Estadística descriptiva multivariada

Campo Elías Pardo

Departamento de Estadística
Facultad de Ciencias
Universidad Nacional de Colombia
Sede Bogotá

Índice general

Índice de tablas

Índice de figuras

Introducción

Este texto se ha escrito con dos objetivos: servir de guía para el curso de *Estadística descriptiva multivariada* de la Carrera de Estadística de la Universidad Nacional de Colombia; y como consulta para los profesionales de distintas áreas interesados en abordar la descripción de sus tablas de datos, teniendo en cuenta las relaciones entre varias variables de manera simultánea.

El objeto de entrada a los métodos estadísticos, que se abordan en estas notas, 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 en base a encuestas son un ejemplo. Otras veces, los datos provienen de sistemas de información administrativos o de transacciones (bancarias, de servicios públicos, supermercados, etc); pero requieren un proceso metodológico de selección, depuración, concatenación y transformación, casi siempre de búsqueda de nuevos datos, para llegar a la tabla objetivo de un análisis.

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: cualitativas (de escala nominal u ordinal), y continuas (de intervalo o de razón).

Nos situamos en el caso en que todos o algunos de los objetivos del estudio se cumplen realizando análisis descriptivos y exploratorios 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 varias variables, que en los métodos básicos deben ser del mismo tipo: continuas o cualitativas.

Las descripciones multivariadas, que recurren a las gráficas para comprender los datos, son mucho más difíciles que las univariadas porque requieren, para su interpretación correcta, del conocimiento de los procedimientos y conceptos para su construcción. Los usuarios de diferentes áreas del conocimiento requieren, por lo menos, de una comprensión intuitiva de los métodos, para 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: espacios vectoriales en los reales con producto interno, del Álgebra Lineal.

Los métodos descriptivos y exploratorios multivariados básicos buscan 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 estas notas se muestran los principales métodos en ejes principales como aplicación de la geometría Euclidiana y del Álgebra Lineal, su lenguaje matemático. La simbología que se adopta es la usual en muchos textos y es la siguiente: las letras mayúsculas en negrilla hacen referencia a matrices ($\mathbf{A}$), la minúsculas en negrilla a vectores ($\mathbf{a}$), 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 al conjunto y a su cardinalidad (número de elementos).

El primer capítulo, denominado preliminares, se ocupa de mostrar el panorama de los métodos abordados en el curso, el lenguaje estadístico R como "calculadora" gráfica y de

Álgebra Lineal y los principales paquetes de R a utilizar en el curso.

El lector de estas notas debe instalar el R (R Core Team 2014) en su computador, leer el manual *An Introduction to R* (Venables, Smith & R Development Core Team 2015), disponible en la consola de R, una vez instalado. Instalar el paquete, programado en R, `FactoClass` (Pardo & Del-Campo 2007), que complementa estas notas: tiene la mayoría de los datos que aquí se utilizan, 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 2014), para exportar tablas a LaTeX en el entorno *tabular*. Estas notas están editadas en LaTeX (The-LaTeX-Project-Team 2017).

Para la edición de las gráficas se utiliza el programa de uso libre *xfig* (Sutanthavibul et al. 2016), ya que R permite exportar a ese formato y, a su vez, *xfig* exporta a los formatos de gráficas más conocidos. Los planos factoriales necesitan, casi siempre, edición posterior para destapar etiquetas que quedan superpuestas, en estas notas esa tarea se hace con *xfig*. Como complemento y referencia para la ejecución de los métodos se utiliza el programa *DtmVic* (Lebart 2015), de uso libre académico.

Este texto consta de seis capítulos y está pensado para trabajar a razón de un capítulo por semana y dejando 4 semanas para las presentaciones de los trabajos del curso y las revisiones del aprendizaje mediante exámenes.

Capítulo 1

Preliminares

Se hace una presentación de los métodos abordados en el texto, del Lenguaje Estadístico de R (R Core Team 2014) y de otros elementos necesarios para abordar el estudio de la lógica de los métodos y su aplicación. También se hace una presentación de la descripción bivariada, que se utiliza luego el algunos aspectos de los métodos multivariados abordados.

1.1. Introducción a los métodos

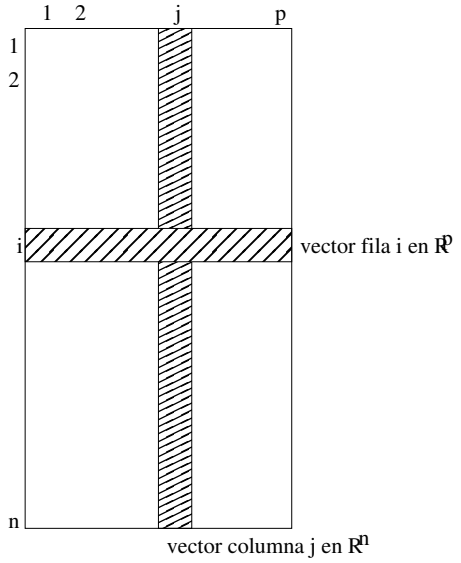
Los métodos básicos de la estadística descriptiva multivariada son de dos tipos: factoriales y de clasificación. En ambos casos se hace una representación geométrica de las tablas de datos, transformadas según el método. La representación se hace utilizando la geometría Euclidiana multidimensional, que se apoya en los conceptos de Álgebra Lineal, necesarios, para definir espacios vectoriales en $\mathbb{R}^n$ con producto interno.

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 $\mathbb{R}^n$.

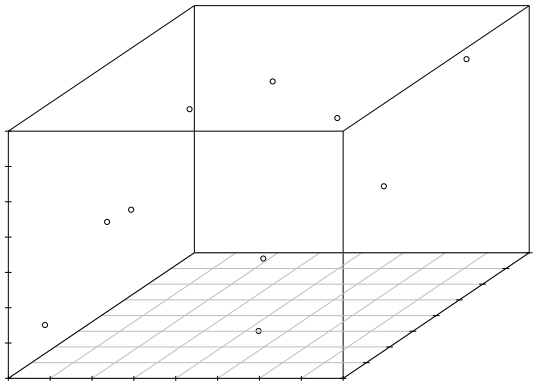
En el primer caso, la representación geométrica es un conjunto de n puntos en un sistema de p ejes ortogonales, cada eje asociado a una variable, y las coordenadas de un individuo son los p valores que toma para las variables. Esta representación es abstracta pero tiene

las mismas propiedades de las representaciones en dos dimensiones, que se denominan diagramas de dispersión (ver figura 1.1).

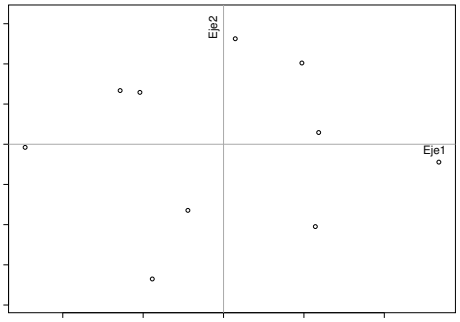
Dos individuos están cercanos en $\mathbb{R}^p$ si tienen más o menos las mismas coordenadas, es decir valores similares para las p variables. Entonces la representación geométrica es útil para comparar a los individuos entre si y observar la estructura de la "nube de individuos", es decir los n puntos,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



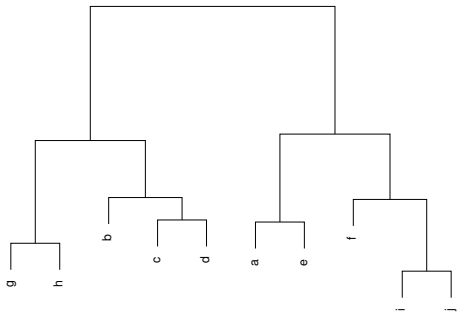
Esquema de la tabla



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

Para hacerlo 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 buscan descubrir grupos en los datos, de tal manera que los individuos se parezcan lo más posible cuando pertenecen a un mismo grupo y que sean lo más diferentes posible cuando pertenezcan a diferentes grupos.

Al 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. Con algunas transformaciones los ángulos entre los vectores variables representan su correlación y la longitud de las flechas sus desviaciones estándar. Mediante proyecciones sobre planos se puede ver de manera conjunta las relaciones entre todas las variables.

Los métodos en ejes principales o factoriales básicos se muestran esquemáticamente en la figura 1.2. Éstos se introducen en el capítulo 2, donde se presenta el análisis en componentes principales (ACP), empleado en la exploración de tablas de individuos descritos por variables continuas.

En el capítulo 3 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 covarianzas y las de correlación, cuando no se dispone de los datos con las que fueron calculadas. También el método de análisis en coordenadas principales (ACO), utilizado para obtener imágenes Euclidianas para las matrices de distancia 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 4 se presenta el análisis de correspondencias simples (ACS), utilizado para la descripción de tablas de contingencias, 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.

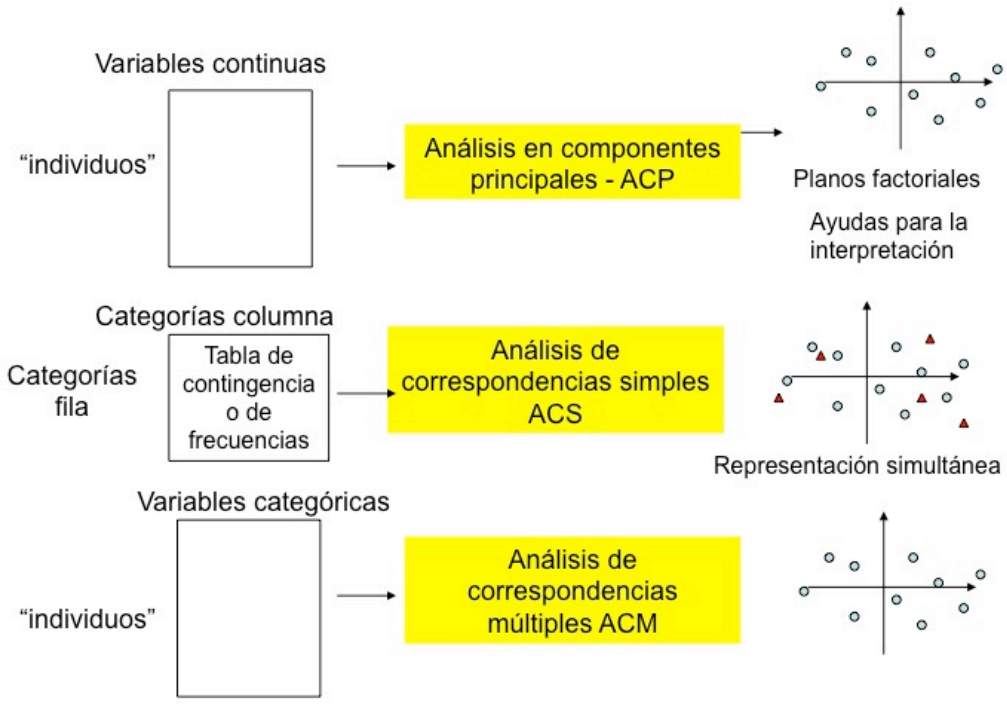


Figura 1.2: Esquema de los métodos factoriales básicos

El capítulo 5 se dedica al análisis de correspondencias múltiples (ACM), usado en la descripción de tablas de individuos descritos por variables cualitativas. El ACM se muestra como una generalización del ACS, que es el análisis de correspondencias de la tabla indicadora de las particiones asociadas a las variables categóricas, que tiene n individuos en las filas por p categorías en las columnas.

Los métodos de clasificación automática se presentan en el capítulo 6 y corresponden a la búsqueda de grupos de las nubes de individuos mediante dos tipos de algoritmos, que se combinan entre si y con los métodos factoriales, para completar la descripción multivariada básica de una tabla de datos.

1.2. El lenguaje estadístico R

Si aceptamos que el profesional en Estadística debe manejar al menos tres lenguajes: el Español, el Inglés y el Matemático-Estadístico; el R se convierte en un cuarto lenguaje. El lenguaje estadístico R permite, en primer lugar, hacer los cálculos del Álgebra Lineal, parte de matemáticas fundamental en la construcción de la Estadística, en segundo lugar realizar gráficas que están siempre presentes en la Estadística y en tercer lugar llevar a cabo métodos específicos de la Estadística. 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

Para 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. Para ahorrar palabras en la secuencia de instrucciones se utiliza, en este documento, el símbolo $\longrightarrow$ y se debe hacer clic en la palabra que aparece a continuación de él, en el menú que aparezca. 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* $\longrightarrow$ *Mirror* y escoja uno.
- Clic en *Download R for Windows* $\longrightarrow$ *base* $\longrightarrow$ *Download R 3.3.3 for Windows (71 megabytes, 32/64 bit)*

  Los usuarios de Linux y OS X de Mac pueden seguir las instrucciones de la página

Web de R.

- Guarde el archivo en un directorio $\longrightarrow$ *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 ellos se pueden producir con el menú.
- Para leer el manual de introducción: clic en Help $\longrightarrow$ *An Introduction to R*.

Es muy conveniente para quien empiece con R leer esta introducción, ya que redundará en el mejor uso del lenguaje y a la postre en ahorro de tiempo. Existe una versión en español: http://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf. El usuario de R también está llamado a leer esta introducción, al menos de manera parcial, de acuerdo con sus intereses.

Ahora la consola de R espera comandos y los primeros usos, en la lectura de este texto, son los de calculadora graficadora y calculadora matricial. R diferencia mayúsculas y minúsculas, `<-` indica asignación (también se puede utilizar `=`) y el símbolo `#` se utiliza como comentario (el texto que aparece a la derecha de `#` se presenta pero R no lo interpreta).

El lenguaje R está enmarcado dentro de la programación orientada a objetos y 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, sobre todo cuando es en un paquete que requiere instalación. Por ejemplo `plotct{FactoClass}` indica que la función `plotct` está en el paquete `FactoClass`.

1.2.2. Instalación de paquetes

Para este documento se requiere instalar en R los paquetes: `ade4`, `scatterplot3d` y `FactoClass`, sin embargo este último requiere los otros dos, de modo que R instala todos al instalar `FactoClass`. Un procedimiento para hacerlo es mediante el comando:

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

Otra forma de instalar uno o varios paquetes en R la barra de control de la consola:

- Clic en *Paquetes* (en la barra de menú situada en la parte superior) $\longrightarrow$ *Instalar paquetes*, aparece lista de *Mirrors*.
- Seleccione un *Mirror* $\longrightarrow$ *OK*, aparece lista de paquetes.
- Selecciones los paquetes (con la tecla *Control* oprimida cuando se quieren varios) `ade4`, `ade4TkGUI`, `scatterplot3d` y `FactoClass` $\longrightarrow$ *OK*.

1.2.3. RStudio y documentación con Sweave y Markdown

RStudio es el primer IDE (*Integrated Development Environment*) para R http://www.rstudio.com/products/rstudio/features/ y facilita la documentación de los trabajos realizados con R, integrando *Sweave* y *R Markdown*, por lo tanto es recomendable utilizarlo.

"*Sweave* (Leisch & R-core 2016) es una herramienta que permite integrar el código de R para llevar a cabo el análisis de datos dentro de documentos LaTeX. "LaTeX es *de facto* el estándar para la comunicación y publicación de documentos científicos" http://www.latex-project.org/. Este documento se ha editado en LaTeX.

"*R Markdown* es un formato de edición que permite la creación fácil de documentos dinámicos, presentaciones y reportes desde R" http://rmarkdown.rstudio.com/.

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 para facilitar su uso y que ha denominado DtmVic; está disponible bajo Windows y es de uso libre para propósitos académicos y de investigación. El DtmVic se utiliza como programa de referencia para el curso.

El DtmVic se baja de la página http://www.dtmvic.com, se descarga el archivo `dtm_software.zip` (http://www.dtmvic.com/DEA/dtm_software.zip), se descomprime la carpeta y se ejecuta haciendo clic en `DtmVic_5_6.exe`. Por otro lado se descarga el archivo de ejemplos: `DtmVic_Examples` (http://www.dtmvic.com/DEA/DtmVic-Examples.zip).

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` (http://mcj.sourceforge.net/), 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 exportan directamente a este formato (`.fig`) y desde `Xfig` a casi cualquier formato gráfico. Una manera de exportación a R 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 en la edición de las gráficas de estas notas, 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 2 de Morrison (1990): *Matrix Algebra*, disponible en http://www.stat.duke.edu/courses/Spring10/sta345/morrison/Mori1990a.pdf. Se recomienda utilizar R para verificar los ejemplos y ejercicios numéricos de este capítulo;
- el capítulo 2 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. La tabla se constituye en un producto intermedio dentro de un proyecto de investigación y puede tener distintos orígenes. La tabla por si sola no tiene ningún interés de análisis sino que lo tiene en cuanto forma parte de un contexto de investigación. En algunos casos llegar a ella puede costar, por ejemplo, el 80 % del presupuesto de una investigación. El análisis de la tabla de datos está orientado por el contexto de la investigación.

El estudiante de este curso debe situarse en el papel de investigador dentro del contexto de la investigación de donde 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 queramos abordar es compleja y no es posible entenderla en su totalidad. Tenemos que aceptar que lo que observemos de la realidad será casi siempre parcial y lo que describamos dependerá de los objetivos que planteemos. Es fundamental entender el contexto de la realidad que queremos estudiar y plantear de manera clara los objetivos que queremos resolver.

La información obtenida en una investigación se almacena en una base de datos, la cual está 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 requiere de una o más tablas de datos, que luego se describirán por uno o más métodos de los abordados en este curso. Las decisiones que hay que tomar, en el sentido de las técnicas estadísticas a usar y los aspectos al interior de ellas, requieren, además del conocimiento

de éstas, el del contexto en el que se enmarca la tabla de datos.

Todo 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 haciendo. Sin embargo existen, en la literatura, muchos textos guía para ir mejorando esas competencias, por ejemplo el de Hernández, Fernández & Baptista (2006). En este curso se hace un trabajo con el doble propósito de mejorar las competencias, tanto en metodología como en el uso apropiado de los métodos básicos de la Estadística descriptiva multivariada.

Los métodos estadísticos abordados en este curso son especialmente útiles en ciencias sociales y humanas, por lo cual forman parte de la *Metodología de la investigación cuantitativa* de estas ciencias (Briones 1996).

1.7. Trabajo del curso

El avance de los estudiantes en metodología de la investigación y en la aplicación de los métodos abordados en este curso se hace elaborando un trabajo utilizando datos existentes. Los estudiantes deben leer sobre metodología de la investigación y sobre el contexto de donde enmarca el trabajo que desean realizar. Se debe entender que no existe "una metodología de la investigación" sino una serie de guías que nos ayudan a construir nuestra metodología. Las guías proveen una secuencia de actividades generalmente bajo el supuesto de que vamos a empezar una investigación desde su génesis. Para este curso se debe disponer de los datos, lo que muchas veces significa que se está en la fase final de una investigación. En algunas ocasiones se parte de investigaciones ya concluidas y lo que se busca es una explotación nueva de los datos, lo que generalmente requiere la definición de objetivos que no estaban en el estudio original.

El trabajo del curso se realiza en grupo y se concreta a través de tres informes escritos: propuesta, proyecto y trabajo final.

