27.0	13	0.161	3.9	59.9	21	0.135	3.2	88.4
6	0.181	4.3	31.3	14	0.159	3.8	63.7	22	0.132	3.2	91.5
7	0.180	4.3	35.7	15	0.158	3.8	67.5	23	0.129	3.1	94.6
8	0.176	4.2	39.9	16	0.155	3.7	71.2	24	0.121	2.9	97.6

Figura 6.10. Histogramas de valores propios y del criterio de Benzécri ACM de consumo cultural y tabla de valores propios




```
plot(acm,Trow=FALSE,gg=TRUE,xlim=c(-1,1),ylim=c(-1,1.3),
cframe=1.1),col.col="black",cex.global=0.8)
```

Figura 6.11. Primer plano factorial del ACM de frecuencia de lectura de niños, mostrando las categorías activas.

Aunque el tercer eje permite ver algunas diferencias, mejor que en los dos primeros ejes, se puede prescindir de este en una síntesis del consumo cultural de los niños.

6.5.3.5. Primer plano factorial

Se observa claramente el efecto Gutttman (forma de parábola en las variables ordinales) en todas las variables, es decir, las categorías están ordenadas: arriba a la derecha están las que indican que no hay consumo; a medida que se desciende en dirección a la izquierda, el consumo cultural va aumentando; luego arriba a la izquierda se ubican las categorías de mayor consumo cultural.

Como las categorías se ubican en el promedio de las coordenadas de los niños que las asumen (alejado por el inverso de la raíz del valor propio), los niños quedan ordenados de la misma forma, según su consumo cultural.

Tabla 6.7. Coordenadas y ayudas para la interpretación de las categorías del ACM de frecuencia de lectura en niños

	peso	Coordenadas			Contribuciones abs.			Cosenos cuadrados			contr.
Categoria	%	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3	inercia
Teat.sema	0.5	-0.498	-0.076	-0.307	0.4	0.0	0.2	0.8	0.0	0.3	3.9
Teat.mens	0.7	-0.921	0.523	0.284	2.0	0.8	0.3	3.8	1.2	0.4	3.8
Teat.trim	1.1	-0.872	-0.560	-0.280	2.8	1.5	0.4	5.4	2.2	0.6	3.7
Teat.seme	0.9	-0.905	-0.368	-0.512	2.5	0.5	1.1	4.7	0.8	1.5	3.8
Teat.anua	1.4	-0.932	0.111	0.076	**4.0**	0.1	0.0	7.8	0.1	0.1	3.7
Teat.no	12.1	0.327	0.038	0.051	**4.3**	0.1	0.1	28.2	0.4	0.7	1.1
Libr.diar	2.7	-0.847	0.778	-0.113	**6.5**	**7.1**	0.2	14.1	11.9	0.2	3.3
Libr.vdse	5.1	-0.067	-0.434	0.365	0.1	**4.1**	**3.3**	0.2	8.3	5.9	2.8
Libr.sema	2.6	0.021	-0.482	-0.307	0.0	2.6	1.2	0.0	4.2	1.7	3.4
Libr.mens	1.1	-0.576	-0.254	-0.673	1.2	0.3	2.4	2.3	0.5	3.2	3.7
Libr.trim	0.7	0.462	0.036	-0.162	0.5	0.0	0.1	0.9	0.0	0.1	3.8
Libr.anua	0.5	-0.067	-0.889	-0.162	0.0	1.6	0.1	0.0	2.3	0.1	3.9
Libr.no	4.0	0.734	0.494	0.040	**7.2**	**4.2**	0.0	17.1	7.8	0.0	3.0
Cine.mens	1.5	-0.992	-0.001	0.879	**4.8**	0.0	**5.4**	9.5	0.0	7.5	3.6
Cine.trim	1.8	-0.723	0.177	-0.384	3.0	0.2	1.2	6.1	0.4	1.7	3.6
Cine.seme	1.2	-0.390	-0.512	-0.332	0.6	1.4	0.6	1.2	2.1	0.9	3.7
Cine.anua	1.4	-0.250	0.100	-0.101	0.3	0.1	0.1	0.6	0.1	0.1	3.7
Cine.no	10.8	0.328	0.017	-0.007	**3.9**	0.0	0.0	19.9	0.1	0.0	1.4
Vide.diar	1.9	-1.067	1.235	0.379	**7.3**	**12.6**	1.3	14.8	19.9	1.9	3.5
Vide.vdse	6.0	-0.284	-0.505	0.542	1.6	**6.6**	**8.5**	4.6	14.5	16.7	2.5
Vide.sema	3.4	-0.149	-0.390	-0.822	0.3	2.2	**11.1**	0.6	3.9	17.4	3.2
Vide.mens	0.8	-0.090	0.011	-1.464	0.0	0.0	**8.5**	0.0	0.0	11.2	3.8
Vide.no	4.5	0.975	0.450	0.004	**14.1**	**3.9**	0.0	34.7	7.4	0.0	2.9
Radi.diar	3.3	-0.724	0.709	-0.103	**5.7**	**7.1**	0.2	12.9	12.4	0.3	3.2
Radi.vdse	3.3	-0.087	-0.720	0.475	0.1	**7.4**	**3.6**	0.2	12.9	5.6	3.2
Radi.sema	0.8	-0.352	-0.619	-1.867	0.3	1.4	**14.0**	0.7	2.0	18.4	3.8
Radi.no	9.2	0.322	0.063	0.035	**3.2**	0.2	0.1	12.8	0.5	0.1	1.8
Musi.diar	2.7	-1.037	1.067	0.286	**9.7**	**13.2**	1.1	20.8	22.0	1.6	3.4
Musi.vdse	5.9	-0.219	-0.787	0.443	1.0	**15.9**	**5.6**	2.7	34.3	10.9	2.6
Musi.sema	1.3	-0.537	-0.451	-2.151	1.3	1.2	**29.2**	2.5	1.7	39.8	3.7
Musi.no	6.7	0.717	0.357	-0.085	**11.5**	**3.7**	0.2	34.6	8.6	0.5	2.4

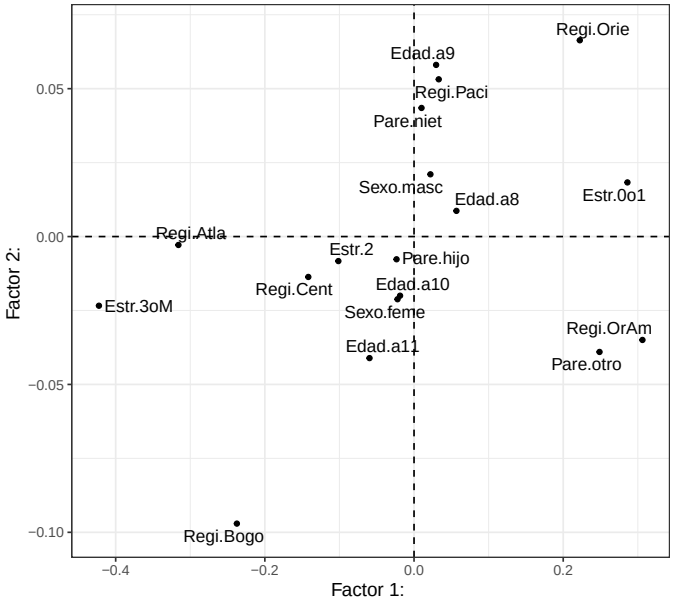
```
ayu<-inertia.dudi(acm,,T); table(cbind(peso=acm$cw*100,acm$co,
ayu$col.abs/100,abs(ayu$col.rel)/100),digits=c(0,1,rep(3,3),rep(1,7))
```

6.5.3.6. Variables suplementarias

En la figura 6.12 se muestran las categorías de las variables sociodemográficas, proyectadas como suplementarias en el primer plano factorial y las coordenadas y ayudas para la interpretación sobre los tres primeros ejes factoriales. El plano debe verse como un *zoom* del centro de la figura 6.11.

Las categorías que tienen valores test inferiores a -2 y superiores a 2 (están en negrilla), son las que se pueden interpretar. Solo es pertinente interpretar el primer eje (horizontal) del plano factorial.

Entonces, las diferencias del consumo cultural se pueden explicar en parte por el nivel socioeconómico (el estrato es una aproximación): el consumo cultural aumenta a medida que mejora el nivel socioeconómico. Pero también por el parentesco con el jefe del hogar: los niños que viven con personas diferentes a los padres o abuelos tienen menor consumo cultural, y por la región geográfica donde viven: los niños de las regiones Atlántica, Bogotá y Central suelen tener mayor consumo cultural que los de las regiones Orinoquía-Amazonía y Oriental.




```
sc<-supqual(acm,Ys)
plotfp(as.data.frame(sc$coor),col="black",cframe=1,gg=TRUE)
```

Figura 6.12. Proyección de categorías suplementarias sobre el primer plano factorial del ACM de consumo cultural

Tabla 6.8. Coordenadas y ayudas para la interpretación de las categorías suplementarias del ACM de consumo cultural

Cate	%	dis	Coordenadas			Valores test			Cosenos cuadrados		
goria		tan^2	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3
Pare.hijo	42.4	0.336	-0.023	-0.008	0.031	-1.794	-0.588	**2.394**	0.002	0.000	0.003
Pare.niet	10.7	4.298	0.010	0.043	-0.088	0.215	0.931	-1.887	0.000	0.000	0.002
Pare.otro	3.6	14.895	0.249	-0.039	-0.108	**2.860**	-0.449	-1.238	0.004	0.000	0.001
Sexo.masc	28.5	0.993	0.022	0.021	0.019	0.979	0.937	0.846	0.000	0.000	0.000
Sexo.feme	28.3	1.007	-0.022	-0.021	-0.019	-0.979	-0.937	-0.846	0.000	0.000	0.000
Edad.a8	13.4	3.221	0.057	0.009	-0.022	1.406	0.215	-0.539	0.001	0.000	0.000
Edad.a9	13.6	3.158	0.030	0.058	-0.025	0.739	1.450	-0.632	0.000	0.001	0.000
Edad.a10	14.7	2.865	-0.019	-0.020	0.047	-0.491	-0.524	1.241	0.000	0.000	0.001
Edad.a11	15.0	2.790	-0.060	-0.041	-0.004	-1.586	-1.092	-0.100	0.001	0.001	0.000
Regi.Atla	7.7	6.354	-0.316	-0.003	-0.125	**-5.561**	-0.050	**-2.209**	0.016	0.000	0.002
Regi.Orie	12.9	3.380	0.222	0.066	0.092	**5.365**	1.602	**2.220**	0.015	0.001	0.003
Regi.Cent	9.4	5.009	-0.142	-0.014	-0.069	**-2.811**	-0.271	-1.378	0.004	0.000	0.001
Regi.Paci	8.8	5.441	0.033	0.053	0.088	0.630	1.012	1.672	0.000	0.001	0.001
Regi.Bogo	8.9	5.358	-0.238	-0.097	0.006	**-4.554**	-1.861	0.119	0.011	0.002	0.000
Regi.OrAm	8.9	5.379	0.306	-0.035	-0.045	**5.861**	-0.669	-0.853	0.017	0.000	0.000
Estr.0o1	23.8	1.378	0.286	0.018	0.010	**10.819**	0.692	0.378	0.059	0.000	0.000
Estr.2	22.0	1.580	-0.101	-0.008	-0.043	**-3.582**	-0.293	-1.513	0.007	0.000	0.001
Estr.3oM	10.9	4.214	-0.422	-0.023	0.065	**-9.133**	-0.506	1.398	0.042	0.000	0.001

```
xtable(data.frame(por=sc$ncat/34.76,dis2=sc$dis2,coor=sc$coor,
vt=sc$tv,cos2=sc$cos2),digits=c(0,1,rep(3,10)))
```

6.5.4. Conclusiones del análisis

Los niños entre 8 y 11 años de la muestra del Dane (2014) tienen un bajo consumo cultural: menos de la mitad asisten a teatro, cine o escuchan radio; casi la cuarta parte dice no leer libros (figura 6.8).

Las categorías se ordenan en el primer plano factorial, en forma de parábola, lo que muestra la tendencia general de que los niños que tienen bajo consumo cultural, lo tienen simultáneamente en las seis actividades. Entonces, el primer plano factorial es un indicador del consumo cultural y se pueden vislumbrar grupos de niños: los ubicados arriba a la izquierda son los de mayor consumo y los de arriba a la derecha los de menor consumo. En el centro del plano hacia abajo se ubican los de consumo medio (6.11).

Si se desea ordenar a los niños por su consumo cultural, se pueden utilizar las coordenadas del primer eje, pues este es también un indicador de consumo cultural.

El nivel de consumo cultural está explicado en parte, por el nivel socioeconómico (aproximado con el estrato), la relación con el jefe del hogar y la región donde viven los niños. Los niños suelen tener más consumo cultural a mayor nivel socioeconómica o si viven con padres o abuelos y si son de las regiones Atlántica, Bogotá o Central (6.12).

6.6. Ejercicios

1. Obtenga las fórmulas del ACM como el ACS de la tabla disyuntiva completa.

2. Obtenga e interprete la distancia entre dos individuos.

3. Demuestre que la distancia entre dos categorías j y k es igual a $d^2(j,k) = \frac{n}{n_j n_k}(b+c)$, donde b es el número de individuos que asumen la categoría j pero no la k y c cumple la condición contraria, es decir, $(b+c)$ es el número de individuos que asumen una y solo una de las dos categorías.

4. Demuestre que la subnube de categorías de una misma variable tiene el mismo centro de gravedad de la nube completa.

5. Demuestre que la nube de categorías está contenida en un subespacio de dimensión $p - s$, siendo p el número de categorías y s el número de variables.

6. Obtenga e interprete las inercias: de una categoría, de una variable y total.

7. Demuestre que el valor test asociado a la coordenada sobre el eje de una variable suplementario j es: $t_s(j) = \sqrt{\frac{n_j(n-1)}{n-n_j}} G_s(j)$.