1.7.1. Propuesta

Los miembros del grupo acuerdan el tema que desean trabajar y buscan en la literatura y su entorno, un trabajo ya realizado a condición que se tenga acceso a los datos o a una base de datos administrativa sobre la que se desee hacer un análisis. Es ideal que exista al menos un profesional interesado en el trabajo que van a realizar los estudiantes, dispuesto a colaborar.

El contenido de la propuesta es:

- Título: debe dar una idea clara de lo que se va a hacer en el trabajo, desde el punto de vista práctico, es decir del contexto de la investigación de donde se han tomado los datos.
- Introducción: se describe el contexto de la investigación, sus objetivos, y se introduce lo que se desea obtener en este trabajo.
- Descripción de los datos: descripción de la base de datos y el procedimiento como se llegó a ella.
- Referencias.

1.7.2. Proyecto

Se repite el contenido de la propuesta, mejorando de acuerdo a las observaciones y concretando los objetivos que se quieren cumplir con el trabajo. El cuerpo del escrito debe contener las siguientes secciones:

1. Introducción: la de la propuesta pero mejorada según las observaciones y las discusiones del grupo.
2. Objetivos: objetivos específicos del trabajo. Estos deben ser prácticos, es decir del contexto, y no deben contener pasos metodológicos, lo cual implica que no se incluyen términos estadísticos en su redacción.
3. Descripción de la tabla (o tablas) de datos que se utilizarán en el trabajo. En la descripción de una tabla de datos se debe expresar claramente quiénes son los “indi-

viduos" y cómo se llegó a ellos: ¿son una población?, ¿son una muestra?, ¿cómo fue el procedimiento de muestreo?, ¿se dispone de factores de expansión?. Las variables se deben colocar en una tabla, organizándolas por grupos temáticos e incluyendo su resumen univariado: media, desviación estándar, mínimo y máximo, para las variables continuas, porcentaje de cada categoría para las variables categóricas.

4. Metodología: el procedimiento estadístico con el que van a dar respuesta a cada uno de los objetivos: métodos, variables activas, variables ilustrativas, etc.

5. Referencias.

1.7.3. Trabajo final

Se presentan de nuevo las secciones del proyecto y se incluyen los resultados y sus análisis y las conclusiones. Entonces el cuerpo del trabajo final contiene las siguientes secciones:

1. Introducción.

2. Objetivos.

3. Descripción de los datos.

4. Metodología.

5. Análisis de resultados.

6. Conclusiones.

7. Referencias.

Desde hace algunos años se vienen editando los documentos estadísticos en LaTeX y se dispone de una plantilla para los artículos de la Revista Colombiana de Estadística, que se ha extendido para los documentos del Simposio de Estadística, para los trabajos de grado y para los trabajos de cursos como éste. La plantilla para los trabajos de este curso, *Report.zip*, se encuentra en: https://sites.google.com/site/eccubidesg/latex/.

1.8. Preparación de los datos para el análisis

En la versión descriptiva de los métodos multivariados las variables juegan dos papeles complementarios: se denominan *activas* a las que se seleccionan para la construcción de ejes factoriales y clases; e *ilustrativas* a la que juegan el papel de explicar o ilustrar los ejes o clases obtenidos. Las variables activas requieren más atención en su preparación para el análisis multivariado por su influencia en la estructuración de 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 una tabla con muchas columnas, de donde se debe obtener una tabla para el análisis específico que se desea abordar. Como ejemplo para esta y otras secciones del texto se utiliza parte de una consulta del Sistema de Información Académico de la Universidad (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 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. Los resultados se presentan estandarizados con media 10 y y desviación estándar 10. 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; para el primer semestre de 2013.

La hoja de datos retenida para el ejemplo tiene las columnas correspondiente la carrera, los resultados del examen en cada área y global. 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 lecto-escritura y matemáticas básicas; se incluyen las dos variables con las categorías *si* y *no*.

La hoja de datos se incluye en el objeto `admi{FactoClass}`, que tiene 445 admitidos y 15 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*) o origen geográfico (*orig*), se recoficaron con menos categorías de 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. Se incluyeron para utilizarlas en los ejemplos de transformación de variables.

Código R. Para obtener la figura 1.3:

```
library(FactoClass) # cargar FactoClass
data(admi)  # cargar la tabla
par(mfrow=c(3,2),mai=c(0,0,0,0)) # para poner 6 tortas sin espacios
    entre gráficas
for (i in c(1,8,14,10,12,13)){
  cat<-attributes(admi[,i])$levels; per<-tabulate(admi[,i])/445*100
  paste(cat,round(per,1),sep="%")->eti; pie(summary(admi[,i]),eti)
}
dev.print(device = xfig) # grabar la gráfica como Rplot001.fig
```

Código R. Para obtener la figura 1.4, que incluye los diagramas de caja y bigotes para edad y 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(2:7,15)) hist(admi[,i],main=names(admi)[i])
boxplot(admi$age,main="age");boxplot(admi$exam,main="exam")
dev.print(device = xfig)
```

1.8.1. Transformación de variables cualitativas

Las descripciones multivariadas requieren un trabajo previo sobre las variables cualitativas. Los datos faltantes o no respuestas se suelen codificar como una categoría adicional; lo mismo se suele hacer con los *no aplica* cuando están presentes. En algunas variables se deben agrupar categorías, buscando que no haya categorías con frecuencia muy baja y que 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

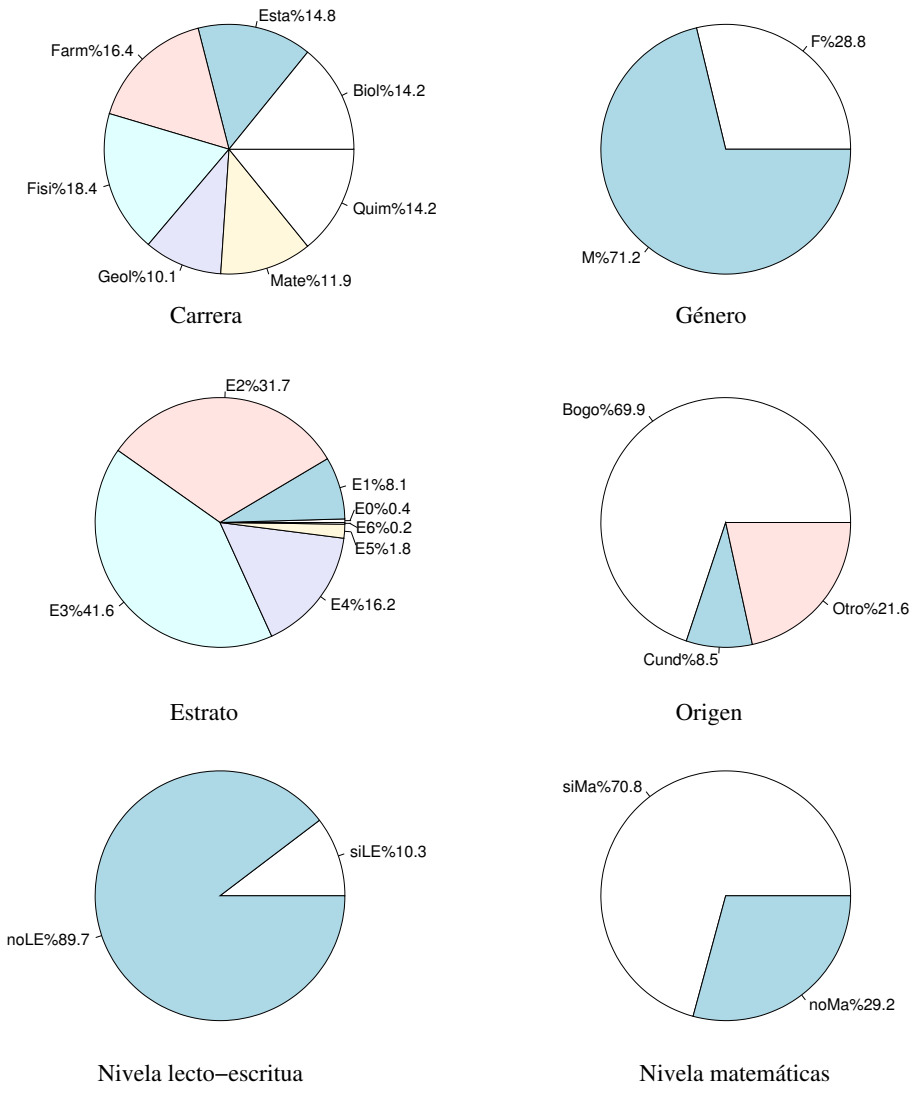


Figura 1.3: Tortas mostrando la distribución de las categorías de las variables cualitativas de los admitidos a la Facultad de Ciencias.

frecuencia para casi todos. Se dejaron 3 categorías: Bogotá, Cundinamarca y Otro, con el resto de departamentos. La distribución de los admitidos en las tres categorías se muestra en la torta *Origen* de la figura 1.3. La variable estrato del admitido *stra* requiere una agrupación de categorías, porque como puede verse en la figura 1.3, los estratos 0, 1, 4, 5 y 6 tienen frecuencias muy bajas. Se agrupan entonces en 3 categorías: *bajo* (0, 1 y 2), *medio* (3) y *alto* (4, 5 y 6), en la variable `estr` de `admi{FactoClass}`. En los comandos de R siguientes se muestra una manera de obtener la nueva variable `estr`. En el código R se observa la distribución de frecuencias luego de `summary(estr)`.

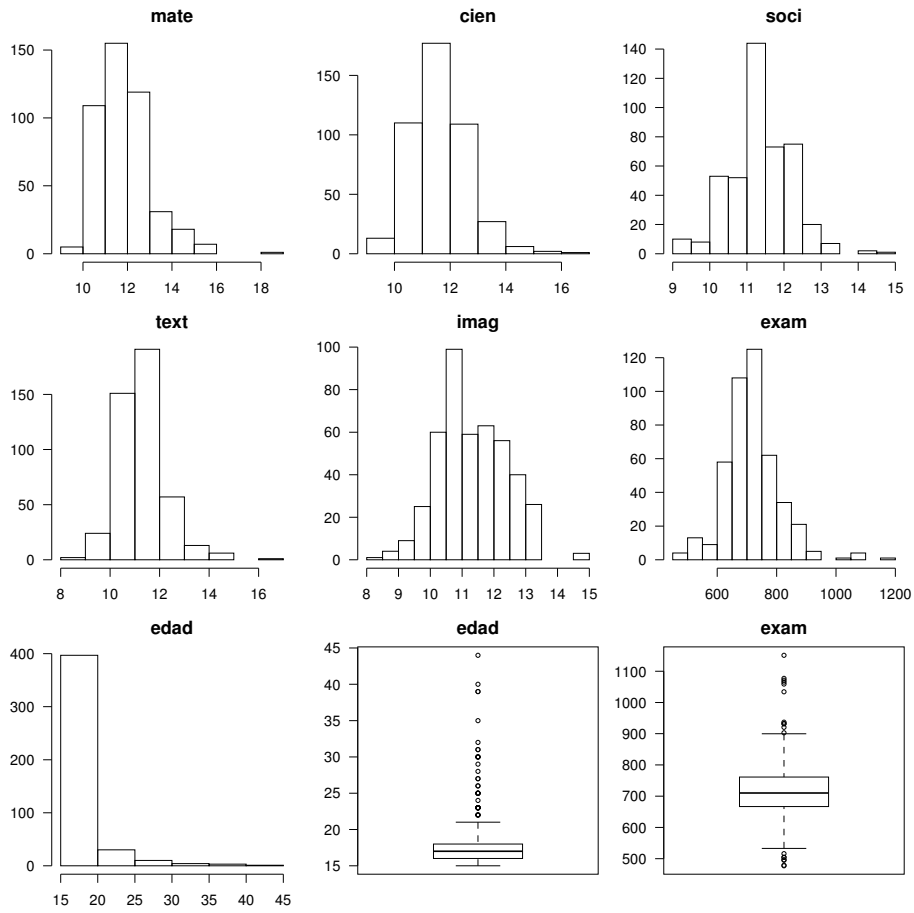


Figura 1.4: Histogramas de los puntajes obtenidos en el examen de los admitidos a la Facultad de Ciencias.

Código R. 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
```

1.8.2. Codificación en clases de variables continuas

Hay diferentes razones para que algunas veces se decida convertir algunas variables continuas en una ordinales. Una razón es para utilizarlas como variables activas en un análisis de correspondencias múltiples. Por ejemplo, la edad en años se suele pasar a ordinal por

su naturaleza de variable sociodemográfica, y 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, de la teoría de la información y de las propiedades del análisis de correspondencias múltiples. Se pueden resumir en cuatro:

1. Los límites de las clases deben respetar argumentos del contexto, por ejemplo, en el caso de la edad, 18 años porque se alcanza la mayoría de edad, 60 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 de las variables activas, en un análisis de correspondencias múltiples, sea más o menos igual. La influencia de las variables en el análisis es proporcional al número de categorías.

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 utilizando los cinco números de Tukey, que divide 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.

Código R. 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
```

1.9. 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.

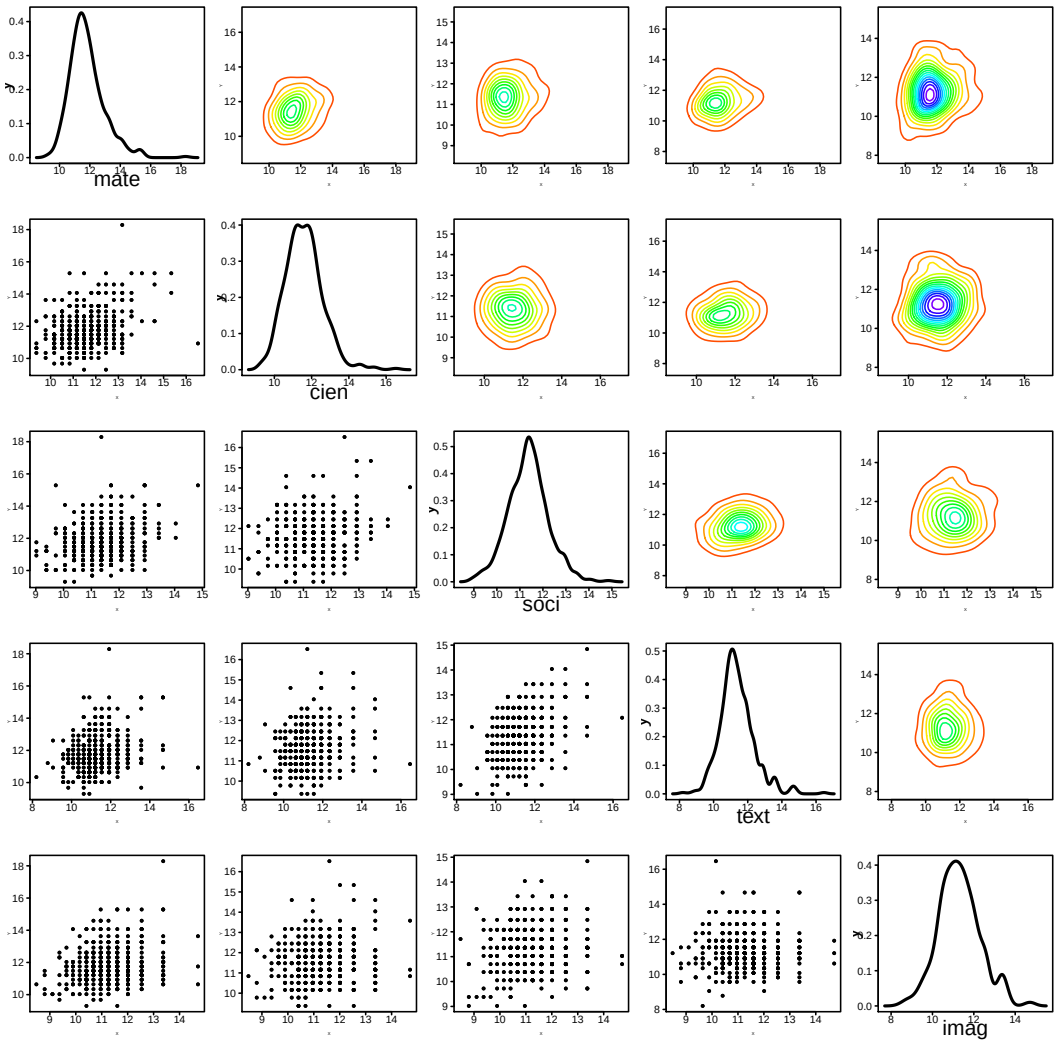
1.9.1. Descripción de parejas de variables continuas

La descripción de dos variables continuas se logra con los diagramas de dispersión y se complementa con las covarianzas y 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}`, ensambla en una gráfica: las densidades *kernel* univariadas en la diagonal, las densidades *kernel* bivariadas en el para triangular superior, y los diagramas de dispersión en la parte triangular inferior. Las varianzas y covarianzas y las correlaciones se arreglan en sendas matrices simétricas, que tienen en la diagonal las varianzas de las variables (1 en el caso de la matriz de correlaciones) y las covarianzas o correlaciones por fuera de la diagonal.

Los diagramas de dispersión de los resultados del examen de admisión de los aspirantes admitidos a las Carreras de la Facultad de Ciencias, según las áreas: matemáticas, ciencias, sociales, lenguaje e imagen; se presentan en la figura 1.5, junto con las matrices de varianzas y covarianzas y correlaciones que se presentan debajo de los diagramas de dispersión. Se observan mayores relaciones lineales entre matemáticas y ciencias, textual y sociales, ciencias y sociales.

La relación del puntaje total y los puntajes parciales se da por construcción, porque el total es un resumen de los puntajes por áreas, en efecto los coeficientes de correlación son importantes, siendo mayores con matemáticas y ciencias:

```
> round(cor(admi$exam,admi[,2:6]),3)
      mate  cien  soci  text  imag
[1,] 0.753 0.653 0.593 0.519 0.458
```




```
plotpairs(admi[,2:6]); dev.copy2pdf(file="AdmiPairs.pdf")
```

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

```
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 1.5: 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.

1.9.2. 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 entre los 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 1.6. Se observa que a las carreras de Geología, Física y Matemáticas ingresan, en promedio estudiantes con mejores resultados en el examen, en esta cohorte los admitidos de menores puntajes son los de Química, Farmacia y Estadística. Individualmente los mejores puntajes están en Física, Matemáticas y Geologia 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.

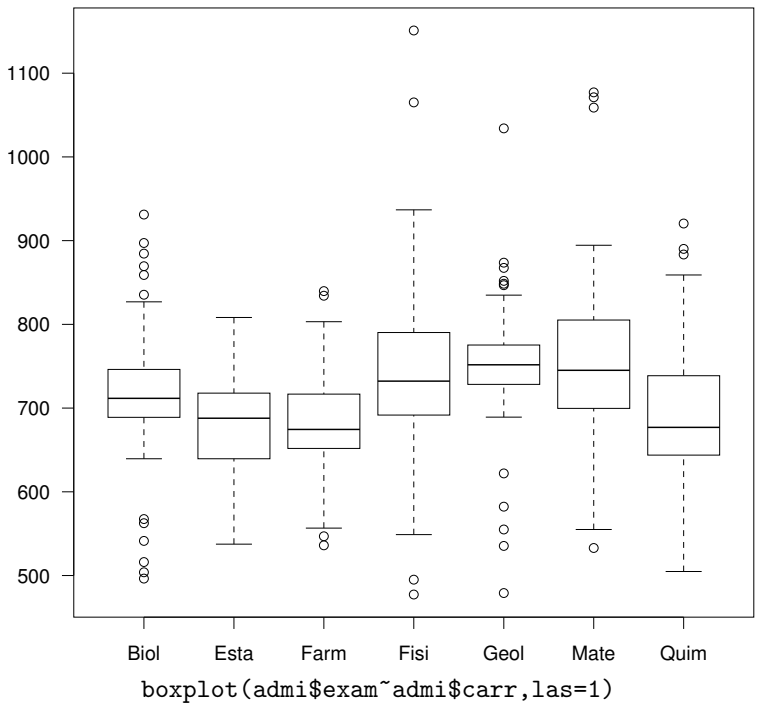


Figura 1.6: Distribuciones del puntaje del examen obtenido por los admitidos según carreras.

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 inter* y *varianza total*. La razón de correlación se define como el cociente entre varianza inter y varianza entre. La varianza inter es:

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

donde $\overline{X}_k = \sum_{i \in I_k} \frac{1}{n_k} X_i$, 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. Entonces la razón de correlación, entre la variable cuantitativa X y la variable cualitativa Y es:

$$\eta^2(X,Y) = \frac{Var_X(inter)}{Var(X)} \tag{1.1}$$

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

```
> round(centroids(admi[,2:7],admi$carr)$cr*100,2)
      mate cien soci text imag  exam
[1,] 15.86 4.31 3.22 2.34 4.42 11.87
```

Las razones de correlación son bajas pero se destacan las de matemáticas y el puntaje global (*exam*) sobre las demás. En la sección siguiente se caracterizan mejor las carreras según los resultados del examen de los admitidos.

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 y categorías que más caracterizan a un grupo de n_k individuos como subconjunto de los n individuos. Un valor test es un índice de ordenamiento obtenido mediante la metodología

de una prueba de hipótesis, aunque no se trata de hacer inferencias estadísticas. Para un grupo de n_k individuos, el valor test se obtiene suponiendo que los n_k individuos se extraen, aleatoriamente y sin reemplazamiento, del conjunto de los n individuos de la tabla de datos. En el caso de una variable continua se trata de la comparación de la media del grupo con la media global. Entonces la hipótesis nula es: $H_0 : \mu_k = \mu$ y la hipótesis alterna: $H_1 : \mu_k \neq \mu$.

Para contrastar esta hipótesis se utiliza la estadística de prueba:

$$T_k = \frac{\bar{X}_k - \mu(X)}{\sigma_k(X)} \quad con \quad \sigma_k^2(X) = \frac{n - n_k}{n - 1} \frac{\sigma^2(X)}{n_k}$$

$\frac{n - n_k}{n - 1}$ es el factor de corrección por tamaño de población finita. Los n valores de la variable X juegan el papel de la población, entonces los valores de $\mu(X)$ y $\sigma(X)$ son conocidos, ya que se calculan con los n datos. Bajo H_0 la distribución de T_k es normal estándar y el valor test se interpreta como un cuantil de esta distribución. El valor test corresponde a una estimación de la estadística y es:

$$t_k = \frac{\bar{x}_k - \bar{x}}{\sigma_k(X)} \quad con \quad \sigma_k^2(X) = \frac{n - n_k}{n - 1} \frac{\sigma^2(X)}{n_k} \tag{1.2}$$

Valores test más grandes, en valor absoluto, indican mayor diferencia entre las medias del grupo correspondiente a la categoría y la de todos los individuos.

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.

Por ejemplo el valor test para el puntaje total del examen en la Carrera de Estadística: $n = 445$, $n_{esta} = 66$, $\bar{x}_{esta} = 680$, $\bar{x} = 718$, $var(X) = 8039$, entonces:

$$s_{esta}^2(X) = \left(\frac{445 - 66}{444}\right) \frac{8039}{66} = 104 \quad t_{esta} = \frac{680 - 718}{\sqrt{104}} = -3.73$$

Código R. Para caracterizar las carreras según los resultados de sus admitidos en el examen por áreas y puntaje total (tabla 1.1):

```
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 1.1: Caracterización de las carreras según los resultados del examen de admisión por áreas y global

categoria	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 1.1 se puede observar que los admitidos a Biología obtienen en promedio menores puntajes en matemáticas, los de Estadística obtienen en promedio menores puntajes 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 mejores puntajes en matemáticas y ciencias, los de Geología en Matemáticas y los de Matemáticas en matemáticas e imagen.

1.9.3. Descripción de dos variables cualitativas

Para describir la asociación de dos variables cualitativas se construyen tablas de contingencia (TC), que son la clasificación de los individuos por las categorías de las dos variables simultáneamente. Por ejemplo, para ver la asociación entre edad y estrato de los admitidos se construye la TC (tabla 1.2), con el código siguiente para poner totales fila y columna y la tabla de frecuencias relativas, también con las marginales.

Código R. Para obtener la tabla 1.2:

```
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 1.2: Tabla de contingencia edad×estrato de los admitidos, tabla de frecuencias relativas y código R para obtenerlas

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

La asociación entre categorías fila y columna se visualiza con los histogramas de perfiles fila y columna, puestos como barras del 100 %, cuyas franjas de colores representan el porcentaje de cada categoría en el histograma de la fila o columna. Los perfiles son distribuciones condicionales cuando se asimila la tabla de frecuencias relativas como la distribución de probabilidad conjunta entre las dos variables.

Código R. Para obtener la figura 1.7, utilizando la función `plotct{FactoClass}`:

```
par(mfrow=c(2,1),mai=c(0.4,1,0.3,0.1))
plotct(t(tc),"row",col=c("white","yellow","green","blue"))
plotct(tc,"row",col=c("white","yellow","green","blue"))
```

Se observa, en la figura 1.7 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.

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 1.8.

Dos medidas de asociación entre variables cualitativas

La estadística χ^2 utilizada para contrastar la hipótesis de independencia se puede utilizar como medida de asociación entre dos variables cualitativas. En una tabla de contingencia 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, $n_{j\cdot}$ son que asumen la categoría j (marginal fila), $n_{\cdot k}$, los que asumen la categoría k (marginal columna) y n el total de la tabla (total

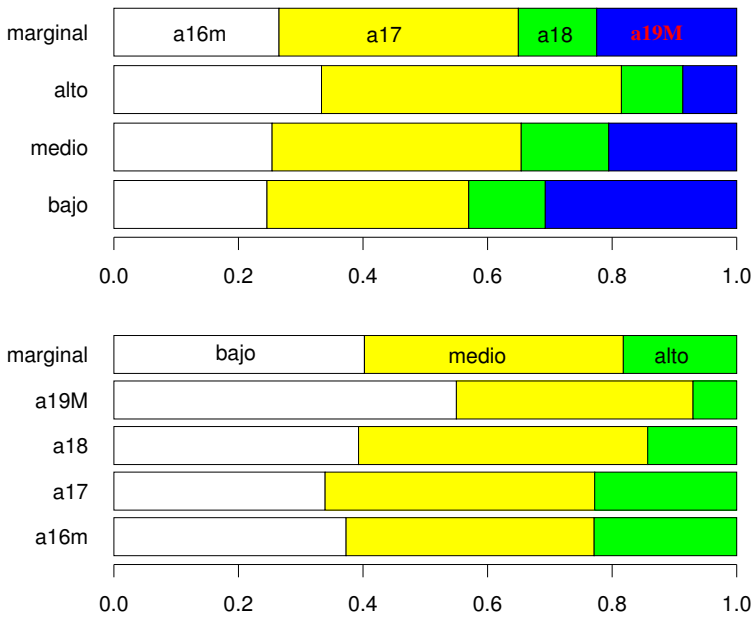


Figura 1.7: Perfiles fila y columna de la TC edad×estrato.

de individuos). La estadística χ^2 se expresa (ver por ejemplo Canavos (1988, p. 370)):

$$\chi^2 = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{jk} - \frac{n_{j\cdot}n_{\cdot k}}{n}\right)^2}{\frac{n_{j\cdot}n_{\cdot k}}{n}} \tag{1.3}$$

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 $\frac{n_{j\cdot}n_{\cdot k}}{n}$ es inferior a 5 (Agresti 2002, p. 78).

Aquí se utiliza como índice descriptivo, calculando además su valor p ($P(\chi^2 \geq \chi^2_c)$, donde χ^2_c 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 $Z \sim N(0,1)$.

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

$$\phi^2 = \frac{\chi^2}{n} \tag{1.4}$$

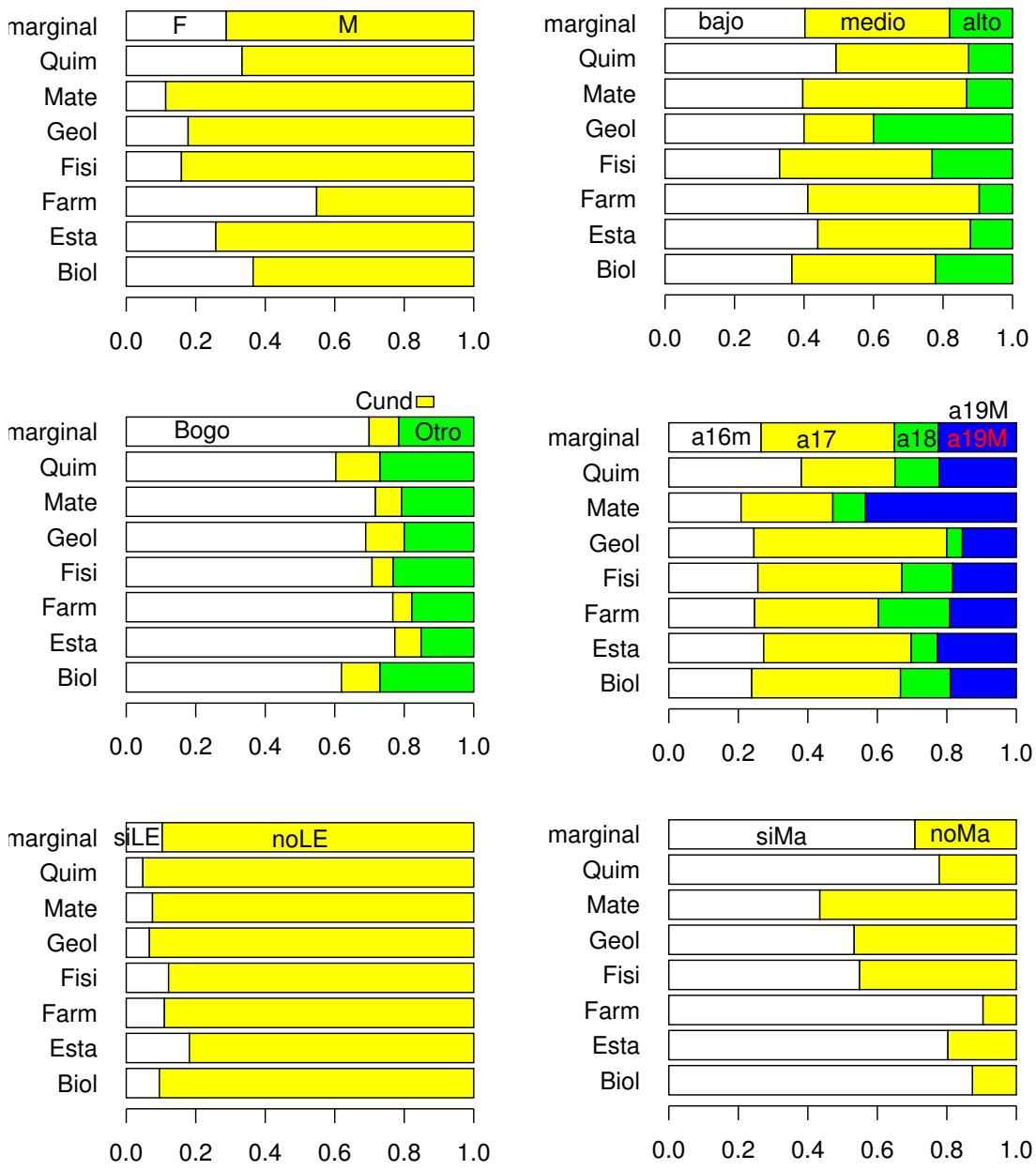


Figura 1.8: Perfiles de las carreras según variables cualitativas

Con la función `chisq.carac{FactoClass}` se pueden obtener los índices de asociación presentados de una variable cualitativa (a caracterizar) con otras variables cualitativas. Se muestra, como ejemplo, la asociación de las carreras con las variables sociodemográficas de los admitidos: *gene*, *estr*, *orig* y *edad*.

Código R. Para calcular índices de asociación entre *carrera* y variables sociodemográficas:

```
> round(chisq.carac(admi[,8:11],admi$carr,decr = FALSE),3)
       chi2 dfr  pval   tval  phi2
gene 44.108   6 0.000  5.264 0.099
estr 29.190  12 0.004  2.679 0.066
orig  9.676  12 0.644 -0.370 0.022
edad 33.553  18 0.014  2.189 0.075
```

Sólo entre carreras y origen de los admitidos no se muestra asociación. A continuación se muestras las categorías responsables de las relaciones entre carreras y las otras variables a través de sus categorías.

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 se recurre, de nuevo, a calcular los valores test siguiendo la metodología de una hipótesis. En la urna hay n "bolas"con n_j que tienen la característica j, de la urna se extrae una muestra sin reemplazo de tamaño n_k, se define la variable aleatoria $N =$"número de bolas con la característica j al extraer n_k bolas". Esta variable aleatoria tiene distribución hipergeométrica de parámetros n, n_j, n_k. Para el resultado de la "estimación" n_{kj} se calcula $P(N \geqslant n_{kj})$, si $\frac{n_{kj}}{n_k} \geqslant \frac{n_j}{n}$ y el valor test es el cuantil de la normal estándar que deja a la derecha un área igual a la mitad de esta probabilidad. Si la desigualdad es en el otro sentido el valor test es el cuantil de la normal estándar que deja a la izquierda una área igual a la mitad de la probabilidad $P(N \leqslant n_{kj})$ calculada con la distribución hipergeométrica. La función `cluster.carac{FactoClass}` hace este trabajo presentando, para cada una de las categorías de la variable que se está caracterizando, las categorías de las demás variables que la caracterizan, ordenadas por los valores test. La función presenta solamente las categorías que tienen valores test superiores, en valor absoluto, al umbral $v.lim = 2$ (ver ayuda de la función: `?cluster.carac`). En la tabla 1.3 se muestra la salida asociada a la figura 1.8, que en este caso se puede tomar como ayuda a la lectura de la figura, pero generalmente se usan estas salidas, sin las gráficas.

Tabla 1.3: Caracterización de las carreras según algunas variables cualitativas

categoria	carrera	v.test	p.valor	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

Para mostrar un cálculo tomemos la única variable característica de la carrera de Biología, que muestra una frecuencia más alta de los que tienen que nivelar matemáticas (87.3 %) con respecto al porcentaje global (70.8 %). La tabla de contingencia de $niMa \times carr$ y sus marginales se muestran en la tabla 1.4.

Entonces, los parámetros para la distribución hipergeométrica son: $n = 445$, $n_k = n_{biol} = 63$, $n_j = n_{siMa} = 315$ y $n_{kj} = 55$. Hay que calcular primero la probabilidad $P(N \geq 55)$, $N \sim H(445, 315, 63)$.

Código R. Para calcular el valor test:

```
> vp<-phyper(54, 315, 130,63,lower.tail=FALSE);vp
[1] 0.000862394
> qnorm(vp/2,lower.tail=FALSE)
[1] 3.331951
```

En la figura 1.9 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.

La tabla 1.4 sirve para entender bien la lectura de los porcentajes de la tabla 1.3. En la primera fila: el porcentaje de los que tienen que nivelar matemáticas dentro de los admitidos a la Carrera de Biología es 87.3 %, mientras que el porcentaje de todos los admitidos a la Facultad que tienen que hacerlo es 70.8 % (ver en perfiles columna=carreras). 17.5 % (ver perfiles fila=niMa) es el porcentaje que hay en Biología de los que tienen que nivelar Matemáticas. El valor test de 3.332 está indicando que 87.3 % es suficientemente mayor que 70.8 % para concluir que la Carrera de Biología se caracteriza por tener más proporción de admitidos que tienen que nivelar Matemáticas. Como niMa, solo tiene 2 categorías el valor test para los que no tienen que nivelar es -3.332, es decir que Biología se caracteriza por tener menos proporción que el global de los que no tienen que nivelar matemáticas, sin embargo no hay necesidad de decirlo porque es el resultado complementario.

Para obtener la tabla 1.3 se utiliza la función `plotct{FactoClass}`, que además de hacer las gráficas de perfiles fila y columna, retorna las tablas que aparecen.

Código R. Para obtener la tabla 1.3:

```
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))
```

Tabla 1.4: 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	
noMa	6.2	10.0	5.4	28.5	16.2	23.1	10.8	
marg	14.2	14.8	16.4	18.4	10.1	11.9	14.2	

Perfiles columna

siMa	87.3	80.3	90.4	54.9	53.3	43.4	77.8	70.8
noMa	12.7	19.7	9.6	45.1	46.7	56.6	22.2	29.2

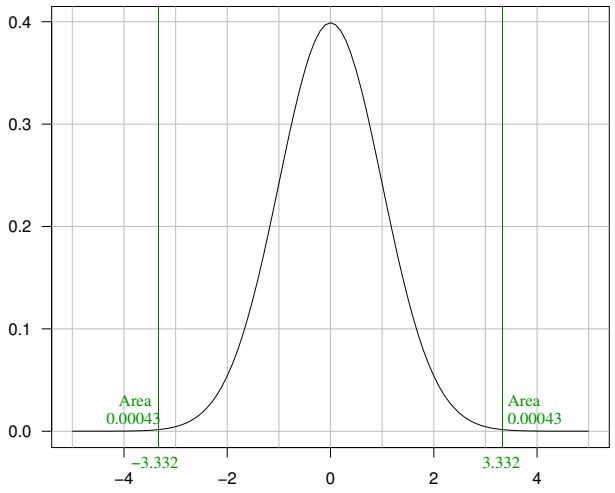


Figura 1.9: Ilustración de la obtención del valor test a partir de una probabilidad (área bajo la curva normal estándar).

1.10. 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 no igual?

   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)`

5. Repita los análisis realizados en la sección 1.9 utilizando el archivo *EDMbivariadoAdmi.Rmd* y *RStudio*.

1.11. Taller: caracterización de una variable cualitativa por variables nominales y cuantitativas

El Departamento de Estadística de la U.N. realizó, hace algunos años, una encuesta a 424 jueces de la República, para conocer los problemas de la justicia y las opiniones de los jueces sobre el asunto.

Objetivo

Caracterizar el *tipo de juzgado* y la *región* por las demás variables, se espera que el estudiante resuelva las siguientes preguntas:

1. ¿Hay diferencias en opinión entre los diferentes tipos de jueces (respectivamente, entre regiones)? ¿Cuáles son las más destacadas?