8. Demuestre que un ACM cuando todas las variables tienen dos categorías es equivalente a un ACP normado de una de las categorías por cada una de las variables.

9. Demuestre que el criterio de Benzécri para un ACM con dos variables coincide con las tasas de inercia del ACS.

10. Demuestre la igualdad de la ecuación (6.12).

6.7. Talleres de ACM

6.7.1. ACM de razas de perros

Este taller se encuentra en Fine (1996) y es un ejemplo pequeño para entender los conceptos del ACM.

Objetivo

Seleccionar las razas de perros de acuerdo con la función para la que se utilizan: compañia, caza o utilidad (salvamento, defensa, perro lazarillo o policía, etc.).

Los datos se encuentran en el paquete `FactoClass` como `DogBreeds`. Para cada una de las veintisiete razas estudiadas se registran seis variables que miden las cualidades físicas o psíquicas de la raza y la variable función, que es ilustrativa en el análisis:

Variables	Categorías		
Tamaño	Pequeño	Medio	Grande
Peso	Liviano	Medio	Pesado
Velocidad	Baja	Media	Alta
Inteligencia	Pequeña	Media	Grande
Afectividad	Pequeña	Grande	
Agresividad	Pequeña	Grande	
Función	Compañía	Caza	Utilidad

Realice el ACM y resuelva las preguntas que aparecen a continuación.

Preguntas

1. A partir del archivo de datos responda:

   a) ¿Cuáles son las características del perro CANI?

   b) Identifique los pares de razas ("individuos") que presentan características idénticas.

2. Construya la tabla disyuntiva completa (TDC) y observándola responda:

   a) ¿Qué categorías presenta la raza bóxer para cada una de las variables?

   b) ¿Cuántas razas de perros se caracterizan por poseer una inteligencia media y cuáles son?

3. Construya la tabla de Burt (se puede pedir en el ACM) y observándola responda:

   a) ¿Cómo se distribuyen las razas de perros según la variable peso?

   b) ¿Cuántas razas de perros son muy inteligentes y poco afectuosas?

   c) ¿Cuántas razas de perros tienen inteligencia media o superior y gran tamaño?

4. ¿Cuántos ejes factoriales considera razonable interpretar?

5. ¿Cuáles son las categorías que constituyen el primer eje? (Contribución mayor que el promedio)

6. ¿Qué categorías tienen coordenadas importantes en el primer eje y de qué signos son?

7. ¿Cuáles son las razas que se encuentran más alejadas del origen? ¿Cuáles son sus coordenadas sobre el primer eje?

8. ¿Cuáles son las categorías más contributivas al segundo eje?

9. Observando la coordenada de la categoría baja de la característica velocidad. ¿En qué dirección del segundo eje se encontrarán los perros poco veloces? Teniendo en cuenta esto y razonando sobre el espacio de los individuos, ¿qué razas de perros podrían considerarse poco veloces?

10. ¿Es posible distinguir grupos de categorías en el primer plano factorial? ¿Cuántos grupos? ¿Qué categorías integran cada uno de ellos?

11. En el gráfico de las categorías activas e ilustrativas (función) sobre el primer plano factorial, ¿a qué grupo de categorías activas se encuentran vinculadas cada una de las categorías de la variable suplementaria?

12. En el gráfico simultáneo de individuos y categorías, ¿qué razas de perros corresponden a cada una de los grupos de categorías identificados?. Es decir, ¿qué razas de perros conforman cada grupo?

13. Para cada grupo de razas de perros que usted ha definido, calcule los perfiles de las características observadas. Es decir, ¿cuáles son las características de cada uno de los grupos de razas?

14. Compare los perfiles de los grupos de razas y exprese en unas pocas frases las conclusiones.

6.7.2. Comparación de AC

Caso de dos variables

Para las variables *carr* y *estr* realice el ACS, el ACM y el AC de la tabla de Burt respectiva. Verifique que se cumplen cada una de las fórmulas dadas en Lebart, Piron & Morineau (2006) para la comparación.

Escriba la tabla 5.2 reemplazando las fórmulas por los resultados numéricos del ejercicio. Haga un resumen de la comparación. Responda las siguientes preguntas:

1. En el ACS de la TC, las coordenadas del centro de gravedad de la nube de los perfiles fila son:

   _____

2. En el ACS, la nube está soportada en: _____ dimensiones.
3. La inercia asociada al ACS es: _____ ¿Qué significa? _____

   _____

4. En el ACS, el porcentaje de inercia retenido por el primer plano es: _____%.

5. En el ACM equivalente al ACS, la TDC tiene _____ filas y _____ columnas.
6. En el ACM de la TDC las nubes de "individuos" y de categorías tienen, respectivamente, _____ y _____ puntos.
7. En este ACM las nubes están soportadas en _____ dimensiones.
8. La inercia asociada al ACM es: _____ y su significado estadístico es:

   _____

9. La tabla de Burt asociada a la TDC tiene _____ filas y _____ columnas y su total es: _____
10. Según el criterio de Benzécri, en el ACM para el histograma de los τ, se deben tener en cuenta _____ ejes y el primer plano retiene el _____% de la suma de los τ.

ACM y AC de la tabla de Burt.

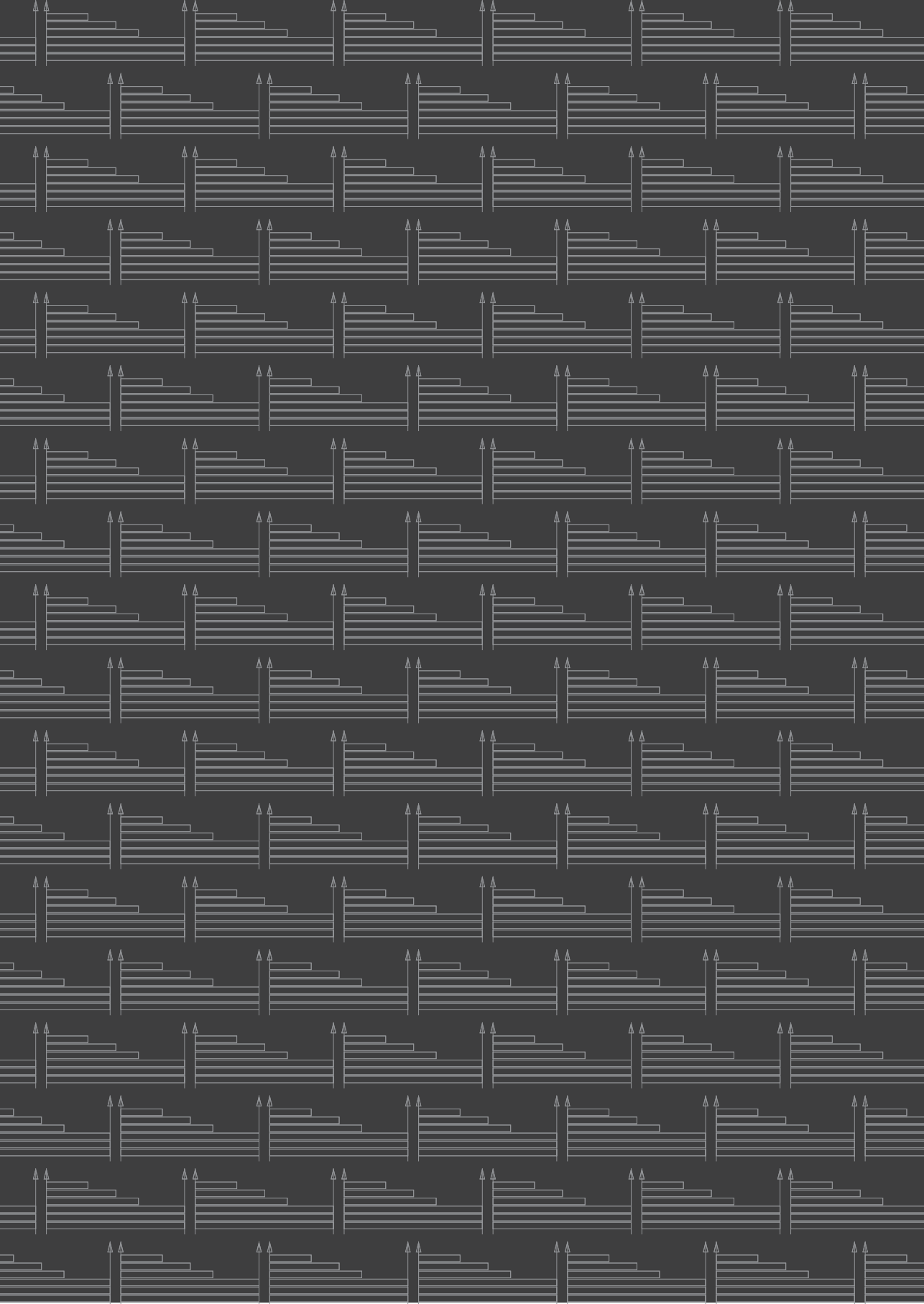
Ejecute el ACS de la tabla de Burt del taller *razas de perros* y compare los resultados con el ACM ya realizado, verificando las fórmulas de comparación dadas en Lebart, Piron & Morineau (2006). Haga un resumen de la comparación.

Realice el histograma del Benzécri para el ACM de *razas de perros* y utilícelo en la decisión de cuántos ejes interpretar.

Capítulo

siete

Métodos de clasificación



Los métodos de clasificación que se abordan en este documento, son los denominados en inglés *Cluster Analysis*, expresión que se puede traducir como "métodos de agrupamiento", cuyo objetivo es descubrir patrones en los datos en forma de grupos bien diferenciados, que tengan individuos homogéneos en su interior. En las áreas de minería de datos, aprendizaje automático y reconocimiento de patrones, se conocen con el nombre de *métodos de clasificación no supervisada*. La literatura francesa de análisis de datos los denomina *métodos de clasificación automática*.

En el sentido matemático un algoritmo de agrupamiento busca una partición de un conjunto de n elementos en K subconjuntos, que es lo mismo que definir una variable cualitativa emergente de los datos.

En primera instancia, se conocen dos tipos de métodos:

1. Los que permiten obtener una partición directa mediante un algoritmo. El más conocido y utilizado es el *K-means*.

2. Los que construyen una sucesión de particiones anidadas, que se representan mediante un árbol o dendrograma. Se conocen como *métodos de clasificación jerárquica*. Los más utilizados son los de clasificación jerárquica aglomerativa, que parten de todos los individuos, como n clases de un elemento, y se van uniendo en pasos sucesivos hasta llegar a un solo grupo o clase de n individuos.

Los algoritmos de clasificación utilizan medidas de similitud, disimilitud o distancia entre individuos y entre grupos. Las similitudes, disimilitudes o distancias entre grupos constituyen los criterios de agregación de los métodos de clasificación jerárquica aglomerativa.

En este documento se sigue la propuesta proveniente de la literatura francesa, que es combinar los dos tipos de métodos de clasificación, para obtener una mejor partición, y además combinarlos con los métodos en ejes principales (Lebart, Morineau & Piron, 1995; Lebart, Piron & Morineau, 2006). Para el estudio general de los métodos de agrupamiento se recomienda el texto de Everitt, Landau, Leese & Stahl (2011).

7.1. Obtener una partición directa

En estos algorítmos se dan el número de clases, los puntos iniciales requeridos para empezar el algoritmo y un criterio de parada. Uno de los métodos más conocidos es el *K-means* (Ball & Hall, 1965; MacQueen, 1967), que está relacionado con la geometría utilizada en los métodos en ejes principales

porque recurre a la distancia euclidiana entre individuos; y la distancia entre grupos se calcula como la distancia euclidiana entre sus centros de gravedad.

Los criterios de homogeneidad intra grupos y de heterogeneidad entre grupos, implícitos en el método *K-means*, están definidos a partir de la inercia: en una nube de puntos dotada de una partición en K clases.

7.1.1. Descomposición de la inercia

Sea una nube de n puntos N_n en $\mathbb{R}^p$ con una partición en K clases. Entonces, la inercia total de la nube de puntos con respecto a su centro de gravedad $\mathbf{g}$ se puede descomponer en inercia entre clases e inercia intraclases:

$$Inercia(N_n) = \sum_{i=1}^{n} p_i d^2(i, \mathbf{g}) = \sum_{k=1}^{K} p_k d^2(\mathbf{g}_k, \mathbf{g}) + \sum_{k=1}^{K} \sum_{i \in I_k} p_i d^2(i, \mathbf{g}_k) \quad (7.1)$$

donde:

p_i: peso del individuo i, $\sum_{i=1}^{n} p_i = 1$.

$\mathbf{g}$: centro de gravedad de la nube de puntos, $\mathbf{g} = \sum_{i=1}^{n} p_i \mathbf{x}_i$, $\mathbf{x'}_i$, es la fila i de la matriz de coordenadas de los puntos $\mathbf{X}$ con n filas y p columnas.

p_k: peso de la clase k, $p_k = \sum_{i \in I_k} p_i$.

$\mathbf{g}_k$: centro de gravedad de la clase k, $\mathbf{g}_k = \sum_{i \in I_k} \frac{p_i}{p_k} \mathbf{x}_i$.

$d^2(.,.)$ es la distancia euclidiana canónica.

El primer término de la fórmula (7.1) es la inercia entre clases y el segundo, la inercia intraclases. En el cálculo de la inercia ha intervenido la distancia euclidiana canónica. Así pues, la medida de disimilitud entre individuos ya está seleccionada.

El método *K-means*, como se verá en la sección siguiente, busca una partición en K clases que tenga inercia intraclases mínima.

7.1.2. Agregación alrededor de centros móviles: $K-means$

Dentro de los algoritmos que permiten obtener particiones del número de clases deseado, el método *K-means* es uno de los más utilizados y forma parte de los métodos, conocidos en la literatura francesa, como *de agregación alrededor de centros móviles*.

A continuación, se resume el procedimiento descrito en Lebart *et al.* (2006) utilizando también su notación, que es útil para entender la lógica e interpretación del *K-means*. Los algoritmos implementados en los programas estadísticos son un poco diferentes, ya que están optimizados para que sean más eficientes.

Se busca una partición en K clases de un conjunto I de n individuos, descritos por p variables continuas. Se tiene entonces una nube de n puntos-individuos en $\mathbb{R}^p$, dotada de una distancia euclidiana d.

El algoritmo procede de la manera siguiente (en la notación el superíndice indica el número de la etapa o paso dentro del algoritmo y el subíndice, da la clase):

- Paso 0

  Se dan K centros iniciales de las clases: $\{C_1^0, C_2^0, \ldots, C_k^0, \ldots, C_K^0\}$, que inducen a una partición de I en K clases $P^0 = \{I_1^0, I_2^0, \ldots, I_k^0, \ldots, I_K^0\}$. De tal forma que el individuo i pertenece a la clase I_k^0 si el punto i está más próximo de C_k^0 que de todos los demás centros.

- Paso 1

  Se determinan los K centros de gravedad $\{C_1^1, C_2^1, \ldots, C_k^1, \ldots, C_K^1\}$ de las clases $\{I_1^0, I_2^0, \ldots, I_k^0, \ldots, I_K^0\}$.

  Estos nuevos centros llevan a una nueva partición construida con la misma regla: $P^1 = \{I_1^1, I_2^1, \ldots, I_k^1, \ldots, I_K^1\}$.

- Paso m

  Se determinan K nuevos centros de las clases $\{C_1^m, C_2^m, \ldots, C_k^m, \ldots, C_K^m\}$ tomando los centros de gravedad de las clases en el paso $m-1$:

  $\{I_1^{m-1}, I_2^{m-1}, \ldots, I_k^{m-1}, \ldots, I_K^{m-1}\}$.

  Estos nuevos centros inducen a una nueva partición del conjunto I:

  $P^m = \{I_1^m, I_2^m, \ldots, I_k^m, \ldots, I_K^m\}$.

El algoritmo se detiene si la nueva partición no es mejor que la anterior (la inercia intraclases deja de disminuir), o si dos iteraciones sucesivas dan la misma partición, o porque se ha alcanzado un número máximo de iteraciones fijado de antemano. Generalmente la partición obtenida depende de la selección inicial de los centros.

El algoritmo $K-means$ disminuye la inercia intraclases.

Hay que mostrar que la partición $P^m = \{I_1^m, I_2^m, \ldots, I_k^m, \ldots, I_K^m\}$ tiene una inercia intraclases menor o igual a la de la partición P^{m-1} de la etapa anterior.

A cada individuo del conjunto a clasificar, se le asocia un peso $p_i > 0$ tal que $\sum_{i=1}^{n} p_i = 1$. $d^2(i, C_k^m)$ es el cuadrado de la distancia entre el individuo i y el centro inicial de la clase k en la etapa m, que es el centro de gravedad de la clase k en el paso $m-1$. Entonces, la suma de las inercias de las clases de la partición P^m con respecto a los puntos que permitieron construirla es:

$$v(m) = \sum_{k=1}^{K} \sum_{i \in I_k^m} p_i d^2(i, C_k^m) \tag{7.2}$$

Recordemos que en la etapa m, I_k^m es el conjunto de los individuos que están más próximos a C_k^m que de todos los otros centros, y que el centro de gravedad de esta clase se calcula en la etapa $m+1$: $C_k^{m+1} = \mathbf{g}_k^m$.

La inercia intraclases en la etapa m es la cantidad:

$$V(m) = \sum_{k=1}^{K} \sum_{i \in I_k^m} p_i d^2(i, C_k^{m+1}) \tag{7.3}$$

donde C_k^{m+1} es el centro de gravedad de la clase I_k^m, que es el nuevo centro en la etapa $m+1$.

La suma de las inercias con respecto a los puntos que originaron la partición P^{m+1} es

$$\mathbf{v}(m+1) = \sum_{k=1}^{K} \sum_{i \in I_k^{m+1}} p_i d^2(i, C_k^{m+1}) \tag{7.4}$$

$V(m)$ de (7.3) es menor o igual que $v(m)$ de (7.2) porque la inercia con respecto al centro de gravedad es siempre menor a la inercia con respecto a cualquier otro punto.

La cantidad $v(m+1)$ de (7.4) es menor o igual a $V(m)$ de (7.3) debido a que si en la nueva partición al menos un individuo cambia de clase queda más cerca de otro nuevo centro.

Entonces, $v(m+1) \leq V(m) \leq v(m) \leq V(m-1)$, es decir, la inercia intraclases disminuye con cada paso del algoritmo.

Ejemplo "Café"

Como "ejemplo de juego" vamos a utilizar el ejemplo "Café" del capítulo 3, partiendo de las coordenadas sobre el primer plano factorial (figura 3.7).

Buscaremos dos clases ejecutando "a mano" el procedimiento *K-means* partiendo de los cafés: ExCl y O40C, como centros iniciales.

En la tabla 7.1 se muestran los pasos del proceso. Las coordenadas sobre el primer plano factorial son las dos primeras columnas de la tabla 7.1 (abajo), ordenadas por las coordenadas sobre el primer eje. En la figura 7.1 se puede visualizar el procedimiento.

Pasos del $K-means$ de los diez cafés

Aquí se muestran los centros y las particiones de cada paso del algoritmo, extraídos de tabla 7.1.

Paso 0: $K = 2$; centros iniciales: $\begin{cases} C_1^0 = [-0.89, -1.68] \\ C_2^0 = [0.15, 1.31] \end{cases}$

La partición es: $P_0 = \begin{cases} I_1^0 = \{ExOs, ExCl, C20C, C20M\} \\ I_2^0 = \{O20C, O20M, O40C, O40M, C40C, C40M\} \end{cases}$

Paso 1: Centros de gravedad de P_0: $\begin{cases} C_1^1 = [-0.66, -0.89] \\ C_2^1 = [0.44, 0.59] \end{cases}$

Partición: $P_1 = \begin{cases} I_1^1 = \{ExOs, O20C, ExCl, C20C\} \\ I_2^1 = \{O20M, O40C, O40M, C20M, C40C, C40M\} \end{cases}$

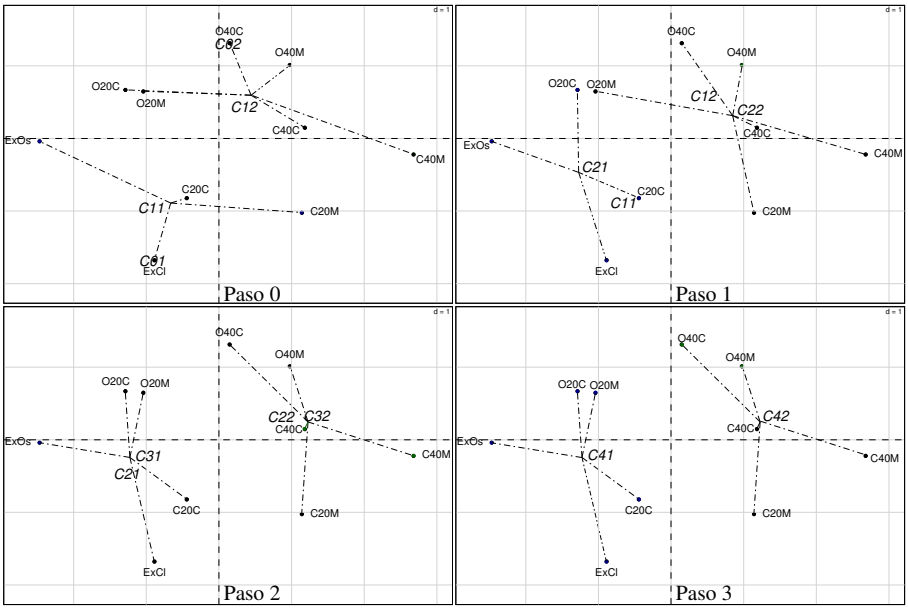


Figura 7.1. Ejemplo de clasificación con *K-means* de los cafés a partir de las coordenadas factoriales sobre los ejes 1 y 2. Los rayos indican la pertenencia a cada clase al unir el centro de gravedad con los puntos. Con los centros iniciales C_1^0 y C_2^0 se construye la partición que se muestra en el paso 0. Los puntos C_1^1 y C_2^1 son los dos centros de gravedad y los puntos iniciales para construir la partición del paso 1. C_1^2 y C_2^2 son los centros de gravedad de la partición del paso 1 y los puntos iniciales del paso 2. En el paso 3 no hay cambios y el proceso termina

Paso 2: Centros de gravedad de P_1: $\begin{cases} C_1^2 = [-1.27, -0.47] \\ C_2^2 = [0.85, 0.31] \end{cases}$

Partición: $P_2 = \begin{cases} I_1^2 = \{ExOs, O20C, O20M, ExCl, C20C\} \\ I_2^2 = \{O40C, O40M, C20M, C40C, C40M\} \end{cases}$

Paso 3: Centros de gravedad de P_2: $\begin{cases} C_1^3 = [-1.23, -0.24] \\ C_2^3 = [1.23, 0.25] \end{cases}$

Como la partición P_3 es igual a la partición P_2, el algoritmo termina. Nótese que los centros de gravedad obtenidos en este paso, son los de las clases de la partición final.

Tabla 7.1. Clasificación "a mano" de los diez cafés con *K-means*: arriba los centros de cada clase en cada paso y abajo las coordenadas de los cafés, las distancias a las clases y la asignación de clase (1 o 2)

Centros

	Paso 0		Paso 1		Paso 2		Paso 3	
Coordenadas	C_1^0	C_2^0	C_1^1	C_2^1	C_1^2	C_2^2	C_1^3	C_2^3
F_1	-0.89	0.15	-0.66	0.44	-1.27	0.85	-1.23	1.23
F_2	-1.68	1.31	-0.89	0.59	-0.47	0.31	-0.24	0.25

Distancias y particiones de pasos 0 y 1

	Coordenadas		Paso 0			Paso 1		
Cafés	F_1	F_2	C_1^0	C_2^0	P_0	C_1^1	C_2^1	P_1
ExOs	-2.47	-0.04	2.28	2.95	1	2.00	2.98	1
O20C	-1.29	0.67	2.38	1.58	2	1.68	1.73	1
O20M	-1.04	0.65	2.33	1.36	2	1.59	1.48	2
ExCl	-0.89	-1.68	0.00	3.17	1	0.82	2.63	1
C20C	-0.44	-0.82	0.97	2.21	1	0.23	1.66	1
O40C	0.15	1.31	3.17	0.00	2	2.34	0.78	2
O40M	0.98	1.01	3.28	0.88	2	2.51	0.68	2
C20M	1.14	-1.02	2.13	2.53	1	1.80	1.76	2
C40C	1.18	0.15	2.76	1.55	2	2.11	0.86	2
C40M	2.68	-0.22	3.86	2.96	2	3.41	2.38	2

Distancias y particiones de pasos 2 y 3

	Coordenadas		Paso 2			Paso 3		
Cafés	F_1	F_2	C_1^2	C_2^2	P_2	C_1^3	C_2^3	P_3
ExOs	-2.47	-0.04	1.27	3.34	1	1.26	3.71	1
O20C	-1.29	0.67	1.14	2.17	1	0.91	2.55	1
O20M	-1.04	0.65	1.14	1.92	1	0.91	2.30	1
ExCl	-0.89	-1.68	1.27	2.64	1	1.48	2.87	1
C20C	-0.44	-0.82	0.90	1.71	1	0.98	1.98	1
O40C	0.15	1.31	2.28	1.22	2	2.08	1.51	2
O40M	0.98	1.01	2.69	0.71	2	2.54	0.80	2
C20M	1.14	-1.02	2.47	1.36	2	2.50	1.27	2
C40C	1.18	0.15	2.53	0.37	2	2.44	0.11	2
C40M	2.68	-0.22	3.96	1.91	2	3.91	1.52	2

Ventajas y desventajas del método $K-means$

El método *K-means* se utiliza ampliamente porque es muy rápido y poco exigente en recursos de cómputo. Sin embargo, tiene dos problemas:

1. Con un método de agrupamiento se pretende descubrir una estructura de clases en los datos y al algoritmo *K-means* se le suministran el número de clases y los puntos iniciales.
2. En general, la inercia mínima que se obtiene depende de los puntos iniciales.

7.2. Métodos de clasificación jerárquica

Son de dos tipos: aglomerativos y divisivos. Los más usados son los primeros, conocidos en la literatura de ciencias naturales, como *métodos de clasificación ascendente jerárquica aglomerativa* (Sokal & Sneath, 1963). Estos métodos construyen una serie de particiones anidadas, empezando por los n individuos, uniendo los dos más cercanos para tener una partición de $n - 1$ clases, calculando la distancia entre el nuevo grupo y los demás individuos, seleccionando de nuevo los dos más cercanos, para conseguir una partición en $n - 2$ clases y continuar aglomerando hasta llegar a una partición de una clase con los n individuos. El proceso de uniones se representa en un árbol de clasificación o dendrograma.

Estos métodos utilizan un índice de similitud, disimilitud o distancia entre individuos. Se dispone de unos cuantos en la literatura dependiendo del tipo de variables y de las aplicaciones. En nuestro contexto se selecciona la distancia euclidiana canónica.

Al conformar grupos se necesita definir una distancia entre ellos, que se denomina *criterio de agregación* y le da nombre a un método específico. Los más sencillos son el de *enlace simple* y *enlace completo*. El primero define la distancia como la que hay entre los dos individuos más cercanos cada uno de diferente grupo y el segundo, entre los dos individuos más lejanos.

Un procedimiento de clasificación jerárquica aglomerativa procede de la siguiente manera:

1. Seleccionar y calcular un índice de disimilitud entre individuos.
2. Seleccionar un criterio de agregación o disimilitud entre grupos.
3. Construir el árbol de clasificación o jerarquía de particiones indexadas:
   a) Buscar el menor valor en $\mathbf{D}$: d_{il}^0: grupo I_{il}^0.
   b) Calcular los índices de disimilitud entre I_{il}^0 y los demás individuos.

c) Eliminar las filas y columnas i y l e incluir la fila y columna I_{il}^0, para colocar las disimilitudes.

d) Volver a 3*a* y repetir hasta tener un solo grupo de n individuos.