2. ¿Hay diferencias entre los grupos de jueces por tipo de juzgado (respectivamente, por región) en edades y en tiempo de servicio a la rama judicial?

Los datos

Los datos están en archivos compatibles con el programa *DtmVic*: *JuezDatos.txt* para los datos y *JuezDiccio.txt* para el diccionario de las variables, que incluye las etiquetas para las categorías de las variables nominales y que son las siguientes:

```
-------------------------------------------------------------------------------------------
   1 . Tipo de juzgado                                              (   5 categories )
          JCIV - Jcivil               JLAB - Jlaboral               JFAM - Jfamila
          JPEN - Jpenal               JPRO - Jpromiscuo
 -------------------------------------------------------------------------------------------
   2 . Region                                                       (   6 categories )
          RATL - Ratlantica           RCAF - Rcafetera              RORI - Roriental
          RPAC - Rpacifica            RCEN - Rcentral               RSUO - Rsur-oriental
 -------------------------------------------------------------------------------------------
   7 . Congestion judicial en 1996                                  (   3 categories )
          C6SI - Si congestion 96     C6NO - No congestion 96       C6NR - NR congestion 96
 -------------------------------------------------------------------------------------------
   8 . Recursos fisicos suficientes                                 (   3 categories )
          RFSI - Si recursos fisicos  RFNO - No recursos fisicos    RFNR - NR recursos fisicos
 -------------------------------------------------------------------------------------------
   9 . Herramientas tecnologicas suficientes                        (   3 categories )
          HTSI - Si herramientas tecn HTNO - No herramientas tecn   HTNR - NR herramientas tecn
 -------------------------------------------------------------------------------------------
  10 . Adaptacion a los nuevos cambios                              (   3 categories )
          ACSI - Si adaptacion cambio ACNO - No adaptacion cambio   ACNR - NR adaptacion cambio
 -------------------------------------------------------------------------------------------
  11 . Satisfecho archivos                                          (   3 categories )
          SASI - Si satisfecho archiv SANO - No satisfecho archiv   SANR - NR satisfecho archiv
 -------------------------------------------------------------------------------------------
  12 . Suficientes empleados despacho                               (   3 categories )
          EMSI - Si suficientes emple EMNO - No suficientes emple   EMNR - NR suficientes emple
 -------------------------------------------------------------------------------------------
  13 . Idoneidad apropiada de los empleados                         (   3 categories )
          IESI - Si idoneidad emplead IENO - No idoneidad emplead   IENR - NR idoneidad emplead
 -------------------------------------------------------------------------------------------
  14 . Satisfaccion gestion administrativa                          (   3 categories )
          GASI - Si satis. gestion    GANO - No satis. gestion      GANR - NR satis. gestion
 -------------------------------------------------------------------------------------------
  15 . Toma decisiones basado en PRINCIPIO DE ORALIDAD              (   3 categories )
          POSI - Si principio oralida PONO - No principio oralida   PONR - NR principio oralida
```

```
-------------------------------------------------------------------------------------------
  16 . Administracion de CSJ mejora gestion                        (   3 categories )
           CSSI - Si mejora gest. CSJ    CSNO - No mejora gest. CSJ    CSNR - NR mejora gest. CSJ
-------------------------------------------------------------------------------------------
  17 . Nuevo sistema acusatorio es instrumento eficaz              (   4 categories )
           NASI - Si eficaz sist. acus   NANO - No eficaz sist. acus   NANR - NR eficaz sist. acus
           NANA - NA eficaz sist. acus
-------------------------------------------------------------------------------------------
  18 . Documentacion suficiente fiscalia                           (   4 categories )
           DSSI - Si documentacion suf   DSNO - No documentacion suf   DSNR - NR documentacion suf
           DSNA - NA documentacion suf
-------------------------------------------------------------------------------------------

====================================
       2 variables  (supplementary)
====================================
-------------------------------------------------------------------------------------------
   3 . Edad                                                        ( numerical )
           EDAD - Edad
-------------------------------------------------------------------------------------------
   4 . Tiempo de servicio a la rama judicial                       ( numerical )
           TSER - Tiempo de servicio a
-------------------------------------------------------------------------------------------
```

Trabajo en DtmVic

Para responder a las preguntas utilice el procedimiento DECAT de DtmVic (Lebart 2015). Siga los siguientes pasos:

1. Cree un directorio de trabajo, por ejemplo *Jueces*.
2. Copie los archivos:
   - *JuezDatos.txt*, archivo de datos;
   - *JuezDiccio.txt*, el diccionario de las variables y
3. Corra DtmVic:
   - 2. Create a command file → DECAT → 1. Open a dictionary: *JuezDiccio.txt* → Open a data file: *JuezDatos.txt* → 3. List of variables → Continue
   - Seleccione la(s) variable(s) categórica(s) a describir (tipo de juzgado y región) y las variables descriptoras (Explanatory: las demás exceptuando la clase de opinión) → Continue
   - Seleccione todos los individuos
   - 2. Create a parameter file for ... → Execute

- Lea los resultados en: Main basic numerical results...

4. Escriba un documento de resumen.

Ejercicio

En el archivo de datos hay una variable (19) que es el resultado de un procedimiento de clasificación siguiendo la estrategia sugerida por Lebart et al. (2006) y presentada en el capítulo 6 de este documento. Describa esta variable, que se puede llamar *clases de opinión*, utilizando, para caracterizarla, todas las demás variables del archivo.

Trabajo con R

Utilice la función `cluster.carac{FactoClass}` para realizar el mismo análisis.

Capítulo 2

Análisis en componentes principales (ACP)

El 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 estructura de la "nube de individuos" y detectar grupos de ellos.

2. Describir las relaciones entre las variables. Los textos de análisis multivariado clásicos privilegian este aspecto del análisis. En la versión descriptiva de este texto la descripción de las variables provee las claves para la lectura de las gráficas de los individuos.

3. Reducir la dimensión de la representación. Entre más relación exista entre las variables mayor es la capacidad de síntesis del ACP y unos pocos ejes factoriales podrán resumir a 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. Los comandos de R que

se presentan son para ayudar a entender la lógica del método, ya que para realizar los cálculos y gráficas del ACP existen funciones en varios paquetes de R.

2.1. Ejemplo "Café"

En Duarte et al. (1996) se presenta un experimento, donde 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 12 filas y 16 columnas. El experimento considera tres factores: agregado (sin, maíz, cebada), porcentaje del agregado (20 % y 40 %) y grado de tostación (clara y oscura). Entonces el experimento consta de 10 tratamientos y sobre las tazas de café de cada uno se miden 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 10 primeras filas que corresponden a los tratamientos del experimento. Los valores obtenidos para los 10 tratamientos se muestran en la figura 2.2. Las claves de las etiquetas son: C claro, O oscuro; E excelso, M maíz, C cebada; 20 %, 40 % de agregado; por ejemplo: CM20 quiere decir tostación clara, con agregado de maíz en un 20 %.

La matriz $\mathbf{Y}$ contiene los datos "activos" del ejemplo. Las 10 filas ($n = 10$) de $\mathbf{Y}$ se representan como puntos en $\mathbb{R}^3$ ($p = 3$), imagen que se denomina *nube de individuos* (figura 2.2, izquierda). Las columnas de $\mathbf{Y}$ 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 3 vectores (*color, DA y EA*) constituyen la *nube de variables*.

2.2. 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 (fila de $\mathbf{Y}$). En la figura 2.2, izquierda, se muestra en 3D la nube de los 10 individuos del ejemplo Café.

2.2.1. Centro de gravedad

Sobre la nube de individuos se define el *centro de gravedad*, notado $\mathbf{g}$, que generaliza el concepto de media como una medida de localización multivariada. Cuando los individuos tienen el mismo peso $(1/n)$, el centro de gravedad es la suma de los n vectores individuo, notados $\mathbf{y}_i$, multiplicada por el escalar $1/n$:

$$\mathbf{g} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i \tag{2.1}$$

El centro de gravedad se constituye en un individuo artificial, denominado típico porque es el punto de referencia para comparar a los demás. 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 y por lo tanto es necesario registrar esos valores, que son los promedios de las variables.

Código R. Para cargar los datos y construir la gráfica:

```
library(FactoClass)
data(cafe)
Y <- cafe[1:10,1:3]
# gráfica con el centro de gravedad
par(las=1)
Y3D <- scatterplot3d (Y, main ="Y",type="h",color ="blue",box=FALSE
   ,las=1) # grafica
Y3D$points3d(Y,pch=1)
addgrids3d(Y, grid = c("xy", "xz", "yz"))
cord2d<-Y3D$xyz.convert(Y) # convertir cordenadas 3D a 2D
text(cord2d,labels=rownames(Y), cex=0.8,col="blue",pos=3) # poner
   etiquetas
# Centro de gravedad
n<-nrow(Y);n
g<-(1/n)*rowSums(t(Y));g # igual a colMeans(Y)
# para incluir g en la grafica
Y3D$points3d(t(g),pch=19,col="darkgreen",type = "h")
text(Y3D$xyz.convert(t(g)),labels="g",pos=3,col="darkgreen")
# para tabular de LaTeX
xtable(Y,digits=c(0,0,1,0))
```

2.2.2. Centrado de la nube de individuos

La coordenada de un individuo centrado $\mathbf{y_{C}}_i$, se obtiene restándole las coordenadas del centro de gravedad $\mathbf{g}$:

$$\mathbf{y_{C}}_i = \mathbf{y}_i - \mathbf{g} \tag{2.2}$$

La matriz de datos centrados $\mathbf{Y_C}$ se obtiene mediante:

$$\mathbf{Y_C} = \mathbf{Y} - \mathbf{1_n g'} \tag{2.3}$$

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

Al representar $\mathbf{Y_C}$ en $\mathbb{R}^p$ el origen se traslada al centro de gravedad de N_n, hecho que se muestra en el esquema de la en las figura 2.1.

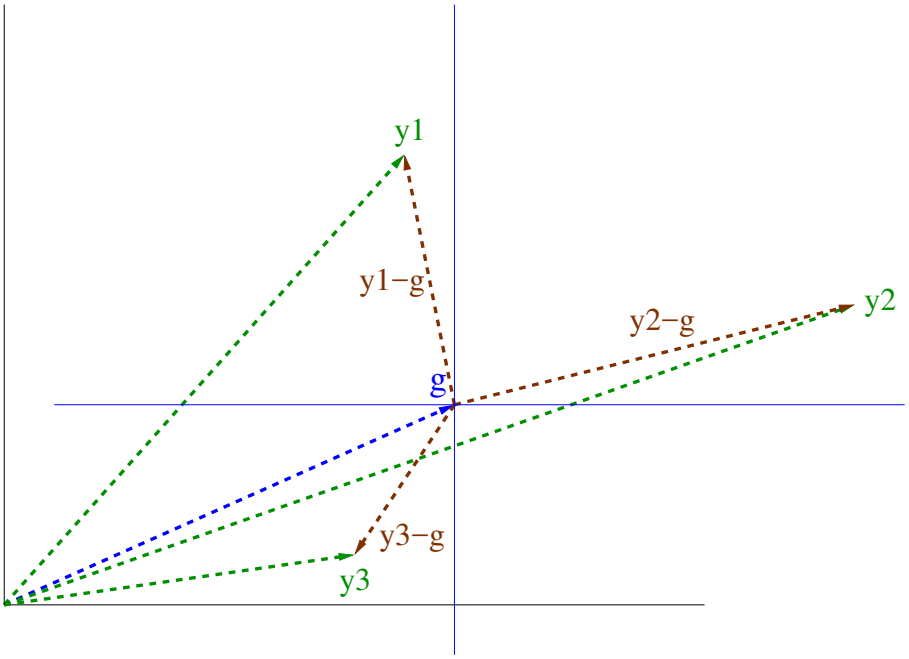


Figura 2.1: Centrado de los individuos en ACP: para representar los puntos $\mathbf{y}_i - \mathbf{g}$, el cero del sistema de coordenadas se traslada a $\mathbf{g}$.

En la figura 2.2, derecha, 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. Los valores para el café típico son 276.70, 401.16 y 35.50, respectivamente, es decir las coordenadas del centro de gravedad en la representación sin centrar. Esta información hay que registrarla porque al centrar los datos se pierde.

Código R. Para centrar los datos y elaborar la gráfica:

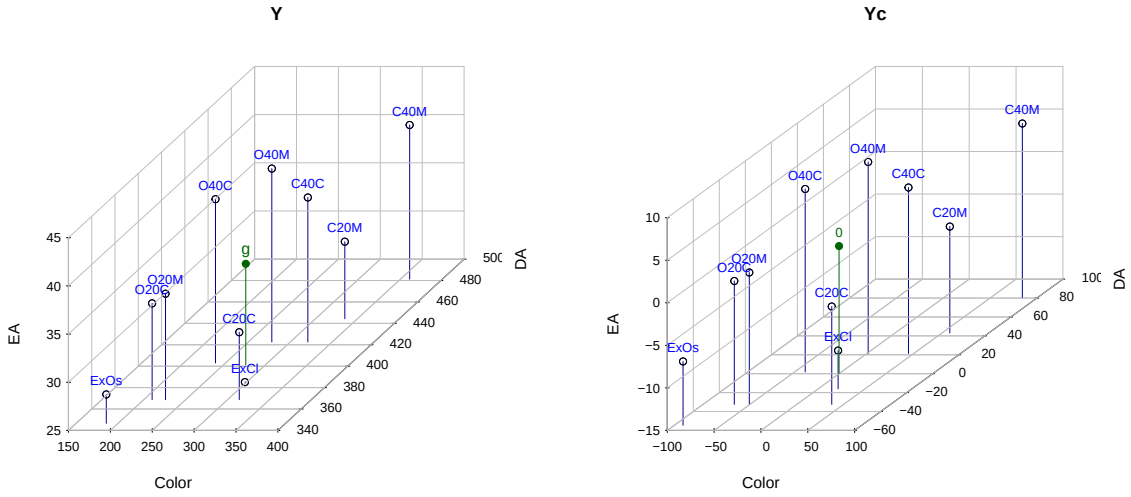
```
par(las=1)
unos <-rep (1,n) # vector de n unos
Yc <-Y- unos %*%t(g);Yc
# grafica de datos centrados
Yc3D <- scatterplot3d(Yc , main ="Yc",type ="h",color ="blue",box=
   FALSE,las=1)
Yc3D$points3d(Yc,pch=1)
addgrids3d(Yc, grid = c("xy", "xz", "yz"))
text (Yc3D$xyz.convert (Yc),labels = rownames (Yc), cex =0.8 , col
   =" blue ",pos =3)
Yc3D$points3d (t(c(0 ,0 ,0)),pch =19 , col =" darkgreen ",type = "h
   ")
text(Yc3D$xyz.convert (t(c(0 ,0 ,0))),labels ="0",pos =3, col ="
   darkgreen ",cex =0.8)
# para tabular de LaTeX
xtable(Yc , digits =rep (1 ,4))
# matriz de distancias
xtable(as.matrix(dist(Y)),digits=rep(1,11))
```

2.2.3. Distancia entre individuos

El parecido entre individuos en la tabla de datos se traslada a la representación geométrica como una distancia (ver figura 2.3). La distancia euclidiana al cuadrado entre dos individuos es:

$$d^2(i,l) = \sum_{j=1}^{p} (y_{ij} - y_{lj})^2 \tag{2.4}$$

Su resultado es igual si se calcula a partir de la matriz de datos centrados. En la figura 2.2 aparecen las distancias entre los 10 cafés. Una distancia de cero indicaría que los dos cafés tienen los mismos valores para las variables. En la gráfica, la pareja de cafés más alejados son $ExOs$ y $C40M$, la distancia entre ellos es 221, la mayor en la tabla. Los más próximos con $C20C$ y $C20M$, con una distancia de 16.



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

$\mathbf{Y}_C$

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

Distancias entre cafés

	ExCl	C40M	C40C	C20M	C20C	ExOs	O40M	O40C	O20M	O20C
ExCl	0	116	46	70	24	118	46	65	75	90
C40M	116	0	71	46	122	221	102	146	176	188
C40C	46	71	0	27	55	155	43	85	109	123
C20M	70	46	27	0	78	178	62	106	133	146
C20C	24	122	55	78	0	130	66	84	88	104
ExOs	118	221	155	178	130	0	120	78	46	33
O40M	46	102	43	62	66	120	0	45	75	87
O40C	65	146	85	106	84	78	45	0	37	45
O20M	75	176	109	133	88	46	75	37	0	16
O20C	90	188	123	146	104	33	87	45	16	0

Figura 2.2: Representación de la tabla de datos del ejemplo Café en 3D. Visualmente las dos figuras son iguales, lo que cambian son las coordenadas sobre los ejes, cuyos valores están en las tablas. Las dos gráficas están representando, también, las distancias entre los cafés.

2.2.4. Inercia de la nube de individuos N_n

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

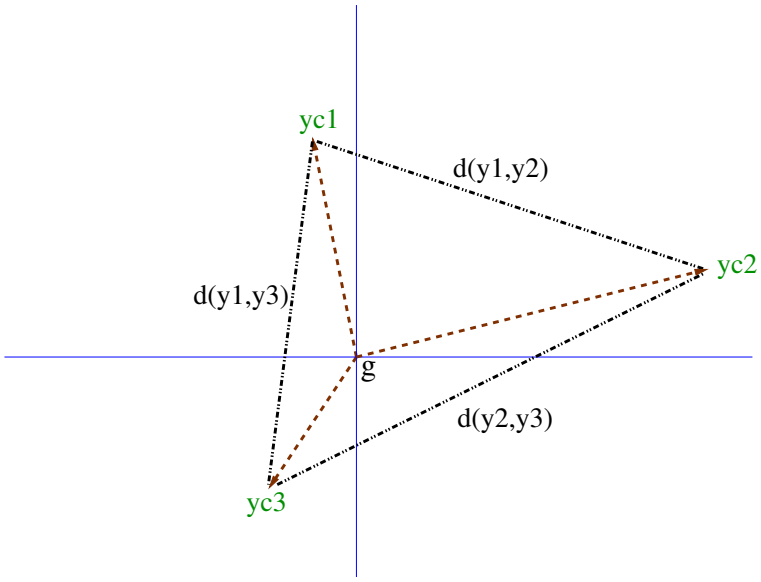


Figura 2.3: Distancias entre individuos.

Si cada individuo i se dota del peso p_i, la inercia de la nube es:

$$Inercia(N_n) = \sum_{i=1}^{n} p_i d^2(i, \mathbf{g}) \tag{2.5}$$

En el caso de pesos iguales para todos los individuos $p_i = 1/n$ y dado que su centro de gravedad se ha trasladado al origen, entonces:

$$Inercia(N_I) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} y_{Cij}^2 = \sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n} y_{Cij}^2 = \sum_{j=1}^{p}\sigma_j^2 \tag{2.6}$$

La información que queda en la nube de puntos es su forma, que está dada por las relaciones de distancias entre los puntos. La fórmula 2.6 permite ver que la inercia es la suma de las varianzas de las variables, por lo que 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 es dividir cada columna de la matriz de datos centrados por la desviación estándar de la variable correspondiente.