7.2.1. Índices de similitud, disimilitud y distancias

Las definiciones de esta sección se han tomado del texto de Jambu (1983). Las *medidas de similitud* evalúan el grado de parecido o proximidad existente entre dos elementos. Los valores más altos indican mayor parecido o proximidad entre los elementos comparados. Un índice de similitud sobre un conjunto E es una aplicación de $E \times E$ que va hacia $\mathbb{R}^+ \cup \{0\}$:

$$\begin{array}{rccc} s: & E \times E & \longrightarrow & \mathbb{R}^+ \cup \{0\} \\ & (i,l) & \longmapsto & s(i,l) \end{array}$$

tal que:

$$\begin{array}{ll} s(i,l) = s(l,i) & \forall\, (i,l) \in E \times E \\ s(i,i) = s(l,l) = s_{max} > s(i,l) & \forall\, i \in E \end{array}$$

Las *medidas de disimilitud* ponen el énfasis en el grado de diferencia o lejanía existente entre dos elementos. Los más altos indican mayor diferencia o lejanía entre los elementos comparados.

Cuando dos elementos coinciden en sus características, la disimilitud es nula. Las medidas de disimilitud son las que han pasado al vocabulario común con la acepción de *medidas de distancia*.

Un índice de disimilitud sobre un conjunto E es una aplicación de $E \times E$ que va hacia $\mathbb{R}^+ \cup \{0\}$:

$$\begin{array}{rccc} d: & E \times E & \longrightarrow & \mathbb{R}^+ \cup \{0\} \\ & (i,l) & \longmapsto & d(i,l) \end{array}$$

tal que:

$$\begin{array}{ll} d(i,l) = d(l,i) & \forall\, (i,l) \in E \times E \\ d(i,i) = 0 & \forall\, i \in E \end{array}$$

A un índice de similitud se le puede asociar un índice de disimilitud mediante la siguiente ecuación:

$$d(i,l) = s_{max} - s(i,l)$$

Teniendo en cuenta las siguientes propiedades se obtienen distintos tipos de índices de disimilitud:

1. $d(i,l) = 0 \rightarrow i = l$
2. $d(i,l) \leq d(i,k) + d(l,k) \qquad \forall i,l,k \in E \qquad (7.5)$
3. $d(i,l) \leq \text{máx}\{d(i,k), d(l,k)\} \forall i,l,k \in E$

Si el índice de disimilitud cumple la propiedad 1, se denomina índice de distancia; si verifica la propiedad 2, se llama desviación; si cumple la 1 y 2, se llama distancia y si cumple las propiedades 1 y 3 (la propiedad 3 implica la 2), se llama *ultramétrica*, que es la distancia asociada a los árboles de clasificación jerárquica.

Teniendo en cuenta las anteriores definiciones se van a mostrar algunos ejemplos de medidas de similitud para tablas binarias y distancias para variables de intervalo.

7.2.2. Índices de similitud para tablas binarias

Se calculan sobre una tabla de n individuos por p atributos de naturaleza binaria. La presencia de un 1 en una celda (i, j) indica que el individuo i tiene el atributo j y la de un cero, que no lo tiene. Entonces, un conteo para comparar los individuos i y l se puede registrar en una tabla de cuatro celdas:

		Individuo l		Suma
		1	0	
Individuo i	1	a	b	$a+b$
	0	c	d	$c+d$
Suma		$a+c$	$b+d$	p

- a: número de atributos presentes en los dos individuos.
- b: número de atributos presentes en el individuo i y ausentes en el individuo l.
- c: número de atributos ausentes en el individuo i y presentes en el individuo l.

- d: número de atributos ausentes en los dos individuos.

Se definen, además:

- $m = a + d$ coincidencias.
- $u = b + c$ no coincidencias.

En la tabla 7.2 se muestran algunos de los índices propuestos en la literatura, que están disponibles en la función `dist.binary{ade4}`. El número de la última columna de la tabla se utiliza para que la función calcule el índice deseado.

Tabla 7.2. Índices de similitud para tablas binarias

Nombre y referencia	Fórmula	`method`[1]
Jaccard (1908)	$S_J(i,l) = \frac{a}{a+b+c}$	1
De coincidencias simple (Sokal & Michener, 1958)	$S_{SM}(i,l) = \frac{a+d}{a+b+c+d} = \frac{m}{p}$	2
Sokal & Sneath (1963)	$S_{ss} = \frac{2a+2d}{2a+b+c+2d} = \frac{2m}{2m+u}$	3
Rogers & Tanimoto (1960)	$S_{RT}(i,l) = \frac{m}{p+u}$	4
Dice (1945)	$S_D(i,l) = \frac{2a}{2a+b+c}$	5
Hamann (1961)	$S_H(i,l) = \frac{m-u}{p}$	6
Ochiai (1957)	$S_o = \frac{a}{\sqrt{(a+b)(a+c)}}$	7
Gowers (Sokal & Sneath, 1963)	$s_8 = \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	8
Pearson (Sokal & Sneath, 1963)	$S_\phi(i,l) = \frac{ad-bc}{(adbc)^{1/2}}$	9
Russell y Rao (Sokal & Sneath, 1963)	$S_{RR}(i,l) = \frac{a}{a+b+c+d} = \frac{a}{p}$	10

[1] Parámetro para seleccionar el índice en la función `dist.binary{ade4}`

7.2.3. Distancias para variables de intervalo

La distancia entre dos individuos i y l se calcula a partir de las filas respectivas de la matriz $\mathbf{X}$, cuyas columnas son p variables cuantitativas. Los individuos están representados como vectores en $\mathbb{R}^p$ en donde se define la distancia como:

$$\begin{array}{rccc} d: & \mathbb{R}^p \times \mathbb{R}^p & \longrightarrow & \mathbb{R}^+ \cup \{0\} \\ & (i,l) & \longmapsto & d(i,l) \end{array}$$

Las distancias disponibles en la función `dist{stats}` se presentan en la tabla 7.3. La distancias euclidiana, de Manhattan, y del máximo son casos particulares de la distancia de Minkowski cuando $r = 2$, $r = 1$ y $r \to \infty$.

Tabla 7.3. Distancias para variables de intervalo

Nombre	Fórmula	`method`[1]
Euclidiana	$d(i,l) = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - x_{lj}\right)^2}$	`euclidian`
Manhattan o Cityblock	$\sum_{j=1}^{p} \lvert x_{ij} - x_{lj} \rvert$	`manhattan`
Minkowski	$\left(\sum_{j=1}^{p} \lvert x_{ij} - x_{lj} \rvert^r \right)^{1/r} ; r \in \mathbb{R}$	`minkowski`
Del máximo o de Chebyschev	$\max_j \{x_{ij} - x_{lj}\}$	`maximum`
Canberra	$\sum_{j=1}^{p} \frac{\lvert x_{ij} - x_{lj} \rvert}{x_{ij} + x_{lj}}$	`canberra`

[1] Parámetro para seleccionar en la función `dist{stats}`.

7.2.4. Criterios de agregación

Para completar un procedimiento de aglomeración jerárquica hay que seleccionar una similitud, disimilitud o distancia entre grupos, que se denomina también *criterio de agregación*. Aquí solo se mencionan tres y en la sección 7.2.7 se introduce el criterio de agregación de Ward. Sean los grupos: A con n_A elementos, y B con n_B elementos.

Enlace simple

La distancia entre dos grupos A y B es igual a la distancia de los dos individuos de diferente grupo más cercanos:

$$d(A, B) = \text{mín}\{d(i, l); i \in A; l \in B\}$$

Este criterio tiende a producir grupos alargados (efecto de encadenamiento), que pueden incluir elementos muy distintos en los extremos.

Enlace completo

La distancia entre los dos grupos es la distancia entre los dos individuos de diferente grupo más alejados:

$$d(A, B) = \text{máx}\{d(i, l); i \in A; l \in B\}$$

El enlace completo tiende a producir grupos esféricos.

Enlace promedio

La distancia entre los dos grupos es el promedio de distancias entre todas las parejas de individuos de diferente grupo:

$$d(A, B) = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{l \in B} d(i, l)$$

En ciencias naturales se conoce como UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*). Tiende a unir grupos con varianzas pequeñas y es intermedio entre los enlaces simple y completo (Everitt *et al.*, 2011).

El procedimiento descrito en la sección 7.2 es posible porque las distancias entre grupos se pueden calcular a partir de la matriz del paso inmediatamente anterior. Sean los grupos: A con n_A elementos, y B con n_B elementos, que se fusionan para crear un grupo AB con $n_{AB} = n_A + n_B$ elementos. Sea otro grupo C con n_C elementos, entonces:

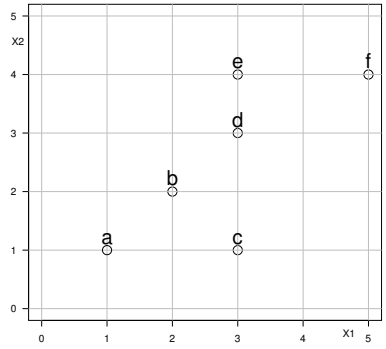
Enlace simple: $d(AB, C) = \text{mín}\{d(A, C), d(B, C)\}$.

Enlace completo: $d(AB, C) = \text{máx}\{d(A, C), d(B, C)\}$.

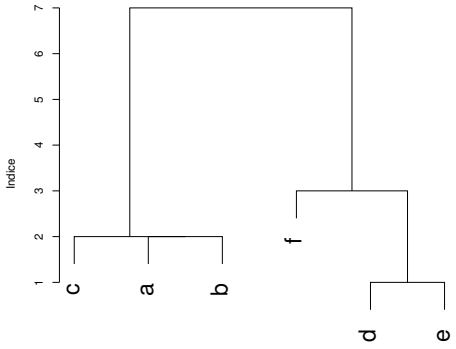
Enlace promedio: $d(AB, C) = \dfrac{n_A d(A, C) + n_B d(B, C)}{n_A + n_B}$.

7.2.5. Ejemplo “de juguete”

Para entender el proceso de construcción de un árbol usaremos un ejemplo de puntos sobre un plano, con la distancia de Manhattan entre puntos y el enlace completo como criterio de agregación, es decir, la distancia entre grupos. En la figura 7.2 se muestra el plano con los puntos y todo el proceso de construcción del árbol. Las distancias de Manhattan se encuentran recorriendo las calles representadas en la grilla de la gráfica, es decir, no hay diagonales. Por ejemplo, la distancia entre a y b es 2, lo mismo que entre a y c y entre b y c.



Coordenadas de los puntos



Árbol con distancia de Manhattan y enlace completo

Distancias de Manhattan y enlace completo:

paso 1

	a	b	c	d	e
b	2				
c	2	2			
d	4	2	2		
e	5	3	3	1	
f	7	5	5	3	2

paso 2

	a	b	c	de
b	2			
c	2	2		
de	5	3	3	
f	7	5	5	3

paso 3

	ab	c	de
c	2		
de	5	3	
f	7	5	3

paso 4

	abc	de
de	5	
f	7	3

paso 5

	abc
def	7

Ultramétrica

	a	b	c	d	e
b	2				
c	2	2			
d	7	7	7		
e	7	7	7	1	
f	7	7	7	3	3

Figura 7.2. Ejemplo “de juego” de una clasificación jerárquica aglomerativa

El proceso de aglomeración es el siguiente:

1. Se unen los puntos d y e a una distancia de 1. Se tiene ahora una partición en 5 clases.

2. Se calculan las distancias entre el grupo de y los demás puntos. La matriz pierde una fila y una columna. La distancia de enlace completo entre dos grupos corresponde a la de los dos puntos más alejados entre sí, uno de cada grupo. Por ejemplo la distancia entre de y a es la de a y e (5), porque e está más alejado de a que d. Se unen a y b a una distancia de 2, y se forma una partición en 4 clases.

3. Se calculan las distancias de enlace completo entre el grupo ab y los demás puntos y grupos. Se une c al grupo ab a una distancia de 2. Ahora la partición es en 3 clases.

4. Se calculan las distancias de enlace completo entre el grupo abc, el grupo de y el punto f. Se une f al grupo de a una distancia de 3. La partición tiene dos clases.

5. La distancia de enlace completo entre los grupos abc y def es de 7. Finalmente se unen estos dos grupos y todos los puntos quedan en una sola clase.

En cada paso del proceso de aglomeración se calculan las nuevas distancias a partir de la matriz del paso anterior, propiedad importante porque no hay que volver más atrás para calcularlas.

Las particiones anidadas que se van construyendo en el proceso de aglomeración quedan registradas en el árbol, partiendo de los puntos hasta llegar a una sola clase con todos los puntos. Se denomina *nodos* a los puntos de unión; e *índices de nivel* a las distancias asociadas, que corresponden a las alturas del árbol.

7.2.6. Ultramétrica asociada a un árbol

Una distancia euclidiana d cumple la desigualdad triangular: sean a, b y c tres puntos en $\mathbb{R}^p$, entonces:

$$d(a,b) \leq d(a,c) + d(b,c) \tag{7.6}$$

Una distancia ultramétrica es más restrictiva y cumple la propiedad:

$$d(a,b) \leq max\{d(a,c), d(b,c)\} \tag{7.7}$$

Si se cumple (7.7) también se cumple (7.6); en la ultramétrica los triángulos son isóceles. Un árbol de clasificación tiene una ultramétrica asociada, definida como la altura mayor del camino que hay que recorrer, en el árbol, para conectar los dos puntos. En la figura 7.2 se muestra un árbol y la ultramétrica asociada a este. La ultramétrica entre a y b es 2 porque hay que subir a esa altura para unirlos. Las ultramétricas entre a y f y entre b y f son iguales a 7.

Se cumple que $d(a,b) \leq max\{d(a,f), d(b,f)\}$, es decir, $2 \leq max\{7,7\}$. También se cumple que $d(a,f) \leq max\{d(a,b), d(b,f)\}$, o sea, $7 \leq max\{2,7\}$.

En taxonomía numérica en ciencias naturales se usa el *coeficiente de correlación cofenética* (Sokal & Rohlf, 1962) para medir la proximidad entre un árbol de clasificación y la matriz de similitudes o disimilitudes utilizada para su construcción. Se define como la correlación entre los valores presentes en la matriz de similitudes, disimilitudes o distancias y los valores correspondientes de las ultramétricas obtenidas en un proceso de clasificación jerárquica aglomerativa.

Se muestra el cálculo del coeficiente de correlación cofenético entre las distancias de Manhattan y las ultramétricas asociadas al árbol de la figura 7.2 obtenido utilizando el enlace completo:

```
dis<-c(D);dis
# [1] 2 2 4 5 7 2 2 3 5 2 3 5 1 3 2
ult <- c(2,2,7,7,7,2,7,7,7,7,7,7,1,3,3)
cor(dis,ult)
# [1] 0.6369261
```

7.2.7. Método de Ward

Para lograr grupos que tengan inercia mínima intraclases se debe utilizar una distancia euclidiana y unir en cada paso del procedimiento los dos grupos que aumenten menos la inercia intraclases, que corresponde al método de Ward (Ward, 1963; Wishart, 1969).

Distancia de Ward entre grupos e individuos

En la figura 7.3 se muestran esquemáticamente tres grupos A, B y C. Si estos grupos están presentes en un proceso de clasificación jerárquica con el método de Ward, hay que tomar la decisión de cuál de las tres parejas de grupos unir. Es decir qué unión es la que causa menos incremento en la inercia intra grupos.

El incremento de inercia al unir A y B es $I_{AB} - I_A - I_B$. A estos incrementos los llamaremos *distancias de Ward entre grupos* y las notaremos W. Entonces, hay que calcular $W(A, B)$, $W(A, C)$ y $W(B, C)$ y la menor de ellas determinará los grupos a unir.

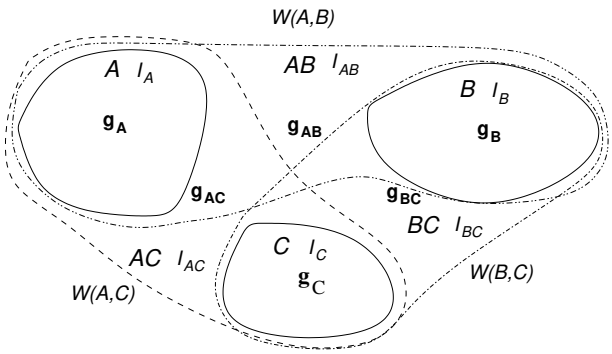


Figura 7.3. Esquema de tres grupos y sus posibles uniones en dos grupos, según el criterio de Ward

Sean A y B dos grupos o clases no vacías y disyuntas y sean p_A, p_B y $\mathbf{g}_A$, $\mathbf{g}_B$, I_A, I_B los pesos, centros de gravedad e inercias de los grupos A y B según el caso.

Al unir los grupos A y B en un grupo AB, este grupo tiene su centro de gravedad $\mathbf{g}_{AB}$ y su inercia intra I_{AB}. La inercia intra de AB es obviamente la suma de las inercias intra de A y B mas la inercia que aparece al considerar un solo centro de gravedad $\mathbf{g}_{AB}$ para los puntos que están en A o en B. Esta inercia es la que hay entre los dos centros de gravedad $\mathbf{g}_A$ y $\mathbf{g}_B$ respecto al centro de gravedad común $\mathbf{g}_{AB}$, es decir:

$$\textit{Inercia entre}(A,B) = p_A d^2(\mathbf{g}_A, \mathbf{g}_{AB}) + p_B d^2(\mathbf{g}_B, \mathbf{g}_{AB}) = p_A \|\mathbf{g}_A - \mathbf{g}_{AB}\|^2 + p_B \|\mathbf{g}_B - \mathbf{g}_{AB}\|^2$$

que es la distancia de Ward entre los grupos A y B.

Reemplazando $\mathbf{g}_{AB} = \frac{1}{p_A + p_B}(p_A\mathbf{g}_A + p_B\mathbf{g}_B)$ en la fórmula anterior, se obtiene:

$$W(A, B) = \frac{p_A p_B}{p_A + p_B} d^2(\mathbf{g}_A, \mathbf{g}_B) \tag{7.8}$$

Este valor es el incremento de la inercia intragrupos al unir los grupos A y B en uno solo. En particular para dos individuos i y l la distancia de Ward es:

$$W(i, l) = \frac{p_i p_l}{p_i + p_l} d^2(i, l) \tag{7.9}$$

Si los pesos son iguales a $1/n$ para los dos individuos, la anterior expresión se reduce a:

$$W(i, l) = \frac{1}{2n} d^2(i, l) \tag{7.10}$$

Fórmula de recurrencia de la distancia de Ward

En los procesos de clasificación jerárquica aglomerativa ascendente, se parte de una matriz de índices de disimilitud o distancias entre todos los individuos, que tiene dimensión $n \times n$. Si se unen los grupos A y B (individuos en los primeros pasos), se eliminan las filas y columnas correspondientes y se inserta una fila y una columna para registrar las distancias entre el grupo AB y los demás, entonces en cada unión la matriz disminuye en una fila y una columna. En los enlaces simple y completo es fácil ver que las distancias entre el grupo conformado en un paso y los demás se pueden calcular a partir de la matriz del paso anterior. En el método de Ward es también posible hacerlo, mediante la fórmula que se presenta a continuación.

Sean A, B y C tres grupos presentes en el mismo paso de construcción del árbol. Si se unen A y B para formar el grupo AB, es necesario calcular la distancia de Ward entre los grupos AB y C. Se conocen las distancias $W(A, B)$; $W(A, C)$ y $W(B, C)$. La distancia $W(AB, C)$ en función de las anteriores es (Pardo, 1992):

$$d(AB, C) = \frac{(p_A + p_C)W(A, C) + (p_B + p_C)W(B, C) - p_C W(A, B)}{p_A + p_B + p_C} \tag{7.11}$$

Una forma de demostrar 7.11 es desarrollando:

$$(p_A + p_B)^2 \|\mathbf{g}_C - \mathbf{g}_{AB}\|^2 = \|p_A(\mathbf{g}_C - \mathbf{g}_A) + p_B(\mathbf{g}_C - \mathbf{g}_B)\|^2$$

Procedimiento del método de Ward

Con los elementos presentados en las subsecciones anteriores, es posible construir un árbol por el método de Ward mediante los pasos siguientes:

1. Calcular la matriz de distancias de Ward entre parejas de individuos con (7.9).
2. Seleccionar la pareja de grupos (individuos en el primer paso) que presente la menor distancia de Ward para conformar el nuevo grupo.
3. Calcular las distancias entre todos los grupos y el grupo recién conformado utilizando la fórmula de distancia de Ward o la fórmula de recurrencia (7.11).
4. Eliminar las filas y columnas correspondientes a los individuos o grupos unidos y adicionar una fila y una columna para registrar las distancias entre el nuevo grupo y los demás.
5. Repetir el proceso hasta llegar a una sola clase.

De inercia entre clases a inercia intraclases y viceversa

Antes de empezar las uniones toda la inercia corresponde a inercia entre clases (cada individuo es una clase), y a medida que se llevan a cabo las uniones, la inercia entre clases va pasando a inercia intraclases, de modo que al terminar, toda la inercia es intraclases (todos los elementos conforman una clase). Por esta razón en el método de Ward la suma de los índices de nivel es igual a la inercia total.

Una vez construido el árbol se puede proceder a cortarlo. Es decir, se parte de una clase, con n individuos, que luego se divide en dos. El incremento de la inercia intra del último paso del procedimiento de clasificación, es la inercia entre las dos clases. Si se corta, de las dos ramas la que se formó a mayor índice, la inercia entre los tres grupos conformados es la suma de los dos últimos índices de nivel. Al continuar los cortes hasta llegar a n clases, cada una de un individuo, toda la inercia intra ha pasado a inercia entre individuos.

Los algoritmos de clasificación jerárquica son robustos: un método para los mismos datos produce los mismos resultados y no requieren de un número de clases preestablecido. Precisamente la mayor utilidad del árbol de clasificación es mostrar la estructura de clases que hay en los datos. El método de Ward implementado en algunos programas estadísticos no tiene en cuenta las propiedades de inercia y no permite la interpretación derivada de ella. Algunos tampoco permiten pesos para los individuos, los cuales son obligatorios cuando se hacen pretratamientos con análisis de correspondencias simples, por ejemplo, o si se desea utilizar factores de expansión muestrales, para que el análisis sea más cercano a la población de donde se ha tomado la muestra.

Algunos de los programas con implementación adecuada del método de Ward son: el de uso libre académico DtmVic (Lebart, 2017) y el paquete FactoClass de R (Pardo & Del-Campo, 2007).

Clasificación con el método de Ward en el ejemplo "Café"

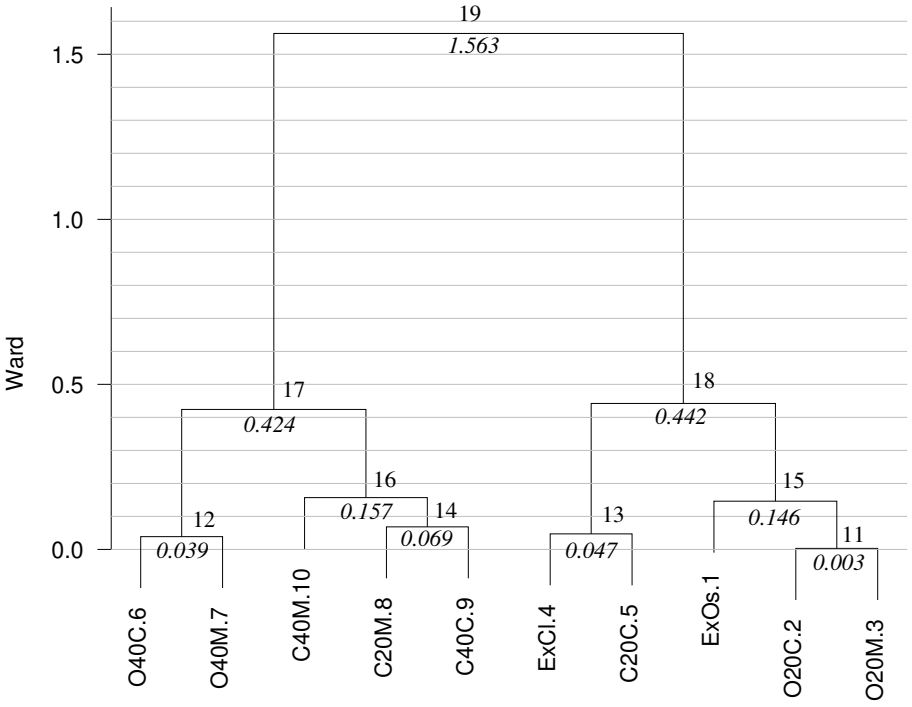
Se retoma el ejemplo de la página 191 para construir el árbol de clasificación con el método de Ward, siguiendo el procedimiento descrito. Sin embargo, no se detallan todos los pasos, sino que se utiliza la matriz inicial de distancias de Ward entre cafés y el árbol de la figura 7.4, para mostrar los aspectos fundamentales del método.

1. Distancia de Ward entre cafés con la fórmula (7.10). Se muestran debajo del árbol en la figura 7.4.

2. Se unen O20C y O20M a una distancia de 0.003, valor que es el incremento de inercia intragrupos al pasar de 10 clases de un café cada una a 9 clases: 8 de un café y 1 de dos cafés. En el árbol cada unión se llama *nodo*, los individuos a clasificar se llaman *hojas* o *nodos terminales*, son los nodos del 1 al 10, y el nodo 11 es la primera unión, lo llamaremos $11 = \{O20C, OC20M\}$.

3. Hay que calcular las distancias entre el grupo 11 y los 8 cafés con la fórmula (7.11).

   El árbol se construyó con la función `ward.cluster{FactoClass}`, que utiliza la función `hclust{stats}`, realizando los cálculos requeridos para que las alturas del árbol sean las distancias de Ward entre grupos, es decir, incrementos de inercia intra al unirlos.




```
> W<-1/20*dist(F)^2;round(W,3)
      ExOs  O20C  O20M  ExCl  C20C  O40C  O40M  C20M  C40C
O20C 0.095
O20M 0.126 0.003
ExCl 0.259 0.284 0.273
C20C 0.236 0.147 0.126 0.047
O40C 0.434 0.124 0.093 0.501 0.244
O40M 0.650 0.263 0.211 0.537 0.268 0.039
C20M 0.700 0.438 0.377 0.228 0.127 0.320 0.207
C40C 0.668 0.319 0.259 0.382 0.178 0.120 0.039 0.069
C40M 1.328 0.828 0.730 0.744 0.505 0.437 0.220 0.151 0.119
      ExOs  O20C  O20M  ExCl  C20C  O40C  O40M  C20M  C40C
```

Figura 7.4. Árbol de clasificación por el método de Ward de los cafés según las coordenadas sobre los dos primeros ejes factoriales. Se muestran los números de los nodos y sus alturas. Abajo se muestra la matriz de distancias de Ward entre cafés

4. Los nodos 12, 13 y 14 son también uniones entre pares de cafés, de modo que las alturas de unión se pueden leer en la matriz de distancias de Ward de la figura 7.4.

5. En el árbol se pueden ver las demás uniones con las distancias de Ward a las que ocurren.

La inercia total de la nube de puntos es 2.889 y corresponde a la retenida en el primer plano factorial del ACP normado de los 10 cafés. Esta inercia antes de la aglomeración es toda la inercia entre los 10 grupos, cada uno de un café, y la inercia intra es cero. En cada nodo se incrementa la inercia intra. Al final toda la inercia es intra del grupo de los 10 cafés. La última unión incrementa la inercia intra en 1.563.