2.2.5. Reducción de la nube de puntos (cambio de escala)

La matriz de varianzas y covarianzas $\mathbf{V}$ asociada a la tabla $\mathbf{Y}$ es: $\mathbf{V} = \frac{1}{n}\mathbf{Y_C'Y_C}$. 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 la media y la desviación estándar de la variable j. En términos matriciales $\mathbf{X}$ se obtiene mediante:

$$\mathbf{X} = \mathbf{Y_C D_\sigma^{-1}} \tag{2.7}$$

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

Código R. Para estandarizar los datos, una vez centrados y hacer la gráfica de $\mathbf{X}$:

```
par(las=1)
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)
# gráfica
X3D <- scatterplot3d(X , main ="X",type ="h",color ="blue",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 , col ="
   blue",pos =3)
X3D$points3d (t(c(0 ,0 ,0)),pch =19 , col ="␣darkgreen␣",type = "h"
   )
text(X3D$xyz.convert (t(c(0 ,0 ,0))),labels ="0",pos =3, col ="
   darkgreen␣",cex =0.8)
# tabla para LaTeX
xtable (X , digits =rep (1 ,4))
```

El valor que un individuo asume para una variable es la diferencia con respecto al promedio, pero ahora medida en el número de desviaciones estándar. Al reducir los datos, la información de las varianzas de las variables se pierde en las gráficas y es necesario también registrarla. 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 2.4 se muestra la gráfica 3D para el ejemplo Café, junto con los valores de $\mathbf{X}$ y las distancias entre cafés.

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 2.11, abajo a la derecha.

La inercia de la nube de puntos, cuando los datos se han estandarizado, es igual al número de variables, ya que cada una de ellas contribuye con 1 a la inercia total. En el ejemplo la inercia es 3. 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.

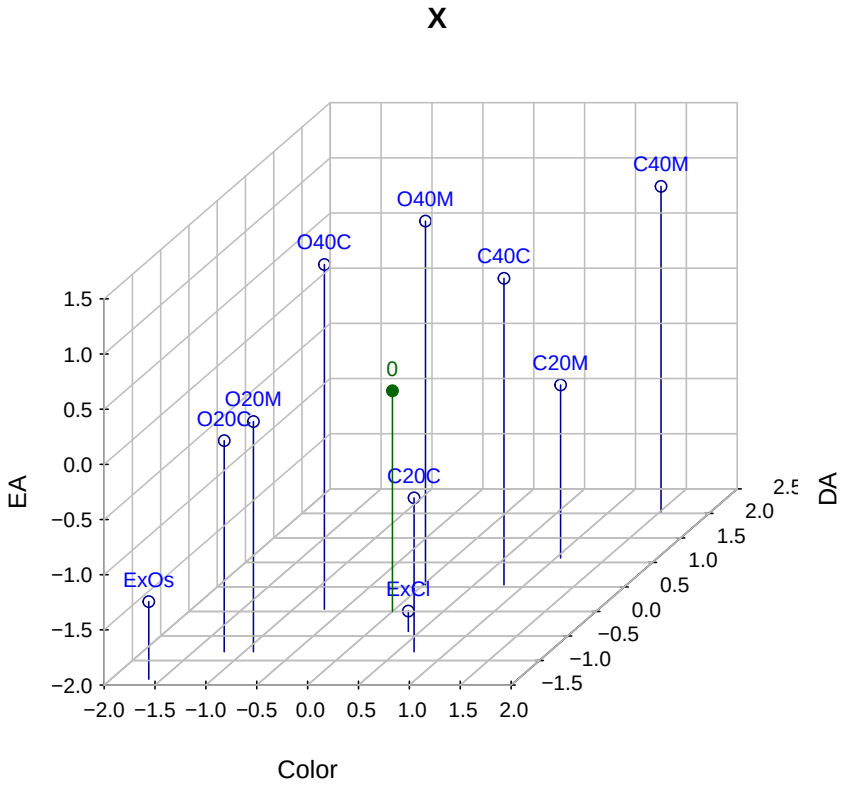
2.2.6. Búsqueda de nuevos ejes: cambio de base

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 tal manera 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 2.5):

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

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



Coordenadas de cafés estandarizados

X

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

Distancias entre cafés estandarizados

	ExC	C4M	C4C	C2M	C2C	ExO	O4M	O4C	O2M	O2C
ExCl	0.0	3.9	2.8	2.1	1.3	2.3	3.3	3.2	2.3	2.4
C40M	3.9	0.0	1.7	1.7	3.4	5.2	2.1	3.0	3.8	4.1
C40C	2.8	1.7	0.0	1.4	1.9	3.7	0.9	1.6	2.3	2.6
C20M	2.1	1.7	1.4	0.0	2.0	3.7	2.1	2.6	2.8	3.0
C20C	1.3	3.4	1.9	2.0	0.0	2.5	2.4	2.4	1.7	1.9
ExOs	2.3	5.2	3.7	3.7	2.5	0.0	3.6	3.0	1.7	1.4
O40M	3.3	2.1	0.9	2.1	2.4	3.6	0.0	0.9	2.0	2.3
O40C	3.2	3.0	1.6	2.6	2.4	3.0	0.9	0.0	1.4	1.6
O20M	2.3	3.8	2.3	2.8	1.7	1.7	2.0	1.4	0.0	0.3
O20C	2.4	4.1	2.6	3.0	1.9	1.4	2.3	1.6	0.3	0.0

Figura 2.4: 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. Los cafés más alejados son $ExOs$ y $C40M$, distancia 5.2; los más cercanos $O20C$ y $O20M$, distancia 0.3

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

$$\sum_{i=1}^{n} \frac{1}{n}(\mathbf{x}_i'\mathbf{u})^2 = \frac{1}{n}(\mathbf{Xu})'\mathbf{Xu} = \mathbf{u}'\frac{1}{n}\mathbf{X}'\mathbf{Xu} \tag{2.8}$$

Encontrar el eje de mayor inercia proyectada equivale a encontrar la dirección $\mathbf{u}$ que maximice (2.8) sujeto a la restricción $\mathbf{u}'\mathbf{u} = 1$. Una manera de resolver este problema es

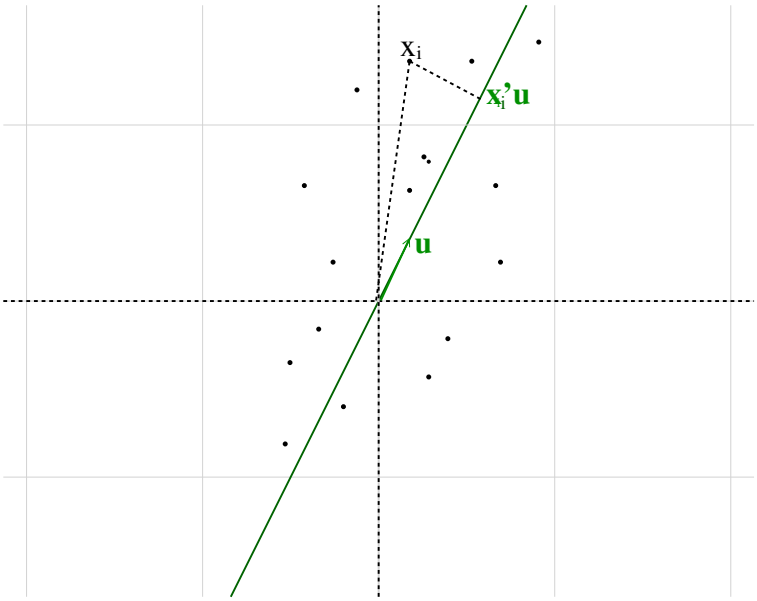


Figura 2.5: Proyección sobre la recta generada por $\mathbf{u}$. Se busca la dirección de la recta $\mathbf{u}$ que tenga la suma de los cuadrados de las proyecciones mayor.

introduciendo un multiplicador de Lagrange λ, entonces se debe maximizar:

$$f(\mathbf{u}) = \mathbf{u}'\frac{1}{n}\mathbf{X}'\mathbf{X}\mathbf{u} - \lambda(\mathbf{u}'\mathbf{u} - 1) \tag{2.9}$$

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:

$$\frac{1}{n}\mathbf{X}'\mathbf{X}\mathbf{u} = \lambda\mathbf{u} \tag{2.10}$$

Las soluciones de (2.10) 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 por $\mathbf{u}'$ se obtiene de nuevo la cantidad que se quiere maximizar (2.8):

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

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 (2.8). El problema ahora es maximizar $\frac{1}{n}\mathbf{X}'\mathbf{X}$ 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:

$$f'(\mathbf{u}) = 2\frac{1}{n}\mathbf{X}'\mathbf{X}\mathbf{u} - 2\lambda\mathbf{u} - \mu\mathbf{u}_1 = \mathbf{0} \tag{2.11}$$

El segundo multiplicador μ debe ser 0, lo que se puede ver premultiplicando la ecuación (2.11) por $\mathbf{u}_1'$. Entonces se obtiene de nuevo la ecuación (2.10) 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 se requieren para obtener las mejores proyecciones de la nube de puntos. Es decir, el mejor eje para proyectar los puntos es el eje 1, el cual es 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:

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

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.

Código R. Para obtener la matriz de correlaciones, a partir de los datos estandarizados, y sus valores y vectores propios:

```
V <- t(X) %*% X/n;V # calculo de matriz correlaciones
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, los valores propios son: $\lambda_1 = 2.067$, $\lambda_2 = 0.822$ y $\lambda_3 = 0.111$ y los vectores propios:

$\mathbf{u_1} = (0.58 \;\; 0.67 \;\; 0.46)'$, $\mathbf{u_2} = (-0.57 \;\; -0.07 \;\; 0.82)'$ y $\mathbf{u_3} = (0.58 \;\; -0.74 \;\; 0.35)'$ 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.

En el ejemplo Café, la primera componente principal es una variable nueva, que resume

las tres propiedades físicas y su expresión es:

$$
\begin{aligned}
F_1 &= 0.58X_{color} + 0.67X_{DA} + 0.46X_{EA} \\
F_1 &= 0.58\left(\frac{Y_{color} - \overline{Y}_{color}}{\sigma_{color}}\right) + 0.67\left(\frac{Y_{DA} - \overline{Y}_{DA}}{\sigma_{DA}}\right) + 0.46\left(\frac{Y_{EA} - \overline{Y}_{EA}}{\sigma_{EA}}\right) \\
F_1 &= 0.58\left(\frac{Color - 276.7}{55.7}\right) + 0.67\left(\frac{DA - 401.2}{39.5}\right) + 0.46\left(\frac{EA - 35.5}{5.8}\right) \\
F_1 &= 0.0104Color + 0.0170DA + 0.0795EA - 12.5
\end{aligned}
$$

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 10 preparaciones de cafés, así:

```
>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 café ExCl se debe a errores
    de redondeo.
```

En la figura 2.6 se puede ver el orden de los cafés de izquierda a derecha.

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 las todas coordenadas sobre un eje cuando le convenga.

2.2.7. 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 2.6, donde se incluyen los valores de las coordenadas y ayudas a la interpretación.

Código R. Para dibujar el plano factorial y la tabla que se muestran en la figura 2.6:

```
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="blue",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 interpretación
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 en entorno tabular de LaTeX
```

La interpretación 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, mientras que al lado positivo del segundo eje los de mayor valor en *EA* y al lado negativo los de mayor valor en *Color*.

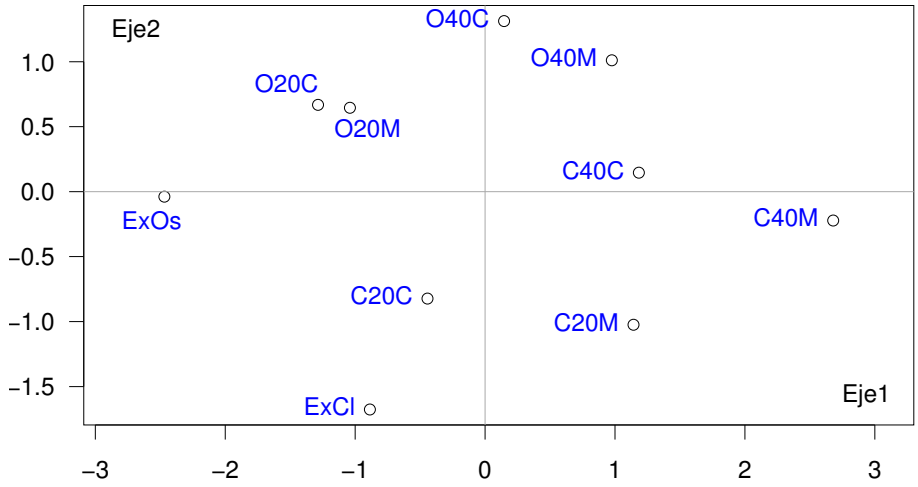
Distanca al origen

La distancia de un punto al origen, en el espacio completo, es un buen complemento en la lectura de los ejes factoriales, está dada por la norma del vector-individuo en $\mathbb{R}^p$. En las salidas de algunos programas, se presenta la distancia al cuadrado.

$$d^2(i, \mathbf{g}) = d^2(i, \mathbf{0}) = \|\mathbf{x}_i\|^2$$

Calidad de la representación o coseno cuadrado

Un plano factorial es una aproximación de la nube de puntos y como tal tendrá puntos bien representados, pero podrá contener puntos con mala calidad de proyección. La calidad de la proyección o representación sobre un eje se mide con el coseno cuadrado que se define,



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 2.6: 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 2.4, 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 10 cafés; en el primer eje el $O40C$ está mal representado.

para un punto, como el cuadrado de la relación entre la norma de la proyección sobre 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.

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

Por ejemplo el coseno cuadrado del café $CM40$ sobre el primer eje es:

$$Cos_1^2(CM40) = \frac{2.68^2}{7.31} = 0.98$$

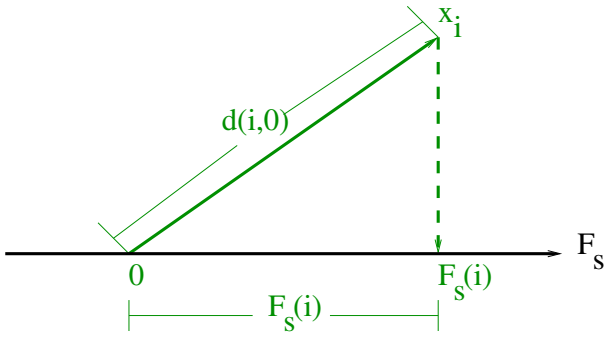


Figura 2.7: Calidad de la proyección sobre un eje s: cociente al cuadrado entre $F_s(i)$ y $d(i,0)$. Cuando se acerca a 1, la longitud de la proyección se aproxima a la distancia original; si se acerca a cero, la proyección conserva muy poco de la distancia original.

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) y se denomina contribución absoluta $Ca_s(i)$. La suma de las contribuciones de todos los individuos es 1.

$$Ca_s(i) = \frac{p_i F_s^2(i)}{\lambda_s}; \qquad \sum_i Ca_s(i) = 1$$

Por ejemplo, la contribución del café $CM40$ a la varianza del primer eje es:

$$Ca_1(CM40) = \frac{1}{10} * \frac{2.68^2}{2.067} = 0.347 = 34.7\,\%$$

En la figura 2.6, abajo, se encuentran los valores de las ayudas a la interpretación para el ejemplo Café.

2.2.8. 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 proyectando sobre la recta generada por el vector propio $\mathbf{u}_s$. Es decir $F_s(i^+) = \mathbf{x}'_{i^+}\mathbf{u}_s$, utilizando el signo + para indicar que es un individuo suplementario.

En el ejemplo de Café, se prepararon tazas con dos cafés comerciales y se le 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 2.8), 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 proyección.

Código R. Para el cálculo de las coordenadas factoriales de los dos cafés comerciales y su proyección sobre el primer plano factorial:

```
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="blue",pos=1)
abline(h=0,v=0,col="darkgrey")
points(Fsup,col="darkgreen",pch=20)  # cafes comerciales
text(Fsup,labels=c("Com1","Com2"),col="darkgreen",pos=2)
```

2.2.9. 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 la variable tipo de contaminación: excelso (sin contaminación), con cebada y con maíz.

Código R. Para calcular las coordenadas y proyectarlas sobre el primer plano factorial:

```
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)
# dev.print(device = xfig)
```

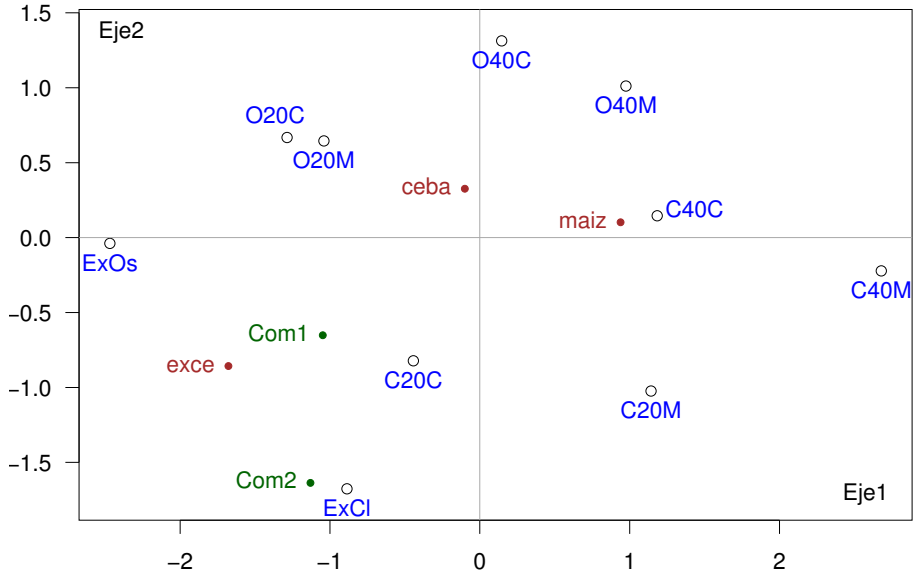


Figura 2.8: Primer plano factorial del ACP del ejemplo Café, mostrando dos cafés comerciales y las categorías del tipo de contaminante. Los cafés comerciales se ponen en el marco de referencia del experimento, su posición permite decir que son de buena calidad (están cerca de los cafés no contaminados). La posiciones de los centros de gravedad: *excelso*, *cebada* y *maiz*, muestran que el maíz afecta más la calidad del café.

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 1.9.2, página 21, 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 son centrados 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 ecuación (1.2):

$$t_s(k) = \frac{F_s(k)}{\sigma_s(k)} \quad con \quad \sigma_s^2(k) = \frac{n-n_k}{n-1}\frac{\lambda_s}{n_k} \Longrightarrow t_s(k) = \sqrt{\frac{(n-1)n_k}{(n-n_k)\lambda_s}}\, F_s(k) \tag{2.12}$$

2.3. 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.

2.3.1. Significado de la media y del centrado de una variable 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.

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.

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

Significado de la media de una variable j

La media $\bar{Y}_j$ es:

$$\bar{Y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} = \frac{1}{n}\mathbf{Y}'_j\mathbf{1_n} = \langle \mathbf{Y}_j, \mathbf{1_n} \rangle_{\frac{1}{n}\mathbf{I}_n} \tag{2.13}$$

La media es entonces la coordenada de la proyección de la variable sobre la primera bisectriz, es decir la recta generada por el vector $\mathbf{1}_n$ (ver figura 2.9). Se puede definir un vector que repite el valor de la media n veces: $\bar{\mathbf{Y}}_j = \bar{Y}_j\mathbf{1_n}$.

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 (ver figura 2.9).

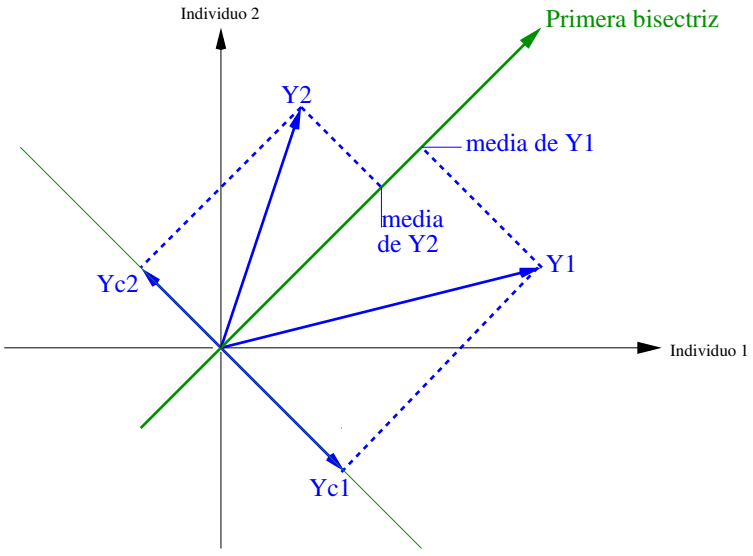


Figura 2.9: Significado geométrico de las medias y del centrado de las variables: el valor de la media de una variable es la coordenada de su proyección sobre la primera bisectriz. El vector centrado de la variable está contenido en el subespacio $\mathbb{R}^{n-1}$ ortogonal a la primera bisectriz. En está gráfica en $\mathbb{R}^2$ ese subespacio es la recta que pasa por (-1,1).

2.3.2. Significado de las varianzas y covarianzas

La varianza de una variable j se puede expresar como:

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

igual a la la norma al cuadrado del vector variable centrado. Es decir que la desviación estándar de una variable es: $\sigma_j = \|\mathbf{Y_C}_j\|_{\frac{1}{n}\mathbf{I}_n}$, corresponde a la norma del vector variable centrado.

La covarianza entre dos variables Y_j y Y_k es:

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

es decir el producto punto entre los dos vectores que representan a las variables centradas.

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

Una columna j de $\mathbf{X}$ 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. En la fórmula (2.7) se muestra el cálculo de la matriz $\mathbf{X}$ y, debajo de ella, un código en R para hacerlo.

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.

2.3.4. Significado de la correlación entre dos variables

La correlación entre dos variables j y k es:

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

Es decir que la correlación entre dos variables es igual al coseno entre los dos vectores variables centradas y es también el coseno entre los dos vectores variables centradas y reducidas: $< \mathbf{X_j}, \mathbf{X_k} >_{\frac{1}{n}\mathbf{I_n}}$, ya que $\|\mathbf{X_j}\|_{\frac{1}{n}\mathbf{I_n}} = 1$ (norma=desviación estándar = 1).

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

2.3.5. 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:

$$Inercia(N_p) = \sum_{j=1}^{p} d^2(\mathbf{Y_{Cj}}, \mathbf{0}) = \sum_{j=1}^{p} Var(Y_j) \tag{2.14}$$

La contribución de una variable a la inercia es su varianza.

En caso del ACP normado cada variable contribuye con 1 a la inercia y la inercia total es igual al número de variables.

2.3.6. Búsqueda de los nuevos ejes

La proyección de una variable $\mathbf{X}_j$ sobre un eje $\mathbf{v}$ en $\mathbb{R}^n$ es: $\mathbf{X'}_j \left(\frac{1}{n}\mathbf{I_n}\right) \mathbf{v} = \frac{1}{n}\mathbf{X'}_j\mathbf{v}$.

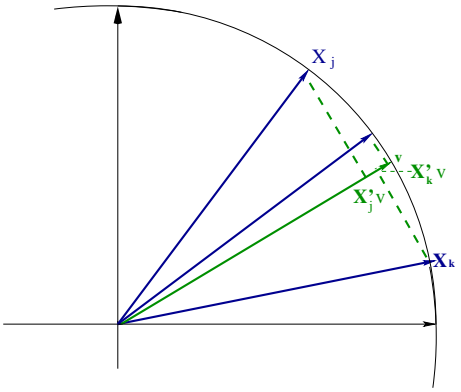


Figura 2.10: Proyección de variables sobre el eje generado por $\mathbf{v}$. La dirección de $\mathbf{v}$ es la maximiza la suma de cuadrados de las proyecciones de los vectores variables sobre $\mathbf{v}$.

La inercia de las p variables proyectadas sobre el eje $\mathbf{v}$ es:

$$\sum_{j=1}^{p}\left(\frac{1}{n}\mathbf{X'}_j\mathbf{v}\right)^2 = \frac{1}{n^2}\mathbf{v'XX'v} = \frac{1}{n}\mathbf{v'}\frac{1}{n}\mathbf{XX'v} \tag{2.15}$$

El eje de mayor inercia proyectada se encuentra maximizando (2.15) sujeto a la restricción $\mathbf{v'}\left(\frac{1}{n}\mathbf{I_n}\right)\mathbf{v} = 1$.

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

derivando con respecto al vector $\mathbf{v}$ e igualando a 0:

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

se obtiene:

$$\frac{1}{n}\mathbf{XX'v} = \mu\mathbf{v} \tag{2.16}$$

La ecuación (2.16) corresponde a la expresión de valores y vectores propios de la matriz $\frac{1}{n}\mathbf{XX'}$ y por lo tanto, la solución está dada por uno de los dos vectores $\mathbf{v}$ asociados al valor propio más grande μ de la matriz $\frac{1}{n}\mathbf{XX'}$, que se notan $\mathbf{v}_1$ y μ_1, respectivamente. Sin embargo, los valores propios de $\frac{1}{n}\mathbf{XX'}$ que son mayores que cero, son iguales a los de $\frac{1}{n}\mathbf{X'X}$, 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 $\frac{1}{n}\mathbf{XX}'$.

2.3.7. Círculo de correlaciones y ayudas a la interpretación

Un plano factorial de las variables estandarizadas se denomina círculo de correlaciones (figura 2.11), 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 visualmente al dibujar un círculo de radio uno en el plano factorial.

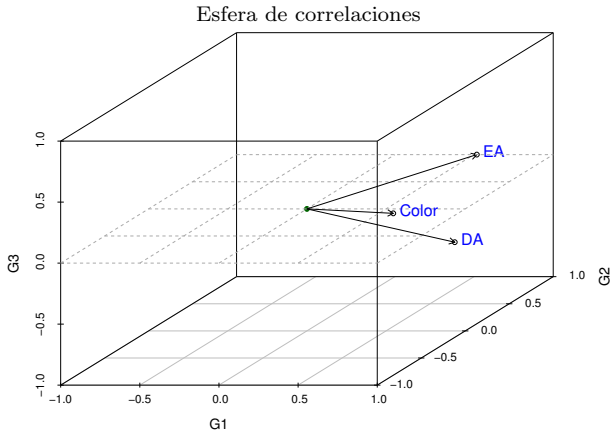
El vector de coordenadas sobre un eje s, $\mathbf{G_s}$, se obtiene mediante $\frac{1}{n}\mathbf{X_j'v_s}$ y coincide con la correlación entre la variable j y el eje s. Los valores de la variable sintética representada por el eje s están en el vector F_s, que contiene las coordenadas de los individuos sobre el eje factorial s.

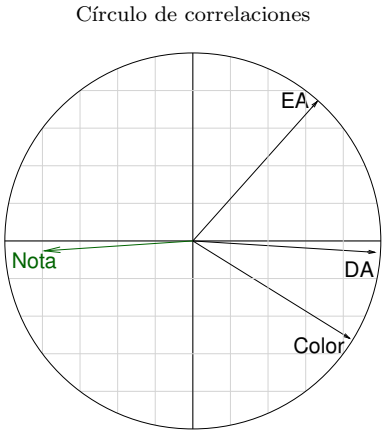
La contribución de cada variable a un eje s sirve para seleccionar las variables que dan más significado al eje. En la figura 2.11 se muestra la esfera y el círculo de correlaciones, la matriz de correlaciones, las coordenadas y las ayudas para la interpretación de las variables, para el ejemplo café.

Código R. Para obtener la esfera y el círculo de correlaciones (figura 2.11), 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)
G3D <- scatterplot3d(G,main="G",xlim=c(-1,1),ylim=c(-1,1),zlim=c
   (-1,1))
coord <- G3D$xyz.convert(G)
text(coord,labels=rownames(G), cex=0.8,col="blue",pos=4)
G3D$plane(0,0,0,col="darkgrey")
G3D$points3d(t(c(0,0,0)),pch=19,col="darkgreen")
cero <- G3D$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") # grabar gráfica en
   xfig
s.corcircle(G,clabel=2)
# proyección de nota como variable ilustrativa
Nota <- cafe[1:10,16]; Nota
```

```
Fnota <- cor(Nota,F);Fnota
arrows(0,0,Fnota[1],Fnota[2],col="darkgreen",angle=10)
text(Fnota,"Nota",col="darkgreen",pos=1,cex=2)
dev.print(device = xfig,file="cafeCirculo")  # grabar círculo en
    xfig
```





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

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 2.11: 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.

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.

2.4. Relación entre los espacios de individuos y variables

Las propiedades que se presentan a continuación son fáciles de demostrar, las que se pueden ver, por ejemplo, en el texto de Lebart, Piron & Morineau (2006).

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

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

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: $cor(Y_j, F_s)$.

2.4.1. Variables continuas como suplementarias o 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 2.8).

2.5. Ejemplo de aplicación de ACP

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.

Resultados del examen de admisión de las carreras de la Facultad de Ciencias

Para el primer semestre de 2013 fueron admitidos 445 estudiantes a las carreras de la Facultad de Ciencias. El examen de admisión tuvo 5 componentes temáticos: matemático, científico, social, textual e imagen. Se realiza un ACP con el objeto de: 1) validar el puntaje total que es el resumen de los 5 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.

Para cumplir con esos objetivos se realiza una 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.

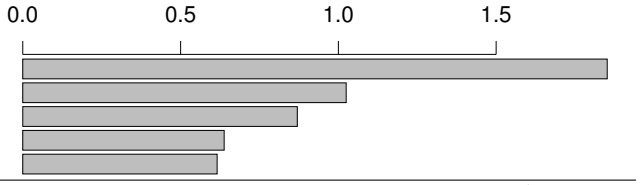
2.5.1. Número de ejes a analizar

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, seleccionar un eje adicional si se considera que suministra información importante que no se ha visto en los anteriores. En la figura 2.12 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 del 57.5 % de la inercia y los tres primeros ejes el 74.9 %.

Código R. Para realizar ACP normado de las notas del examen y obtener la figura 2.12:

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

```
valp<-t(inertia.dudi(acp)$TOT)
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 2.12: 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 uno, el primer plano retiene el 57.5 % de la inercia. La forma del histograma sugiere retener uno o tres ejes para el análisis.

2.5.2. Círculo de correlaciones

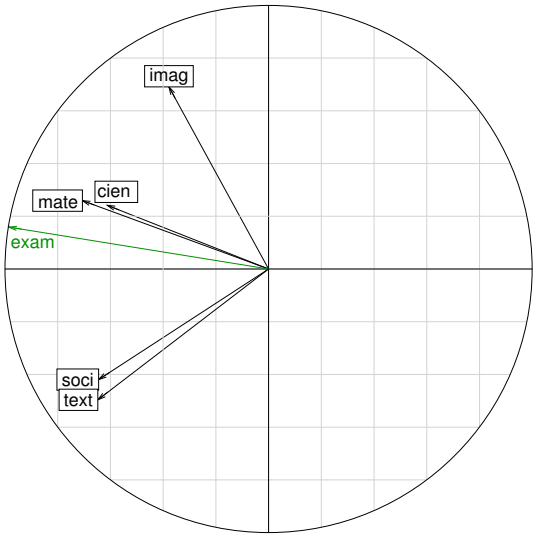
En la figura 2.13 se observa un primer eje de tamaño con alta correlación con el puntaje total y con los mejores puntajes con coordenadas negativas (lado izquierdo del eje). El factor tamaño se presenta cuando todas las correlaciones entre las variables activas son positivas y se observan porque tienen 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 cambiarle de signo a todas las coordenadas sobre el primer eje.

El segundo eje contrapone los resultados 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 imagen (positivo) y científico (negativo).

Código R. Para obtener el círculo de correlaciones (figura 2.13):

```
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 2.13: 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.

2.5.3. Primer plano factorial 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 2.14 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.

Código R. Para obtener la figura 2.14 y la tabla 2.1:

```
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$coor,col="darkgreen")
text(sup$coor,labels=rownames(sup$coor),col="darkgreen",pos=1)
#dev.print(device = xfig,file="acpExaAdmiCatSup.fig")
# tabla de coordenadas y ayudas para la interpretación
xtable(cbind(wcat=sup$wcat,d2=sup$d2,sup$coor,sup$t,sup$cos2),
    digits = rep(3,12))
```

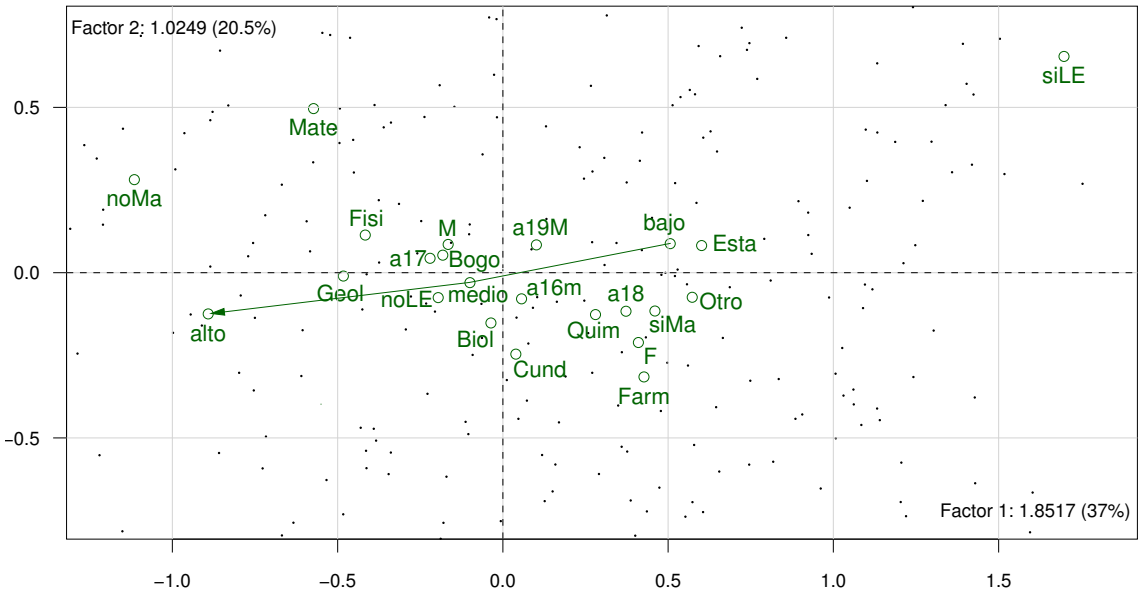


Figura 2.14: Primer plano factorial de los admitidos mostrando las variables cualitativas ilustrativas. Los puntos sin etiqueta corresponden a los admitidos.

Tabla 2.1: Coordenadas, valores test y cosenos cuadrados de las categorías suplementarias

Cate.	Peso	dis^2	Coordenadas			Valores test			Cosenos cuadrados		
			Eje1	Eje2	Eje3	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3
Biol	0.142	0.117	-0.036	-0.152	-0.080	-0.228	-1.287	-0.730	0.011	0.198	0.054
Esta	0.148	0.403	0.601	0.082	-0.109	3.885	0.711	-1.024	0.897	0.017	0.029
Farm	0.164	0.346	0.427	-0.315	0.068	2.928	-2.908	0.684	0.526	0.287	0.013
Fisi	0.184	0.214	-0.416	0.114	0.108	-3.064	1.127	1.160	0.812	0.061	0.055
Geol	0.101	0.244	-0.483	-0.010	-0.071	-2.507	-0.071	-0.540	0.956	0.000	0.021
Mate	0.119	0.751	-0.573	0.496	-0.014	-3.261	3.799	-0.114	0.437	0.328	0.000
Quim	0.142	0.106	0.280	-0.127	0.036	1.763	-1.071	0.331	0.745	0.152	0.012
F	0.288	0.229	0.410	-0.211	0.117	4.036	-2.793	1.675	0.733	0.194	0.059
M	0.712	0.037	-0.166	0.085	-0.047	-4.036	2.793	-1.675	0.733	0.194	0.059
bajo	0.402	0.267	0.506	0.088	0.035	6.430	1.496	0.642	0.961	0.029	0.005
medio	0.416	0.021	-0.099	-0.030	-0.068	-1.300	-0.530	-1.292	0.468	0.043	0.217
alto	0.182	0.878	-0.891	-0.125	0.078	-6.512	-1.225	0.834	0.905	0.018	0.007
Bogo	0.699	0.040	-0.182	0.053	-0.062	-4.286	1.685	-2.129	0.828	0.071	0.096
Cund	0.085	0.080	0.039	-0.246	-0.025	0.186	-1.567	-0.174	0.019	0.755	0.008
Otro	0.216	0.382	0.573	-0.075	0.210	4.654	-0.814	2.493	0.860	0.015	0.116
a16m	0.265	0.052	0.057	-0.079	0.195	0.526	-0.989	2.649	0.062	0.121	0.738
a17	0.384	0.079	-0.220	0.044	-0.118	-2.697	0.717	-2.115	0.616	0.024	0.178
a18	0.126	0.162	0.372	-0.117	-0.068	2.188	-0.921	-0.585	0.855	0.084	0.029
a19M	0.225	0.059	0.102	0.084	0.011	0.847	0.942	0.128	0.176	0.120	0.002
siLE	0.103	3.636	1.698	0.654	0.388	8.927	4.624	2.973	0.793	0.118	0.041
noLE	0.897	0.048	-0.196	-0.075	-0.045	-8.927	-4.624	-2.973	0.793	0.118	0.041
siMa	0.708	0.309	0.460	-0.116	-0.119	11.093	-3.763	-4.171	0.685	0.044	0.046
noMa	0.292	1.814	-1.115	0.281	0.287	-11.093	3.763	4.171	0.685	0.044	0.046

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 17 años quienes tienden a obtener mejores resultados. En género, los hombres tienen 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 otro lugar 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, matemático y científico. Son pocos los admitidos los que tienen que nivelar lecto-escritura (siLE), y por eso se ubican más lejos, al lado derecho arriba, donde se sitúan los de peores resultados en el examen de admisión. En la nivelación de matemáticas la situación es inversa, y los que no tienen que hacerlo están al lado izquierdo, arriba, de mejores resultados en el examen.