Si se hace el ejercicio al revés —es decir, cortando los nodos del árbol de arriba hacia abajo—, la inercia va pasando de intra a entre. Al cortar en el nodo 19, quedan dos clases, con 1.563 de inercia entre clases y 1.327 de inercia intraclases.

Nodo	11	12	13	14	15	16	17	18	19
Ward	0.003	0.039	0.047	0.069	0.146	0.157	0.424	0.442	1.563
SumaWard	0.003	0.042	0.089	0.158	0.304	0.461	0.885	1.327	2.890

La partición que se obtiene al cortar el árbol, para obtener dos clases, resulta ser la misma de la clasificación realizada con *K-means*, pero esto es un caso particular.

7.3. Combinación de métodos

Desde el punto de vista del análisis de datos los métodos a utilizar son el de Ward de aglomeración jerárquica y el *K-means*, porque buscan grupos que tengan inercia intragrupos lo más baja posible. Estos métodos se complementan para subsanar entre sí las desventajas y aprovechar sus ventajas.

Los métodos de clasificación jerárquica tienen dos desventajas: utilizan mayor recurso de cómputo y las particiones obtenidas quedan anidadas.

El *K-means* tiene también dos problemas: hay que darle el número de clases iniciales y los puntos iniciales. El número de clases es precisamente lo que se quiere descubrir en una tabla de datos y el óptimo es local, es decir, depende de los puntos iniciales.

La estrategia descrita en Lebart *et al.* (2006) y programada en DtmVic y FactoClass, combina los dos métodos, ya que sus ventajas y desventajas

son complementarias. Cuando el número de elementos a clasificar no es tan grande y el equipo de cálculo lo permite, se realiza la clasificación jerárquica aglomerativa con el método de Ward. El "histograma de índices de nivel" permite visualizar las mejores alturas de corte del árbol, y por ende, el número de clases. Luego, se disminuye la inercia intraclases de la partición obtenida utilizando *K-means*, con los centros de gravedad de la partición derivada de cortar el árbol como puntos iniciales.

Si el número de elementos a clasificar es demasiado grande, se utiliza *K-means* para hacer un preagrupamiento en un buen número de clases, pueden ser miles, y luego se realiza la clasificación jerárquica con los centros de gravedad de los preagrupamientos.

7.4. Clasificación a partir de coordenadas

Los métodos factoriales se pueden utilizar para transformar los datos antes de realizar procedimientos de clasificación automática. Una de las salidas de un análisis factorial es una tabla de individuos por coordenadas factoriales. Entonces, las columnas de las tablas de entrada de los métodos de clasificación son de la misma naturaleza: coordenadas factoriales, que se asimilan a variables continuas.

En ese sentido, los métodos factoriales pueden cumplir con dos funciones: la primera, en el caso de análisis de correspondencias, es la transformación de unas variables cualitativas en otras continuas; la segunda es una función de filtro, al considerar que los S primeros ejes factoriales contienen la información y los otros son ruido.

En otras palabras, el ACP y los AC, son métodos de pretratamiento de datos para la clasificación que pueden cumplir con dos funciones: cuantificar las variables cualitativas y reducir la dimensionalidad de los datos.

En la figura 7.5 se muestra un esquema de la combinación, para el caso de la clasificación de individuos descritos por variables cualitativas.

7.4.1. Función de transformación o cuantificación

Un programa de ACP normado recibe los datos originales y los estandariza antes de obtener valores y vectores propios, y coordenadas factoriales de individuos y variables.

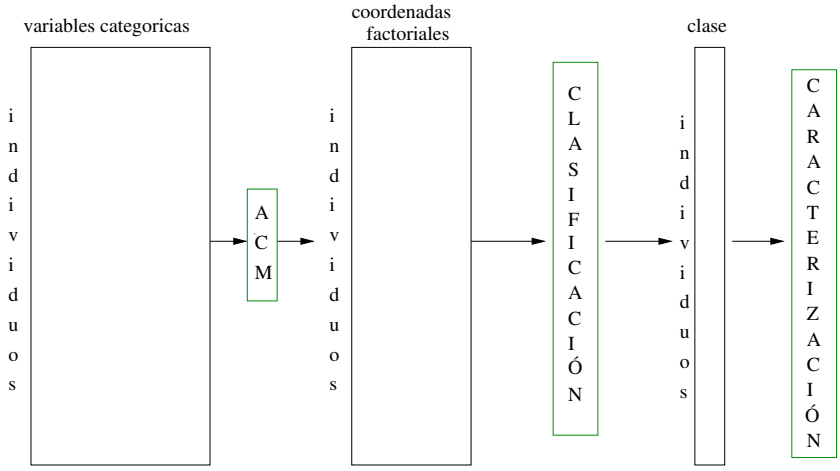


Figura 7.5. Esquema de una estrategia de clasificación con variables cualitativas.

En un análisis de correspondencias las coordenadas factoriales se constituyen en nuevas variables continuas, con las que se puede utilizar la combinación de métodos de la sección 7.3.

7.4.2. Función de filtro

Conectar un método factorial con la clasificación da la posibilidad de seleccionar el número de ejes a utilizar en esta clasificación. En esta decisión se utiliza el histograma de valores propios y otros criterios para la selección del número de ejes, pero haciendo énfasis en el sentido de filtro: aquí seleccionar más ejes puede significar mayor recurso de cómputo, pero no más trabajo para el analista. En problemas pequeños y medianos el recurso de cómputo no tiene importancia. En general, el número de ejes para la clasificación es mayor que el número de ejes seleccionados para analizar en un método factorial. Muchas veces se utilizan todos los ejes para la clasificación, lo que es equivalente a realizar el análisis con las variables originales.

Después del proceso de clasificación se obtiene una partición registrada en una variable cualitativa, que se puede denominar *clase* o *grupo*. Esta variable cualitativa, que emerge de los datos, se puede caracterizar por las variables activas que originaron la estructura de las clases y por variables suplementarias. También se pueden caracterizar por las coordenadas sobre los ejes factoriales.

Las clases se pueden proyectar sobre los planos factoriales como variables suplementarias y se puede visualizar la estructura de clases utilizando colores o símbolos para indicar los "individuos" que pertenecen a cada clase.

7.5. Caracterización automática de las clases

En términos generales, en un procedimiento de clasificación se obtiene una variable cualitativa indicadora de la clase o grupo al que pertenece cada elemento clasificado. Esta variable se puede cruzar con cualquiera de las variables presentes en la tabla de datos correspondiente. Los procedimientos para describir dos variables presentados en el capítulo 2 se pueden utilizar para ese propósito, en particular los ordenamientos por valores test. Las variables continuas o las categorías de las variables cualitativas con diferencias que generan un valor test superior a un umbral, generalmente 2, se dice que caracterizan a la clase. Las que generan valores test inferiores al umbral son de signo negativo (-2, por ejemplo) y se dice que caracterizan negativamente a la clase respectiva.

7.5.1. Descripción con variables continuas

La clase es la variable cualitativa que se desea describir con las variables continuas. Las variables continuas que caracterizan a una clase son aquellas que tienen la media de la clase suficientemente diferente de la media global. Para encontrarlas y ordenarlas se hace la comparación de la media dentro de la clase con la media global, siguiendo el procedimiento de ordenamiento mediante valores test, mostrado en la sección 2.2.2.

En la presentación de los resultados se pueden incluir los *boxplots* de las variables que caracterizan a una o más clases (sección 2.2).

7.5.2. Descripción con variables cualitativas

Una categoría es característica de una clase si su frecuencia dentro de la clase es suficientemente diferente de su frecuencia global. Para encontrarlas y ordenarlas se utiliza el procedimiento de valores test descrito en la sección 2.3.2. Para las variables que tengan categorías que caracterizan a las clases se pueden obtener las gráficas de perfiles de las clases (sección 2.3).

7.6. Una estrategia de clasificación

La estrategia de clasificación que se ha propuesto, desde el punto de vista de la estadística descriptiva multivariada, se resume en los siguientes pasos:

1. Realizar el análisis en ejes principales correspondiente.
2. Seleccionar el número de ejes para la clasificación.
3. Si el número de "individuos" es muy grande, realizar un *K-means* de preagrupamiento en miles de clases.
4. Realizar la clasificación jerárquica con el método de Ward sobre los "individuos" o los grupos del paso anterior.
5. Decidir el número de clases y cortar el árbol.
6. Realizar *K-means* de consolidación partiendo de los centros de gravedad de la partición obtenida al cortar el árbol.
7. Caracterizar las clases.
8. Proyección de las clases sobre los planos factoriales.

En la figura 7.6 se muestra un esquema del procedimiento, el cual está implementado en R en los paquetes `FactoClass`, y en el software de uso libre académico DtmVic, entre otros.

7.7. Ejemplo de aplicación

A continuación se muestra la clasificación de los "Admitidos a las carreras de la Facultad de Ciencias", utilizando `FactoClass`. Este análisis es complementario al realizado en el capítulo 6. Las variables activas son género, edad, estrato y origen, y las variables ilustrativas, los resultados del examen de admisión (continuas) y la carrera. Aunque la función `FactoClass`, se ejecuta primero de forma interactiva, `FactoClass(Y,admi[,1:7])`, aquí se utiliza la función con las decisiones tomadas:

```
fc<-FactoClass(Y,dudi.acm,admi[,1:7],scanFC=FALSE,nf=3,nfcl=6,
      k.clust=8)
```

A continuación se justifican las decisiones y se muestran los resultados.

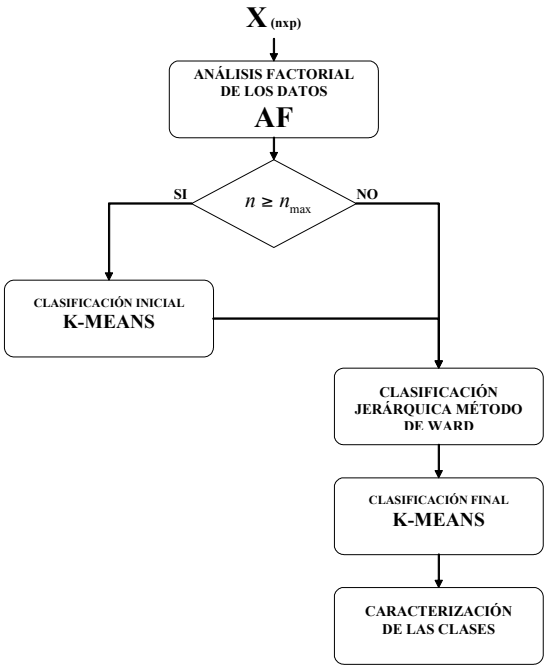


Figura 7.6. Esquema de la estrategia de clasificación

1. Número de ejes para el ACM: en la página 148 se justifica la selección de ejes.

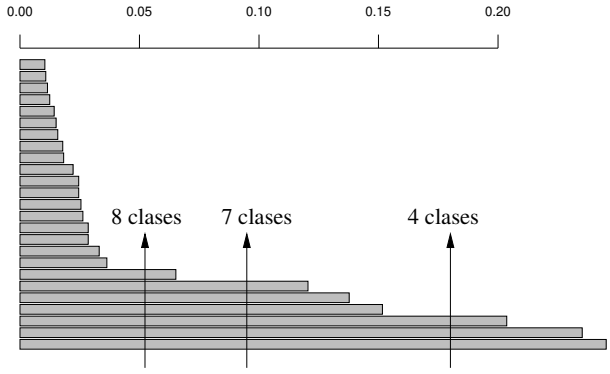
2. Número de ejes para la clasificación: en el histograma de valores propios, figura 6.1, luego de los tres primeros ejes se nota un salto en el eje 6, y se decide utilizar 6 ejes para la clasificación.

3. Número de clases: en la figura 7.7 se observan cambios de inercia entre a intra notorios para justificar el árbol para 4, 7 y 8 clases. Se decidió cortar el árbol para obtener 8 clases, la partición más fina de las tres.

Código para obtener la figura 7.7

```
barplot(fc$indices$Indice[445-25:1],cex.axis=0.6)
# dev.print(device = xfig,
# file="ACMadmiClaHistIndices.fig")
xtable(fc$indices[445-12:1,],digits=c(0,0,0,0,3))
```

Histograma de índices



Últimas 12 uniones

	Nodo	Prim	Benj	Indice
433	878	561	869	0.026
434	879	865	877	0.029
435	880	743	859	0.029
436	881	871	876	0.033
437	882	868	880	0.036
438	883	870	878	0.065
439	884	873	882	0.121
440	885	875	884	0.138
441	886	883	885	0.152
442	887	881	886	0.204
443	888	879	887	0.235
444	889	874	888	0.245

Figura 7.7. Histograma índices de los últimos 25 nodos y últimos 12 nodos

4. Cambios en la consolidación: Las clases 4, 6 y 8 son inestables ya que tuvieron cambios en el proceso de consolidación, la 4 cedió individuos y la 6 y la 8 los recibieron. La inercia intraclases disminuyó de 0.490 a 0.464 (ver tabla 7.4).

Tabla 7.4. Cambios en el proceso de consolidación

Clase	Tamaño		Inercia	
	Antes	Después	Antes	Después
1	68	68	0.100	0.100
2	55	55	0.018	0.018
3	54	54	0.081	0.081
4	77	48	0.107	0.038
5	62	62	0.056	0.056
6	58	66	0.025	0.035
7	38	38	0.082	0.082
8	33	54	0.021	0.054
Total	445	445	0.490	0.464

```
xtable(fc$clus.summ[,1:4],digits=c(0,0,0,3,3))
```

La tercera columna de la tabla 7.4 tiene la cantidad de admitidos en cada clase y los tamaños relativos de las clases son:

```
summary(fc$cluster)->nk
round(nk/sum(nk)*100,1)
   1    2    3    4    5    6    7    8
15.3 12.4 12.1 10.8 13.9 14.8  8.5 12.1
```

5. Caracterización de las clases de admitidos: las clases se describen según las 4 variables cualitativas activas.

   Luego se explora la asociación con la carrera a la que ingresaron, que es ilustrativa en el análisis; y la relación de las clases con los resultados del examen de admisión según las áreas y el puntaje global.

   A continuación se hace un resumen de la descripción de cada una de las clases, incluyendo el número de admitidos y el porcentaje correspondiente.

   Cl 1: 68 (15.3 %). De estrato alto, de 17 o menos años (90 %) y bogotanos (84 %).

Categoria	Vtest	Cl/ca	ca/Cl	Global	n_{cat}
estr.alto	17.4	84.0	100.0	18.2	81
carr.Geol	2.9	31.1	20.6	10.1	45
orig.Bogo	2.9	18.3	83.8	69.9	311
edad.a17	2.5	20.5	51.5	38.4	171
edad.a16m	2.4	22.0	38.2	26.5	118

Esta clase tiene más porcentaje de admitidos a Geología que el global (20.6 vs. 10.1 %), y tienen mejores resultados en el examen que el promedio de los admitidos a la Facultad.

Área	V.test	Clase	Global
exam	6.4	782.4	718.4
cien	5.2	12.2	11.6
soci	5.0	11.8	11.4
mate	3.9	12.3	11.8
text	3.6	11.8	11.4

Cl 2: 55 (12.4 %). De 17 años, estrato medio y bogotanos.

Esta clase escoge Química en proporción menor al promedio (3.6 % vs. 14.2 %).

Categoria	Vtest	Cl/ca	ca/Cl	Global	n_{cat}
edad.a17	10.6	32.2	100.0	38.4	171
estr.medio	10.1	29.7	100.0	41.6	185
orig.Bogo	6.2	17.7	100.0	69.9	311
carr.Quim	-2.6	3.2	3.6	14.2	63

Cl 3: 54 (12.1 %). De 18 años, casi todos bogotanos.

Aumenta la proporción de admitidos a Farmacia (25.9 % vs. 16.4 %).

Categoria	Vtest	Cl/ca	ca/Cl	Global	n_{cat}
edad.a18	17.4	96.4	100.0	12.6	56
orig.Bogo	3.1	15.1	87.0	69.9	311
carr.Farm	2.1	19.2	25.9	16.4	73

Cl 4: 48 (10.8 %). Casi todos son de 16 años o menos (87.5 %), estrato medio (89.6 %), mujeres (70.8 %) y bogotanos (91.7 %).

Disminuye el porcentaje de admitidos a Física (6.2 % vs. 18.4 %).

Categoria	Vtest	Cl/ca	ca/Cl	Global	n_{cat}
edad.a16m	9.4	35.6	87.5	26.5	118
estr.medio	7.3	23.2	89.6	41.6	185
gene.F	6.4	26.6	70.8	28.8	128
orig.Bogo	3.8	14.1	91.7	69.9	311

Cl 5: 62 (13.9 %). De 17 años, con incremento en la proporción de estrato bajo (77.4 %), y de otro departamento (48.4 %).

Categoria	Vtest	Cl/ca	ca/Cl	Global	n_{cat}
edad.a17	11.4	36.3	100.0	38.4	171
estr.bajo	6.4	26.8	77.4	40.2	179
orig.Otro	5.1	31.2	48.4	21.6	96

El resultado en la componente textual es en promedio inferior al global:

Área	V.test	Clase	Global
text	-2.1	11.1	11.4

Cl 6: 66 (14.8 %). De 19 años, bogotanos (98.5 %) y hombres (89.4 %), se incrementa el porcentaje de estrato bajo (51.5 %). Una proporción mayor que el promedio prefiere Matemáticas (19.7 % vs 11.9 %).

Categoria	Vtest	Cl/ca	ca/Cl	Global	n_{cat}
edad.a19M	15.4	66.0	100.0	22.5	100
orig.Bogo	6.4	20.9	98.5	69.9	311
gene.M	3.8	18.6	89.4	71.2	317
estr.bajo	2.2	19.0	51.5	40.2	179
carr.Mate	2.1	24.5	19.7	11.9	53

Cl 7: 38 (8.5 %). Provienen de Cundinamarca, se incrementa la proporción de estrato bajo (57.9 %).

Categoria	Vtest	Cl/ca	ca/Cl	Global	n_{cat}
orig.Cund	15.8	100.0	100.0	8.5	38
estr.bajo	2.4	12.3	57.9	40.2	179

Cl 8: 54 (12.1 %). Son, sobre todo, de otros departamentos (79.6 %), estrato bajo (92.6 %) y de 16 años o menos (75.9 %).

Categoria	Vtest	Cl/ca	ca/Cl	Global	n_{cat}
orig.Otro	9.9	44.8	79.6	21.6	96
estr.bajo	8.6	27.9	92.6	40.2	179
edad.a16m	8.1	34.7	75.9	26.5	118

En promedio tiene resultados inferiores al global:

Área	V.test	Clase	Global
text	-2.2	11.1	11.4
imag	-2.4	11.0	11.3
mate	-2.6	11.4	11.8
exam	-4.2	670.2	718.4
soci	-4.6	10.8	11.4

6. Proyección de las clases en los planos factoriales: en la figura 7.8 se muestran las 8 clases sobre el primer plano factorial del ACM de los admitidos.

   El vector que identifica a qué clase pertenece cada admitido es una variable nominal, que emerge de los datos, y como tal, se proyecta como ilustrativa.

   Esto permite aprovechar el plano para la caracterización de las clases y compararlas.

   Por ejemplo, las clases 1 y 4 se proyectan cerca en el plano: al observar la descripción de las clases ambas clases tienen más porcentaje, que el promedio, de bogotanos y menos de cundinamarqueses. Las clases 2 y 3 se parecen por tener más bogotanos y por no tener admitidos de Cundinamarca.

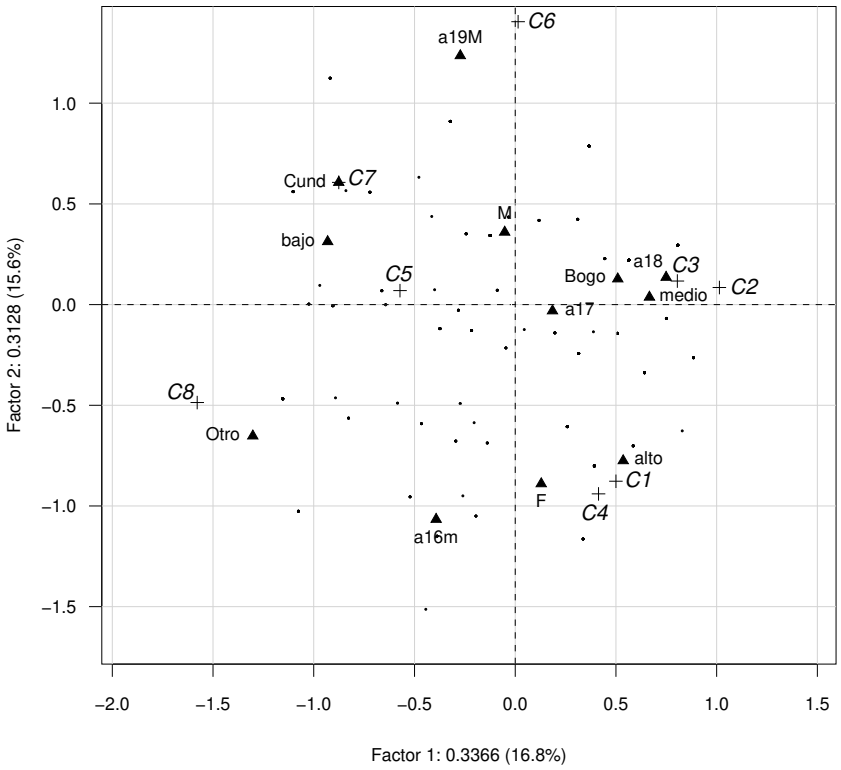


Figura 7.8. Proyección de las clases sobre el primer plano factorial del ACM de admitidos. Sus respectivos puntos se identifican con "+"

Conclusiones del análisis

Las doce categorías de las cuatro variables dan 72 posibilidades de combinación. El proceso de clasificación permite resumirlas en una variable con ocho categorías: las clases, así:

1. 68 (15.3 %). De estrato alto, de 17 o menos años (90 %) y bogotanos (84 %).
2. 55 (12.4 %). De 17 años, estrato medio y bogotanos.
3. 54 (12.1 %). De 18 años, casi todos bogotanos.
4. 48 (10.8 %). Casi todos son de 16 años o menos (87.5 %), estrato medio (89.6 %), mujeres (70.8 %) y bogotanos (91.7 %).
5. 62 (13.9 %). De 17 años, con incremento en la proporción de estrato bajo (77.4 %), y de otro departamento (48.4 %).

6. 66 (14.8 %). De 19 años, bogotanos (98.5 %) y hombres (89.4 %), se incrementa el porcentaje de estrato bajo (51.5 %).
7. 38 (8.5 %). Provienen de Cundinamarca, se incrementa la proporción de estrato bajo (57.9 %).
8. 54 (12.1 %). Son, sobre todo, de otros departamentos (79.6 %), estrato bajo (92.6 %) y de 16 años o menos (75.9 %).

En el primer plano factorial del ACM previo (figura 7.8) se proyecta la variable clase como suplementaria, lo que permite apreciar las características de cada clase según las doce categorías de las variables sociodemográficas, y compararlas entre sí.

7.8. Ejercicios

1. Demuestre la descomposición de la inercia (fórmula 7.1).
2. Muestre que en la distancia de Manhattan los puntos se unen mediante lineas paralelas a los ejes.
3. Verifique la propiedad 3) de una ultramétrica en la matriz de la figura 7.2.
4. Escriba la ultramétrica asociada al árbol del ejemplo "Café" (figura 7.4).
5. Calcule el coeficiente de correlación cofenética entre la ultramétrica y las distancias de Ward del ejemplo "Café".
6. Demuestre que en una ultramétrica los triángulos son isósceles.
7. Demuestre que la propiedad 3) de una ultramétrica implica la propiedad 2): desigualdad triangular.
8. Deduzca la distancia de Ward entre dos grupos A y B (fórmula 7.8).
9. Demuestre la fórmula de recurrencia de Ward (7.11).
10. Muestre que en el método de Ward la inercia es igual a la suma de índices de nivel.
11. ¿Cuáles son las ventajas y desventajas del método *K-means*?
12. ¿Cuáles son las ventajas y desventajas de una clasificación jerárquica aglomerativa?

7.9. Talleres

7.9.1. Clasificación de razas de perros

El objetivo que se busca al realizar el ACM del ejemplo "Razas de Perros", presentado en la sección 7.9.1, se cumple mejor complementando el ACM con la clasificación automática.

Realice una clasificación de las razas de perros utilizando todas las coordenadas factoriales del ACM previo. Obtenga los resultados con el programa estadístico que desee.

Desarrolle los siguientes puntos:

1. Numere los nodos del árbol con los números de la descripción de los nodos (histograma de índices).
2. Describa las tres primeras uniones en la clasificación jerárquica.
3. Justifique la selección de cuatro clases o cambie la decisión.
4. Para la partición en cuatro clases deduzca la inercia entre clases a partir de los índices de nivel.
5. A partir del árbol determine las razas de cada clase.
6. Describa el proceso de consolidación.
7. ¿Qué porcentaje de inercia explica la clasificación?
8. Resuma las características de cada una de las clases.
9. Comente el primer plano factorial del ACM incluyendo las clases obtenidas (centros de gravedad y distinción de las razas de cada clase).
10. Haga un resumen del análisis que responda a los objetivos del ejercicio.

7.9.2. Clasificación de las localidades de Bogotá

En la sección 5.6.1 se realiza el ACS de la TC *localidades*×*estratos*, el objetivo del análisis se cumple mejor al complementar el ACS con la clasificación.

Realice la clasificación de las diecinueve localidades de Bogotá según la distribución de sus manzanas en los seis estratos, utilizando las cinco coordenadas del ACS previo y resuelva los puntos siguientes:

1. La inercia total que entra al procedimiento de clasificación es: _____
2. Trace una linea vertical en el histograma de índices de nivel que indique el corte en cinco clases.
3. La primera unión corresponde a las localidades de _____ y _____
4. Al unirse las localidades de Barrios Unidos (BUni) y Teusaquillo (Teus) el aumento de la inercia intra es _____
5. La inercia intraclases de la partición obtenida es: _____
6. El porcentaje de inercia explicado por la clasificación es: _____
7. La clase 5 está conformada por las localidades:

   _____
8. El perfil de la clase 5 es:

   _____
9. La distribución relativa del estrato 6 en las clases es:

   _____.
10. Ubique los nodos 20 y 21 en el dendrograma.
11. Ubique los nodos 34 a 36 en el dendrograma.
12. Para la partición en cinco clases, se puede obtener la inercia entre clases a partir de los índices de nivel, sumando los valores:

    _____
13. A partir del árbol determine las localidades de cada clase.
14. Las localidades que pertenecen a la clase 2 son:

    _____
15. ¿ Hay cambios en el proceso de consolidación? (Sí/No): _____
16. Para cada clase escriba el estrato más asociado:

    _____
17. De las manzanas de estrato 4 el _____ % pertenecen a la clase 3.

18. Para cada estrato escriba la clase más asociada:

    _____

19. La clase 1 tiene el _____ de sus manzanas en estrato 5.

20. Escriba los porcentajes de manzanas que hay en cada clase:

    _____

21. La clase más grande es la _____ con el _____ % de las manzanas.

7.9.3. Clasificación de adjetivos por colores

El objetivo de encontrar los adjetivos más asociados a cada color se cumple mejor realizando la clasificación de los perfiles de colores derivados de la TC adjetivos × colores.

Realice la clasificación de los adjetivos utilizando todos los ejes factoriales del ACS previo realizado en el taller de la sección 5.6.2, y responda a las siguientes preguntas:

1. ¿Qué adjetivos se unen primero?