En el tercer eje (ver valores test en la tabla 2.1) se observa que en promedio los que tienen edades de 16 años o menos, por un lado, y los que vienen de otra región, por otro, tienen resultados inferiores en la componente de imagen.

2.6. 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:

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

donde $\mathbf{g}$ es el vector centro de gravedad de todos los puntos.

2. Demuestre que el multiplicador de Lagrange μ en la fórmula 2.11 es 0.

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

4. Demuestre que dos vectores propios $\mathbf{u}_s$ y $\mathbf{u}_t$ son ortogonales.

5. Para el ACP normado del ejemplo Café, dibuje las 4 versiones posibles del primer plano factorial, cambiando el sentido de los ejes F_1 y F_2.

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

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

8. Escriba la matriz 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.

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

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

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

12. Calcule las contribuciones a la inercia en $\mathbb{R}^3$ de cada uno de los 10 cafés, en valor y porcentaje.

13. 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 2.6).

14. En R cree un factor con el porcentaje de agregado de granos con los niveles 0, 20 y 40 y proyéctela sobre el primer plano factorial de los cafés.

15. 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.

16. 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 $\mathbf{1}_n$.

17. 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 $\frac{1}{n}\mathbf{I}_n$.

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

19. Muestre que la $\mathbf{M}$ norma de $\mathbf{X}_j$, vector variable centrado y reducido, es 1.

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

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

22. 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 $\frac{1}{n}\mathbf{X}\mathbf{X}'$.

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

24. 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.

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

26. 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: $\tau_S = \frac{\sum_{s=1}^{S} \lambda_s}{\sum_{s=1}^{p} \lambda_s}$

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

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

29. Demostrar que en el espacio de las variables, en el ACP normado, la distancia al cuadrado entre dos variables está entre 0 y 4.

2.7. 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, donde se incluye el código R para realizarlo. En el taller de lactantes se proponen una serie de preguntas sencillas para resolver empleando R u otro programa estadístico.

2.7.1. Análisis en componentes principales gráfico: ACP($\mathbf{Y}$,$\mathbf{I}_2$,$\mathbf{I}_{10}$)

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 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.

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.

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.

11. Calcule los valores y vectores propios de $\mathbf{X}'\mathbf{X}$ y compare con los resultados geométricos.

Contribuciones a la inercia en $\mathbb{R}^2$ y sobre el primer eje. Cosenos cuadrados sobre el primer eje

1. Calcule la inercia de la nube de puntos. Agregue una columna, a la tabla de datos centrados, para registrar la contribución de cada individuo a la inercia y otra para expresar esa contribución en porcentaje. Comente esas contribuciones.

2. Calcule la contribución a la inercia de cada una de las dos variables, en valor y porcentaje y coméntela.

3. Calcule las contribuciones a la inercia proyectada sobre el primer eje de los 10 individuos.

4. Calcule los cosenos cuadrados sobre el primer eje. Comente los valores.

5. Ordene la tabla de datos según las coordenadas sobre el primer eje.

6. Exprese la variable F_1 en función de Y_1 y Y_2.

2.7.2. Ejemplo de ACP: 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 entero). Se dispone además de una variable categórica "categorías", que clasifica las marcas según su contenido de malta (1=Bajo, 2=Estándar, 3=Puro malta) (Fine 1996).

Trabajo

Realice primero un ACP no normado y luego un ACP normado utilizando el software de su preferencia y responda a las preguntas de la siguiente sección. Abajo, se muestra la manera de ejecutar el ACP con el paquete `ade4` de R.

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.

3. Realice el ACP normado, 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 la correlación?. ¿Si 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? ¿Que 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 (lujo, estándar y pura malta).

14. Supongamos que usted desea comprar una botella de Whisky con buena apreciación y que no sea tan cara. De 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?

Resumen del análisis

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

Guía para llevar a cabo el análisis con ade4 R

El paquete `ade4` (Chessel, Dufour & Thioulouse 2004), más antiguo que R, tenía una plataforma propia con gran cantidad de funciones. Afortunadamente los autores decidieron ponerlo como paquete de R. El paquete `FactoClass` (Pardo & Del-Campo 2007) es un complemento especial para `ade4`, en este taller se puede utilizar para graficar el plano factorial de los individuos con la función *plot.dudi* y para producir salidas más organizadas por consola y para documentos en LaTeX con la función *dudi.tex*.

Para leer y preparar los datos, siga el procedimiento:

1. Cree una carpeta de trabajo y copie en ella el archivo `Whisky.txt`
2. Arranque R. Instale los paquetes
3. Cargue los paquetes: `library(FactoClass)` (con esta instrucción se cargan los otros dos).
4. Cambie el directorio a la carpeta de trabajo.
5. Lea la tabla de datos:

   ```
   W<-read.table("Whisky.txt",header=TRUE,row.names=1)
   ```

6. Opcional: cambie la variable categoría por una variable factor:

   ```
   W$Categoria<-factor(W$Categoria,labels=c("Bajo","Estándar","PuroMalta"))
   ```

7. El ACP en `ade4` requiere un `data.frame` que contenga solo las variables activas:

   ```
   Y <- subset(W,select=c(1,2,4,5))
   ```

De ahora en adelante aparecen las instrucciones de R para poder responder a cada pregunta:

1. ACP no normado

   ```
   acpc <-dudi.pca(Y,scale=FALSE)#acp no normado
   ```
   Nota: la consola de R espera una respuesta, mire la ventana gráfica decida (al menos provisionalmente) el número de ejes que desea retener para el análisis, tecléelo en la consola de R y luego Return.

   ```
   inertia.dudi(acpc,,T)#ayuda para las variables
   ```

2. Primer plano factorial de variables en el ACP no normado:

```
s.arrow(acpc$co)#gráfica de variables con ade4
```

3. ACP normado.

```
acp <-dudi.pca(Y)# acp normado
```

4. ¿Cuántos ejes retiene para el análisis? ¿Por qué?. Para sustentar la respuesta ejecute:

```
inertia.dudi(acp)#valores propios
```

```
# gráficas de valores propios
par(mfrow=c(1,2))
    barplot(acp$eig,las=1)
    plot(acp$eig,type='b',las=1)
```

5. ¿Cuál es la variable que más contribuye al primer eje? ¿Cuál la que menos?:

```
inertia.dudi(acp,,T) # debe haber doble coma
dev.new()# para abrir una nueva ventana gráfica
s.corcircle(acp$co) #círculo de correlaciones
```

6. ¿Cuáles son las variables más correlacionadas?. Utilice el círculo de correlaciones y verifique en la matriz: `cor(Y) # matriz de correlaciones`

7. ¿Cuál es la variable mejor representada en el primer plano factorial? ¿Cuál la peor?.

```
inertia.dudi(acp,,T)#debe haber doble coma
```

8. ¿Qué representa el primer eje? ¿Que nombre le asignaría?. ¿Que representa el segundo eje?. `inertia.dudi(acp,,T)#debe haber doble coma`

9. ¿Cuál es el individuo mejor representado en el primer plano factorial? ¿Por qué?. Ubique sobre el gráfico de individuos al peor representado sobre el primer plano factorial.

```
inertia.dudi(acp,T) # ayudas a la interpretación de las filas
dev.new() #otra ventana gráfica
plot(acp,Tcol=FALSE) # individuos sobre el primer plano
#adición de antiguos ejes unitarios
s.arrow(acp$c1,add.plot=TRUE,clabel=0.6)
for (eje in 1:nrow(acp$c1))
abline(0,acp$c1[eje,2]/acp$c1[eje,1],lty=2,col=eje)
```

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): `acp$c1 # vectores propios`

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 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).

```
dev.new() # otra ventana gráfica
plot(acp,Tcol=FALSE)
# crea variable tipo factor
cat <- factor(W$Categoria,labels=c('bajo','estandar','puromalta
    '))
s.class(acp$li,cat,col=c(2:4),add.plot=TRUE)
```

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? `ord2 <-order(acp$li[,2]); W[ord2,]`

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

Ejercicio (opcional): utilice el paquete `scatterplot3d` (Ligges & Mächler 2003) para ver el 3D subespacio principal. Obtenga los planos factoriales 1-2 (primer plano factorial), 1-3 y 2-3 (en todos los casos los puntos deben estar identificados -etiquetados-). Con estos gráficos construya en su mente la imagen 3D y haga un resumen en dos párrafos: uno para individuos y otro para variables. Utilice las salidas del ACP de cualquiera de los paquetes para responder:

1. ¿Qué porcentaje de la inercia se conserva en este subespacio 3D?

2. Para algunos individuos de su interés (al menos 2) indique su calidad de representación en 3D. Identifique los 3 individuos peor representados en el subespacio 3D.

3. Para las variables indique su calidad de representación en el subespacio factorial 3D.

2.7.3. Ejemplo 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 2015).

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. ¿Porque con el ACP se cumplen los objetivos planteados?
2. Describa el bebé promedio según las cinco variables.
3. ¿Cuántos ejes retiene para el análisis? ¿Porque?
4. Según la información disponible ¿cuáles variables son altamente correlacionadas? ¿Cuáles no están correlacionadas?
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 grupos?

9. ¿Se puede decir que hay diferencia entre niños y niñas en este análisis? ¿Cómo son esas diferencias?

10. Escriba un resumen práctico del análisis que satisfaga los objetivos planteados.

Preguntas de lectura en el ejemplo lactantes

Responda a las preguntas siguientes. Los por qué se refieren a explicar la manera como dedujo la respuesta (la ayuda que utilizó, la gráfica que leyó).

1. ¿El ACP realizado es normado o no normado? _____

   ¿Por qué? _____

2. Primer valor propio: _____

3. Primer vector propio: _____

4. Porcentajes de la inercia en: primer eje __, segundo eje __y primer plano factorial __

5. Correlación entre *tetero* y primer factor: _____

6. Variable que más contribuye al primer eje: _____

   ¿Por qué?. _____

7. ¿Las dos variables menos correlacionadas con tetero son: _____

   ¿Por qué? _____

8. Variable mejor representada en el primer plano factorial: _____

   ¿Por qué? _____

9. ¿Características del bebé promedio: _____

10. Coordenadas del bebé promedio sobre el primer plano factorial: _____

11. Los dos bebés que más tetero consumen son: _____

12. Los cuatro bebés que más leche materna consumen son: _____

Cálculos en el ejemplo de lactantes

1. Para el bebé situado en el extremo superior del primer plano factorial escriba las coordenadas sobre los dos primeros ejes factoriales: _____

2. Escriba el peso del punto que representa al bebé anterior: _____

3. Calcule la contribución del bebé anterior a la inercia del segundo eje factorial: ____

4. y la calidad de representación sobre el primer plano factorial: _____

5. Escriba las coordenadas de los antiguos ejes unitarios de las variables *leche* y *tetero* sobre el primer plano factorial:

   _____.

6. Dibuje los antiguos ejes de *leche* y *tetero* sobre el primer plano factorial, indicando los lados positivos y negativos.

Capítulo 3

ACP generalizado $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$

Cada uno de los métodos en ejes principales (ACP, análisis de correspondencias y otros) se pueden ver como un ACP de una matriz $\mathbf{X}$, que contiene los datos a analizar, transformados de acuerdo al método; definiendo 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 $\mathbb{R}^n$, la matriz que define el producto interno es la identidad de dimensión n, notada $\mathbf{I}_n$. En estas notas se utilizan los términos *matriz de métrica* o, simplemente *métrica* para referirse, en cada caso, a una matriz que define un producto interno, de donde se derivan, para vectores: normas, distancias, proyecciones, cosenos.

En los métodos en ejes principales básicos, 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 $\mathbf{D}$ a la matriz diagonal de pesos de los filas y de métrica en el espacio de las columnas.

La matriz $\mathbf{X}$ está centrada con los pesos dados en $\mathbf{D}$, es decir $\mathbf{g} = \mathbf{X}'\mathbf{D}\mathbf{1}_n = \mathbf{0}$, siendo $\mathbf{1}_n$ un vector columna de n unos. Con la definición de la tripleta $(\mathbf{X}, \mathbf{M}, \mathbf{D})$, el análisis en ejes principales queda completamente determinado y se nota $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$.

El ACP normado del capítulo anterior se puede ver como $ACP\left(\mathbf{Y}_c, diag\left(\frac{1}{\sigma_k^2}\right), \frac{1}{n}\mathbf{I}_n\right)$,

la matriz a analizar es la matriz de datos centrados, la métrica $diag\left(\frac{1}{\sigma_k^2}\right)$ 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. El ACP normado también es el $ACP\left(\mathbf{X}, \mathbf{I}_p, \frac{1}{n}\mathbf{I}_n\right)$ si $\mathbf{X}$ es la matriz de datos estandarizados; o el $ACP(\mathbf{Z}, \mathbf{I}_p, \mathbf{I}_n)$, donde $\mathbf{Z}$= $\frac{1}{\sqrt{n}}\mathbf{X}$.

3.1. 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 $\mathbf{D}$ 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.

3.1.1. Coordenadas y pesos de filas

En $\mathbb{R}^p$ las filas de $\mathbf{X}$ se representan como puntos $\mathbf{x}_i$, 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 $p_i = d_{ii}$ de la matriz $\mathbf{D}$. La suma de los pesos es 1: $\sum_{i=1}^{n} p_i = 1$.

3.1.2. Distancias entre filas

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

$$d^2(i,l) = \sum_{j=1}^{p} m_j(x_{ij} - x_{lj})^2 \tag{3.1}$$

donde $m_j = m_{jj}$, puesto que $\mathbf{M}$ es diagonal.

3.1.3. 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:

$$Inercia = \sum_{i=1}^{n} p_i d^2(\mathbf{i}, \mathbf{0}) = \sum_{i=1}^{n} p_i \sum_{j=1}^{p} m_j x_{ij}^2 = \sum_{i,j} p_i m_j x_{ij}^2 \tag{3.2}$$

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.

3.1.4. 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 $\sum_{i=1}^{n} p_i F^2 = \mathbf{u'MX'DXMu}$. 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 $\mathbf{X'DXM}$. Nótese que la inercia (3.2) es también igual a la traza de la matriz de inercia $\mathbf{X'DXM}$.

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. Encontrándose un nuevo sistema de ejes $\mathbf{u_s}$, que son vectores propios$\mathbf{M}$ unitario, asociados a los valores propios de la matriz $\mathbf{X'DXM}$, 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. La aproximación de la proyección de la nube sobre ese subespacio se suele medir con el cociente de inercias:

$$\tau = \sum_{s=1}^{S} \lambda_s \Big/ \sum_{s=1}^{p} \lambda_s \tag{3.3}$$

3.1.5. 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: $\sum_{i=1}^{n} p_i F_s(i)^2 = \mathbf{u_s' M X' D X M u_s} = \lambda_s$

3.2. 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 $\mathbf{D}$ y los pesos están en la matriz $\mathbf{M}$.

3.2.1. Coordenadas y pesos

Las coordenadas de un punto j son los n valores de la columna $\mathbf{X}_j$, su peso es m_j.

3.2.2. Distancias entre columnas

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

$$d^2(j,k) = \sum_{i=1}^{n} p_i (x_{ij} - x_{ik})^2 \tag{3.4}$$

3.2.3. Inercia de la nube N_p

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

$$Inercia(N_p) = \sum_{j=1}^{p} m_j d^2(\mathbf{j}, \mathbf{0}) = \sum_{j=1}^{p} m_j \sum_{i=1}^{n} p_i x_{ij}^2 = \sum_{i,j} m_j p_i x_{ij}^2 \tag{3.5}$$

Nótese que la inercia de las nubes de los dos espacios es igual.

3.2.4. 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 variable j sobre la dirección $\mathbf{v}$. El vector de todas las proyecciones sobre $\mathbf{v}$ es $\mathbf{X'Dv}$ y la cantidad a maximizar es: $\mathbf{v'DXMX'Dv}$ sujeta a la restricción $\mathbf{v'Dv} = 1$. Sin embargo no es necesario realizar la diagonalización de la matriz $\mathbf{XMX'D}$, ya que los ejes factoriales de los dos espacios están relacionados.

3.3. 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 $\mathbf{X'DXM}$; para un eje s las proyecciones se encuentran mediante $\mathbf{XMu_s}$. No se calculan los valores y vectores propios de la matriz $\mathbf{XMX'D}$ 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. (2006).

Notando un valor propio de la matriz $\mathbf{XMX'D}$ 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 $\mathbf{D}$-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 $\sqrt{\lambda_s}\mathbf{u_s}$

3.3.1. 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:

$$x_{ij} = \sum_s \frac{F_s(i)G_s(j)}{\sqrt{\lambda_s}} \tag{3.6}$$

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

$$\mathbf{X} = \sum_s \frac{1}{\sqrt{\lambda_s}} F_s G'_s = \sum_s \sqrt{\lambda_s}\mathbf{v}_s\mathbf{u}'_s \tag{3.7}$$

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

$$\tilde{\mathbf{X}} = \sum_{s=1}^{S} \frac{1}{\sqrt{\lambda_s}} F_s G'_s = \sum_{s=1}^{S} \sqrt{\lambda_s}\mathbf{v}_s\mathbf{u}'_s \tag{3.8}$$

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

$$\tau = \sum_{s=1}^{S} \lambda_s \Big/ \sum_{s=1}^{p} \lambda_s \tag{3.9}$$

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.

3.3.2. Fórmulas del ACP($\mathbf{X}, \mathbf{M}, \mathbf{D}$)

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 $\mathbf{D}$, 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($\mathbf{X}, \mathbf{M}, \mathbf{D}$) (tabla 3.1).

Tabla 3.1: Fórmulas del ACP($\mathbf{X}, \mathbf{M}, \mathbf{D}$)

Espacio	$\mathbb{R}^p$	$\mathbb{R}^n$
Nube	N_n	N_p
Coordenadas	filas de $\mathbf{X}$: $\mathbf{x}'_i$	columnas de $\mathbf{X}$: $\mathbf{X}_j$
Pesos	diagonal de $\mathbf{D}$: p_i	diagonal de $\mathbf{M}$: m_j
Métrica	$\mathbf{M}$	$\mathbf{D}$
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	$traza(\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M})$	$traza(\mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{D})$
Valor propio	λ_s	λ_s
Vector propio	$\mathbf{u_s}$	$\mathbf{v_s}$
Fórmula valor-vector propio	$\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M}\mathbf{u_s} = \lambda_s\mathbf{u_s}$	$\mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{D}\mathbf{v_s} = \lambda_s\mathbf{v_s}$
Coordenadas factoriales	$F_s(i) = \mathbf{x_i'}\mathbf{M}\mathbf{u_s}$	$G_s(j) = \mathbf{X_j'}\mathbf{D}\mathbf{v_s}$
	$G_s = \mathbf{X}'\mathbf{D}\mathbf{v_s} = \sqrt{\lambda_s}\mathbf{u_s}$	$F_s = \mathbf{X}\mathbf{M}\mathbf{u_s} = \sqrt{\lambda_s}\mathbf{v_s}$
Fórmulas de transición	$G_s = \frac{1}{\sqrt{\lambda_s}}\mathbf{X}'\mathbf{D}F_s$	$F_s = \frac{1}{\sqrt{\lambda_s}}\mathbf{X}\mathbf{M}G_s$
	$F_s(i) = \frac{1}{\sqrt{\lambda_s}}\sum_{j=1}^{p} x_{ij}m_jG_s(j)$	$G_s(j) = \frac{1}{\sqrt{\lambda_s}}\sum_{i=1}^{n} x_{ij}p_iF_s(i)$
Fórmula de reconstitución	$x_{ij} = \sum_s \frac{F_s(i)G_s(j)}{\sqrt{\lambda_s}} = \sum_s \sqrt{\lambda_s}\mathbf{u}_s(i)\mathbf{v}_s(j); \quad \mathbf{X} = \sum_s \sqrt{\lambda_s}\mathbf{v}_s\mathbf{u}'_s$	

Escofier & Pagès (1992, Cap.4)

3.3.3. 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 $\mathbf{D}$, composiciones de éstas y las inversas cuando existen. 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 3.1 muestra el diagrama de dualidad asociado al $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$, que tiene, además, utilidad nemotécnica. Algunos ejemplos 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 $\mathbf{X'DX}$, pasando por los espacios F y F^*; la matriz de inercia en E se obtiene con la transformación compuesta de las 4 transformaciones lineales que permiten dar la vuelta al diagrama. $\mathbf{X'DXM}$. 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 del diagrama.