2. Relate las últimas cinco uniones en el proceso de clasificación jerárquica: indique los grupos que se unen en cada nodo y el aumento de la inercia intraclases.

3. ¿Cuánta es la inercia entre clases para una partición en dos clases, usando el método de Ward?

4. De acuerdo con el objetivo del ejercicio, ¿cuántas clases selecciona? ¿Por qué?

5. ¿Cambió el coeficiente de inercia entre / inercia total en el proceso de consolidación? ¿Cuánto?

6. Escriba el valor del coeficiente inercia entre / inercia total después de la consolidación.

7. ¿Qué colores son más frecuentes en cada clase?

8. Construya una tabla de contingencia clases × colores y haga una gráfica de los perfiles fila.

9. Escriba la conclusión del análisis: ¿Qué adjetivos se asocian más a cada color?

10. Produzca y analice los planos factoriales colocando la nueva variable cualitativa *clase* como ilustrativa.

Apéndice

A

La libreria *FactoClass* en R

$n \geq n_{\max}$

En el paquete `FactoClass` de R se implementa la estrategia de clasificación, descrita en este texto, que combina métodos factoriales y de clasificación automática. Se utilizan funciones de `ade4` (Dray & Dufour, 2007), para realizar el análisis factorial de los datos, y de `stats`, para la clasificación no supervisada. Se crean funciones para algunas tareas específicas.

La función `FactoClass` realiza la estrategia completa de clasificación, sin incluir el plano factorial con las clases, ni las salidas en formato LaTeX. Sin embargo, algunas funciones se pueden utilizar separadamente: por ejemplo, la función `cluster.carac`, para caracterizar la partición asociada a una variable cualitativa con variables continuas y nominales presentes en una tabla de datos.

Las funciones del paquete `FactoClass` se presentan en la tabla A.2 y los datos de ejemplos incluidos se relacionan en la tabla A.1. No es necesario extenderse más en la descripción del paquete, ya que luego de instalado y cargado, se puede obtener la ayuda para todas las funciones utilizando ? y el nombre de la función: por ejemplo, `?cluster.carac`.

La versión 1.3.0 es la última de `FactoClas` al editar este texto. Como todo paquete de R, `FactoClass` tendrá un mantenimiento y una evolución que el usuario deberá tener en cuenta. Se solicita, a los lectores de este texto, que informen, al correo electrónico cepardot@unal.edu.co, sobre los *bugs* que encuentren, para corregirlos. La sugerencias para mejorar el paquete también son bienvenidas.

Tabla A.1. Tablas de datos del paquete *FactoClass*

Tabla	Descripción
`admi`	Admitidos a la Facultad de Ciencias 2013-1
Bogota	Localidades por estrato en la ciudad de Bogotá
`cafe`	Degustación y características físicas y químicas de tazas de café
ColorAdjective	Asociaciones entre colores y adjetivos
DogBreeds	Razas de perros
`icfes08`	Tabla de contingencia de departamentos por rendimiento y jornada de colegios
`ninios8a12`	Consumo cultural se niños de 8 a 12 años en Colombia
`Whisky`	Treinta y cinco marcar de Whisky
`Vietnam`	Opinión de los estudiantes estadounidenses sobre la guerra de Vietnam

Tabla A.2. Funciones del paquete *FactoClass*

Función	Descripción
`addgrids3d`	Agrega grillas a una gráfica producida por `scatterplot3d`
`centroids`	Centros de gravedad de clases de una partición
`chisq.carac`	Pruebas chi-cuadrado de una variable cualitativa por varias variables cualitativas
`cluster.carac`	Caracterización de las clases según variables
`dudi.tex`	Tabla de coordenadas y ayudas para la interpretación de los ejes principales de un objeto `dudi` en formato LaTeXutilizando `xtable` (Dahl 2006).
`Fac.Num`	División de un objeto `data.frame` en variables cualitativas y cuantitativas
`FactoClass`	Combinación de métodos factoriales y de clasificación no supervisada
`FactoClass.tex`	Tabla de coordenadas, ayudas para la interpretación de los ejes principales y métodos de clasificación en LaTeX
`ggclass`	Plano factorial de filas mostrando clases y columnas
`kmeansW`	Realiza *K-means* teniendo en cuenta las ponderaciones de los individuos
`list.to.data`	Convierta un objeto `list` a `data.frame`
`plot.dudi`	Gráfico de planos factoriales a partir de un objeto `dudi{ade4}`
`plotcc`	Círculos de correlaciones con `ggplot2`
`plotct`	Gráfica de perfiles fila y columna de una tabla de contingencia.
`plotFactoClass`	Gráfico de planos factoriales con clasificación para objetos `FactoClass`
`plotfp`	Planos factoriales a partir de las coordenadas
`plotpairs`	Modificación de `pairs`
`stableclus`	Grupos estables de varios *K-means* a partir de coordenadas factoriales
`supqual`	Coordenadas y ayudas a la interpretación de variables cualitativas ilustrativas.
`ward.cluster`	Clasificación jerárquica con el método de Ward

Referencias

Agresti, A. (2002). *Categorical Data Analysis*, 2 edn, Wiley, Hoboken.

Ball, G. & Hall, D. (1965). Isodata: A novel method of data analysis and pattern classification, *Technical report*, Stanford Research Institute, Menlo Park.

Benzécri, J. P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [bin. mult.], *Les cahiers de l'analyse des données* **4**(3): 377–378.
URL: http://www.numdam.org/article/CAD_1979__4_3_377_0.pdf

Briones, G. (1996). *Metodología de la investigación cuantitativa en las ciencias sociales*, Vol. Modulo 3 of *Especialización en teoría, métodos y técnicas de investigación social*, Instituto Colombiano para el Fomento de la Educación Superior, ICFES.
URL: https://metodoinvestigacion.files.wordpress.com/2008/02/metodologia-de-la-investigacion-guillermo-briones.pdf

Canavos, G. (1988). *Probabilidad y Estadística. Aplicaciones y métodos*, McGraw-Hill.

Correa, E., De Rosa, C. & Lesino, G. (2006). Monitoreo de clima urbano. Análisis estadístico de los factores que determinan la isla de calor y su aporte al diseño de los espacios urbanos., *Avances en Energías Renovables y Medio Ambiente* **10**: 41–48.

Dahl, D. B. (2016). *xtable: Export tables to LaTeX or HTML*.
URL: http://CRAN.R-project.org/package=xtable

Dalgaard, P. (2008). *Introductory statistics with R*, Springer Science & Business Media.

Dalgaard, P. (2020). *ISwR: Introductory Statistics with R*.
URL: http://CRAN.R-project.org/package=ISwR

Dane (2014). *COLOMBIA - Encuesta de Consumo Cultural - ECC - 2014*, Departamento Administrativo Nacional de Estadística, Bogotá, Colombia.
URL: http://microdatos.dane.gov.co/index.php/catalog/345/get_microdata

Díaz, L. G. (2007). *Estadística multivariada: inferencia y métodos*, 2 edn, Universidad Nacional de Colombia. Facultad de Ciencias., Bogotá.

Dice, L. (1945). Measures of the amount of ecologic association between species, *Ecology* **26**: 297–302. Citado por Sneath and Sokal (1973, p.131).

Dray, S. & Dufour, A. (2007). The ade4 package: implementing the duality diagram for ecologists, *Journal of Statistical Software* **22**(4): 1–20.

Duarte, R., Suarez, M., Moreno, E. & Ortiz, P. (1996). Análisis multivariado por componentes principales, de cafés tostados y molidos adulterados con cereales, *Cenicafé* **47**(2): 65–76.

Escofier, B. & Pagès, J. (1992). *Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación*, Universidad del Pais Vasco, Bilbao.

Everitt, B. S., Landau, S., Leese, M. & Stahl, D. (2011). *Cluster Analysis*, 5 edn, Wiley, London.

Fine, J. (1996). Iniciación a los análisis de datos multidimensionales a partir de ejemplos, *Folleto*, PRESTA: Programme de recherche et a'enseignement en statistique appliquée, Sao Carlos.

González, A. & González, S. (2000). *Introducción a R. Notas sobre R: Un entorno de programación para Análisis de Datos*, version 1.0.1 (2000-05-16) edn.
URL: https://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf

Hamann, U. (1961). Merkmalsbestand und verwandtschafsbeziehungen der farinosae, *Willdenowia* **2**: 639–768. Citado por Sneath and Sokal (1973).

Hardle, W. & Simar, L. (2007). *Applied Multivariate Statistical Analysis*, Springer, Berlin.

Hernández, R., Fernández, C. & Baptista, P. (2006). *Metodología de la investigación*, 4 edn, McGraw-Hill, México.

Hidalgo, P., Manzur, E., Olavarrieta, S. & Farías, P. C. (2007). Cuantificación de las distancias culturales entre países: un análisis de Latinoamérica, *Cuadernos de Administración* **20**(33): 252–272.

Holmes, S. (2008). Multivariate data analysis: the French way, *Probability and statistics: Essays in honor of David A. Freedman*, Institute of Mathematical Statistics, pp. 219–233.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale, *Bull. Soc. Vaud. Sci. Nat.* **44**: 223–270. Citado por Sneath and Sokal (1973, p.131).

Jambu, M. (1983). *Cluster Analysis and Data Analysis*, North-Holland, Amsterdam.

Kassambara, A. & Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*.
URL: https://CRAN.R-project.org/package=factoextra

Lebart, L. (2017). DtmVic: Data and Text Mining - Visualization, Inference, Classification. Exploratory statistical processing of complex data sets comprising both numerical and textual data., Web.
URL: http://www.dtmvic.com/

Lebart, L., Morineau, A. & Piron, M. (1995). *Statisitique exploratoire multidimensionnelle*, Dunod, Paris.

Lebart, L., Piron, M. & Morineau, A. (2006). *Statistique exploratoire multidimensionnelle. Visualisation et inférence en fouilles de données*, 4 edn, Dunod, Paris.

Leisch, F. & R-core (2020). *Sweave User Manual*.
URL: https://stat.ethz.ch/R-manual/R-devel/library/utils/doc/Sweave.pdf

Ligges, U. & Mächler, M. (2003). Scatterplot3d - an R Package for Visualizing Multivariate Data, *Journal of Statistical Software* **8**(11): 1–20.
URL: https://www.jstatsoft.org/article/view/v008i11

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations, *Technical report*, University of California, Los Angeles.

Morales, M. A. (2006). *Generación automática de reportes con R y LaTeX*, Universidad de Córdoba. Departamento de Matemáticas y Estadística, Monteria, Colombia.
URL: https://cran.r-project.org/doc/contrib/Rivera-Tutorial_Sweave.pdf

Morineau, A. (1984). Note sur la caractérisation statistique d'une classe et les valeurs-tests, *Bulletin Technique du Centre de Statistique et d'Informatique Appliquées* **2**(1-2): 20–27.

Morrison, D. (1990). *Multivariate Statistical Methods*, McGraw-Hill series in Probability and Statistics, McGraw-Hill.

Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in japan and its neighbouring regions, *Bull. Jpn. Soc . Sci. Fish* **22**: 526–530. Citado por Sokal and Sneath (1963, p.130).

Pardo, C. E. (1992). *Análisis de la aplicación del método de Ward de clasificación jerarquica al caso de variables cualitativas*, Tesis Magister Scientiae en Estadística, Universidad Nacional de Colombia. Facultad de Ciencias. Departamento de Matemáticas y Estadística, Bogotá.

Pardo, C. E., Bécue-Bertaut, M. & Ortiz, J. E. (2013). Correspondence Analysis of Contingency Tables with Sub-partitions on Rows and Columns, *Revista Colombiana de Estadística* **36**(1): 115–144.
URL: https://bit.ly/33IDJgV

Pardo, C. E. & Del-Campo, P. C. (2007). Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass, *Revista Colombiana de Estadística* **30**(2): 231–245.
URL: https://bit.ly/3iGnnMa

R Core Team (2020). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: https://www.R-project.org/

Rogers, D. & Tanimoto, T. (1960). A computer program for classifying plants, *Systematics Biology.* **132**: 1115–1118.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*, RStudio, Inc., Boston, MA.
URL: http://www.rstudio.com/

Sato, T. & Smith, B. (2018). *Xfig User Manual.*
URL: http://mcj.sourceforge.net/frm_authors.html

Slowikowski, K. (2020). *ggrepel: Repulsive Text and Label Geoms for 'ggplot2'.*
URL: https://CRAN.R-project.org/package=ggrepel

Sneath, P. H. & Sokal, R. R. (1973). *Numerical taxonomy: The Principles and Practice of Numerical Classification*, W. H. Freeman, San Francisco.

Sokal, R. R. & Michener, C. D. (1958). A statistical method for evaluating systematic relationships, *University of Kansas Scientific Bulletin* **28**: 1409–1438. Citado por Sokal and Sneath (1963,p.129).

Sokal, R. R. & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods, *Taxon* **11**: 33–40.

Sokal, R. R. & Sneath, P. H. (1963). *Principles of numerical Taxonomy*, W. H. Freeman, San Francisco.

Tenenhaus, M. & Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, homogeneity analysis and otrer methods for quantifying categorial multivariate data, *Psychometrika* **50**(1): 91–119.

The-LaTeX-Project-Team (2019). *LaTeX – A document preparation system*. URL: https://www.latex-project.org/

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley.

Venables, W. N., Smith, D. M. & R Development Core Team (2017). *An introduction to R*, version 3.4.1 (2017-06-30) edn, R Core Team.

DAPD (1997). *Población estratificación y aspectos socioeconómicos de Santa Fe de Bogotá*, Departamento Administrativo de Planeación Distrital.

Wand, M. (2020). *KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995)*. URL: https://CRAN.R-project.org/package=KernSmooth

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association* **58**(301): 236–244.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York. URL: http://ggplot2.org

Wilkinson, L. (2006). *The grammar of graphics*, Springer Science & Business Media.

Wishart, D. (1969). An algorithm for hierarchical classifications, *Biometrics* pp. 165–170.

Índice analítico