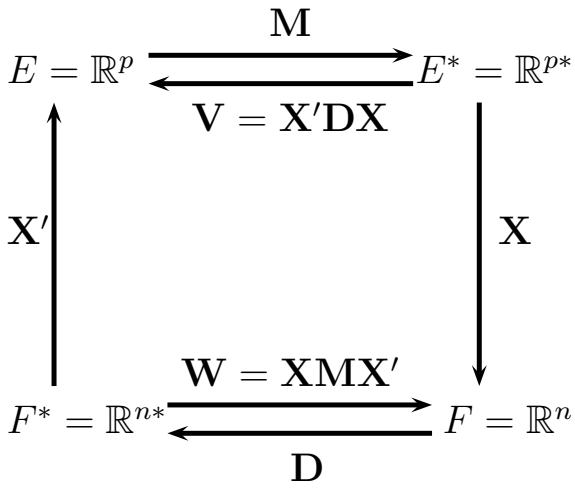


Figura 3.1: Diagrama de dualidad del $ACP(\mathbf{X}, \mathbf{M}, \mathbf{D})$. E es el espacio de las filas y F el de las columnas.

3.4. Ayudas para la interpretación de las gráficas

3.4.1. Calidad de la representación

Un indicador de esa calidad es el coseno, que es una relación entre las magnitudes de la proyección y del vector original. Sin embargo se utiliza el coseno cuadrado, ya que para un punto: la suma de los cosenos cuadrados sobre todos los ejes factoriales es 1; y el coseno cuadrado en un subespacio se obtiene sumando los cosenos cuadrados sobre los ejes factoriales que generan al subespacio. Sobre un eje s el coseno cuadrado de un punto-fila i es:

$$Cos_s^2(i) = \frac{F_s^2(i)}{\|\mathbf{x}_i\|^2} \tag{3.10}$$

Las coordenadas de un vector fila **i** están en la fila i de la matriz **X**. El valor del coseno al cuadrado coincide con la relación de contribuciones del individuo i a la inercia: *contribución a la inercia proyectada sobre el eje s/contribución a la inercia total* y se llama también contribución relativa.

3.4.2. Contribución absoluta

Otro aspecto que ayuda a la interpretación es identificar las filas que más estén contribuyendo a la inercia de un eje s. Este indicador se obtiene dividiendo la inercia proyectada del punto-fila i sobre la inercia total del eje, que es igual al valor propio. Se suele expresar en porcentaje y recibe el nombre de contribución absoluta de la fila i al eje s:

$$Ca_s(i) = \frac{p_i F_s^2(i)}{\lambda_s} \tag{3.11}$$

Para las columnas se definen los mismos indicadores como ayudas para la interpretación de las gráficas.

3.5. 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.

3.6. 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. 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 que se muestra en la figura 3.2. Para una matriz de covarianzas o de correlaciones de orden p la métrica $\mathbf{M}$ es $\mathbf{I}_p$.

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

Figura 3.2: Diagrama cuando solo se conoce la matriz de varianzas o de correlaciones. Parte superior del diagrama de dualidad de la figura 3.1.

.

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 3.1:

$$\mathbf{G}_s = \sqrt{\lambda_s}\mathbf{u}_s \tag{3.12}$$

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 3.9.1.

Código R. 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)
```

3.7. 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 3.3, 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):

$$w_{il} = \frac{1}{2}(d^2_{i\cdot} + d^2_{l\cdot} - d^2_{il} - d^2_{\cdot\cdot}) \tag{3.13}$$

donde: $d^2_{i\cdot} = \sum_{l=1}^{n} p_l d^2_{il}$ y $d^2_{\cdot\cdot} = \sum_{i,l=1}^{n} p_i p_l d^2_{il}$

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

Figura 3.3: Diagrama cuando solo se conoce la matriz de productos internos $\mathbf{W}$. Parte inferior del diagrama de dualidad de la figura 3.1.

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 3.1:

$$\mathbf{F}_s = \sqrt{\lambda_s}\mathbf{v}_s \tag{3.14}$$

El análisis en coordenadas principales se encuentra implementado, entre otros, en el `ade4` en la función `dudi.pco`. Para disimilitudes no euclidianas también se puede obtener este tipo de gráficas, metodología que se conoce con el nombre de *escalamiento multidimensional*. En el taller 3.9.2 se puede ver un ejemplo de aplicación.

3.8. Ejercicios

1. Escriba el diagrama de dualidad y las principales fórmulas del $ACP\left(\mathbf{Y}_c, \mathbf{I}_2, \mathbf{I}_{10}\right)$ del taller ACP geométrico sección 2.7.1 (página 71). Encuentre analíticamente los valores y vectores de la matriz de inercia del taller.
2. Muestre las distancias entre individuos en el $ACP\left(\mathbf{Y}_c, diag\left(\frac{1}{\sigma_j^2}\right), \frac{1}{n}\mathbf{I}_n\right)$ son iguales a las del $ACP\left(\mathbf{X}, \mathbf{I}_p, \frac{1}{n}\mathbf{I}_n\right)$.
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 $\mathbf{X'DXM}$.
4. Muestre que las coordenadas factoriales de las filas sobre un eje son **D**-centradas.
5. Para el ACP generalizado demuestre:
   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 **D**-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. En el ejemplo Café del capítulo 2 encuentre la matriz de rango 2 que mejor se aproxima a los datos. ¿Cuánto es el valor de la calidad de esa aproximación?

7. Encuentra las fórmulas del ACP normado canónico a partir de las fórmulas del ACP generalizado dadas en la tabla 3.1.

8. 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, dándole la vuelta al diagrama hasta llegar a E de nuevo.

9. Dibuje el diagrama de dualidad para el ACP canónico.

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

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

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

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

3.9.1. 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 3.2), responda las preguntas siguientes:

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. (puede utilizar la función *s.corcircle* del paquete *ade4* (Dray & Dufour 2007)).
8. Resuma el análisis respondiendo a los objetivos.

Tabla 3.2: Matriz de correlaciones entre variables de clima en la ciudad de Mendoza

	temp-peratura	vel viento	altura	SVF	emisi-vidad	ancho canal	orien-tacion	nubo-sidad	inercia inercia
temperatura	1.000	0.113	0.411	0.174	0.067	0.134	0.101	0.836	-0.164
vel.viento	0.113	1.000	0.040	0.291	0.009	0.009	0.018	0.123	-0.050
altura	0.411	0.040	1.000	0.102	0.406	-0.293	-0.018	0.157	-0.443
SVF	0.174	0.291	0.102	1.000	-0.118	0.165	-0.200	0.176	0.119
emisividad	0.067	0.009	0.406	-0.118	1.000	-0.361	-0.104	-0.007	-0.966
ancho.canal	0.134	0.009	-0.293	0.165	-0.361	1.000	0.029	0.176	0.290
orientacion	0.101	0.018	-0.018	-0.200	-0.104	0.029	1.000	0.103	0.035
nubosidad	0.836	0.123	0.157	0.176	-0.007	0.176	0.103	1.000	-0.053
inercia	-0.164	-0.050	-0.443	0.119	-0.966	0.290	0.035	-0.053	1.000

Fuente: Correa et al. (2006).

3.9.2. Análisis en coordenadas principales

En Hidalgo et al. (2007) se construye una distancia cultural entre algunos países latinoaméricanos (tabla 3.3), la cual corresponde a una distancia euclidiana elevada al cuadrado. Realice el análisis en coordenadas principales (ACO), sobre la matriz `>sqrt(es)` ($d_{il} = \sqrt{D_{il}}$) utilizando las funciones `dudi.pco` e `inertia.dudi` de *ade4* y responda a las preguntas siguientes:

1. ¿Cuál es la dimensión del espacio de representación (rango de la matriz)?

2. ¿Cuántos ejes selecciona para el análisis? ¿Por qué?

3. ¿Tiene sentido hablar de ayudas para la interpretación de las variables?

4. *Idem* para los individuos.

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

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

7. Analice los planos factoriales. Puede utilizar *plot.dudi* para graficar los planos factoriales que requiera.

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

9. Compare los resultados con los del artículo.

10. Haga un resumen práctico del análisis.

Tabla 3.3: Distancias culturales entre países de Latinoamérica

	Argentina	Bolivia	Brasil	Colombia	Costa Rica	Ecuador	Salvador	Guatemala	México	Venezuela
Argentina	0.000	2.348	1.677	0.796	2.240	2.409	1.490	0.750	2.832	1.060
Bolivia	2.348	0.000	1.736	2.385	1.086	2.848	3.048	2.390	2.339	1.227
Brasil	1.677	1.736	0.000	1.750	1.445	2.392	0.816	1.605	1.992	2.151
Colombia	0.796	2.385	1.750	0.000	2.746	2.317	1.427	1.301	1.840	1.182
CostaRica	2.240	1.086	1.445	2.746	0.000	1.980	3.273	2.867	1.542	3.223
Ecuador	2.409	2.848	2.392	2.317	1.980	0.000	2.511	1.767	1.833	2.500
Salvador	1.490	3.048	0.816	1.427	3.273	2.511	0.000	1.365	2.271	1.182
Guatemala	0.750	2.390	1.605	1.301	2.867	1.767	1.365	0.000	3.599	1.723
Mexico	2.832	2.339	1.992	1.840	1.542	1.833	2.271	3.599	0.000	2.499
Venezuela	1.060	1.227	2.151	1.182	3.223	2.500	1.182	1.723	2.499	0.000

Fuente: Hidalgo et al. (2007).

Capítulo 4

Análisis de correspondencias simples (ACS)

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

- comparar los perfiles fila,
- comparar los perfiles columna y
- 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 et al. (2006).

4.1. 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 4.1). La tabla se obtiene a partir de los datos `admi{FactoClass}`.

Código R. Para construir la tabla a partir de `admi{FactoClass}`:

```
library(FactoClass)
data(admi)
K<-unclass(table(admi$carr,admi$estr))
```

4.1.1. Tabla de contingencia

Siguiendo la misma notación de Lebart et al. (2006), $\mathbf{K}$ 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 $k = 445$.

Tabla 4.1: Clasificación 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

4.1.2. 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 4.1 está la tabla $\mathbf{F}$ 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.

La marginal fila representa a 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 $\mathbf{D}_p = diag(f_{\cdot j})$.

Código R. Para obtener $\mathbf{F}$, $\mathbf{D}_n$ y $\mathbf{D}_p$ y las tabla 4.1:

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

4.1.3. Tabla de perfiles fila

Para cada carrera se tiene una distribución de frecuencias entre los 3 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 $\mathbf{F}$. La marginal columna de la tabla $\mathbf{F}$ se constituye en la distribución promedio de los perfiles fila y es la distribución de todos los 445 admitidos en los 3 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` (ver tabla 4.2).

Se puede observar que el perfil de Geología es el que más difiere de los demás porque tiene más porcentaje de estrato alto, con 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.

4.1.4. Tabla de perfiles columna

Cada estrato tiene su distribución según las 7 carreras (condicionales o perfiles columna). La distribución marginal fila es la distribución de todos los admitidos, en las 7 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 $\mathbf{F}\mathbf{D}_p^{-1}$. En R: `F%*%solve(Dp)` (ver tabla 4.2).

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.

Código R. Para obtener tablas y gráficas de perfiles, tabla 4.2:

```
# plotct es una función de FactoClass
plotct(K,"row",col=c("brown","orange","darkgreen"))
tabs<-plotct(K,"col",col=c("darkgreen","brown","orange","darkblue",
   "gray","white","green"),tables=TRUE)
# para exportar la gráfica a xfig, para su edición
#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))
```

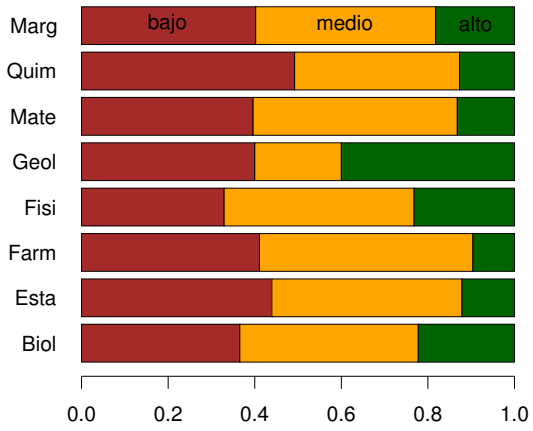
Tabla 4.2: Perfiles fila y columna de la tabla carreras×estratos

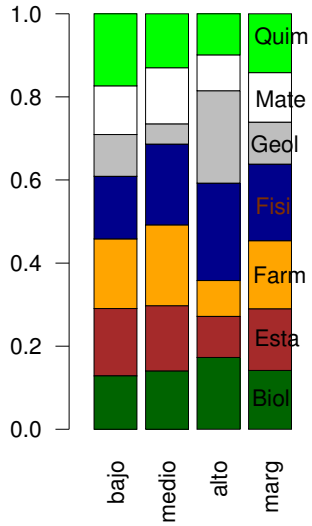
Perfiles de carreras según estratos

	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

Perfiles de estratos según carreras

	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





4.1.5. 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** (ver tabla 4.3). En esta tabla las distribuciones condicionales fila (respectivamente, columna) son todas iguales a la marginal de las columnas (filas) de la tabla **F**.

Las desviaciones al modelo de independencia son **F-A** (ver tabla 4.3).

Tabla 4.3: Tablas de: frecuencias relativas, independencia y diferencia

	F observada			**A** independencia			**F** − **A** diferencia		
	Ebajo	Emedio	Ealto	Ebajo	Emedio	Ealto	Ebajo	Emedio	Ealto
Biología	5.2	5.8	3.1	5.7	5.9	2.6	-0.5	-0.0	0.6
Estadística	6.5	6.5	1.8	6.0	6.2	2.7	0.6	0.4	-0.9
Farmacia	6.7	8.1	1.6	6.6	6.8	3.0	0.1	1.3	-1.4
Física	6.1	8.1	4.3	7.4	7.7	3.4	-1.3	0.4	0.9
Geología	4.0	2.0	4.0	4.1	4.2	1.8	-0.0	-2.2	2.2
Matemáticas	4.7	5.6	1.6	4.8	5.0	2.2	-0.1	0.7	-0.6
Química	7.0	5.4	1.8	5.7	5.9	2.6	1.3	-0.5	-0.8

4.2. El ACS como dos ACP

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

4.2.1. ACP de los perfiles-fila

La tabla que se analiza es la de perfiles fila, es decir que el histograma que representa al 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, suma de las filas de $\mathbf{F}$, y se ordenan en la matriz diagonal $\mathbf{D_n}$, las distancias entre distribuciones condicionales se definen a partir del producto punto dado por la matriz $\mathbf{D_p^{-1}}$. Las imágenes para 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 ($\mathbf{F'D_n^{-1}FD_p^{-1}}$, ver tabla 3.1) 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 se logra una simplificación de las formulas del ACS.

Veamos el anterior párrafo con más detalle:

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 4.2. 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 son las marginales fila de $\mathbf{F}$ (tabla 4.1), para Farmacia es: 0.164, es decir que el 16.4 % de los admitidos a Ciencias son de esta carrera.

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 que el centro de gravedad es la marginal columna de la tabla $\mathbf{F}$, 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, por facilidad, en las fórmulas, el ACP se hace sin centrar y luego se elimina 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.

Distancia entre perfiles fila

En este análisis la matriz de producto interno que genera la métrica es $\mathbf{D^{-1}}_p$, cuyo elemento diagonal es $\frac{1}{f_{\cdot j}}$, con la cual la distancia al cuadrado entre dos perfiles fila i y l es:

$$d^2(i,l) = \sum_{j=1}^{p} \frac{1}{f_{\cdot j}}(x_{ij} - x_{lj})^2 = \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{lj}}{f_{l\cdot}} \right)^2 \tag{4.1}$$

La distancia (4.1), 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 4.3.1; y las relaciones cuasi-baricéntricas, sección 4.3.2.

Inercia de la nube de perfiles fila

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

$$Inercia(N_n) = \sum_{i=1}^{n} f_{i\cdot} d^2(i, \mathbf{g}_p) = \sum_{i=1}^{n} f_{i\cdot} \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j} \right)^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} \tag{4.2}$$

La inercia (4.2) es el coeficiente ϕ^2 (1.4), una medida de asociación entre las dos variables cualitativas.

En las tablas de contingencia se suele probar independencia entre las dos variables categóricas. 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:

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

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 4.3) se puede calcular I, en R: `I=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 con $(7-1)(3-1) = 12$ grados de libertad.

En R, con el comando: `pchisq(29.19,12,lower.tail =FALSE)`, se obtiene 0.0037 y la decisión estadística es rechazar H_0. Recordemos la media y 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 H_0 se rechaza, porque $29.19 > 26.7$.

Código R. Para comparar las dos distribuciones:

```
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
```

Búsqueda de los nuevos ejes en el espacio de perfiles fila

La matriz de inercia $\mathbf{X'DXM}$ es:

$$\mathbf{F'D}_n^{-1}\mathbf{D}_n\mathbf{D}_n^{-1}\mathbf{FD}_p^{-1} = \mathbf{F'D}_n^{-1}\mathbf{FD}_p^{-1} \tag{4.3}$$

El término general de (4.3) es:

$$\sum_{i=1}^{n} \frac{f_{ij}}{f_{i\cdot}} \frac{f_{ij'}}{f_{\cdot j'}} \tag{4.4}$$

Para mostrar que (4.3) 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

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

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, es $\mathbf{D}_p^{-1}$ 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 y es lo que se puede denominar centrado a posteriori; con esto se pierde una dimensión y la nube de puntos queda soportada en el subespacio de dimensión $min(n, p) - 1$. En el ejemplo la nube de perfiles fila está soportada en $\mathbb{R}^2$, es decir que en el primer plano factorial se comparan los perfiles sin perder información.

El valor propio 1 no entra en el análisis, 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 et al. (2006, p.149).

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 4.1).

Código R. Para ACS utilizando las funciones `dudi.coa{ade4}` y `plot.dudi{FactoClass}`:

```
acs<-dudi.coa(K,scannf=FALSE)
plot(acs,Tcol=FALSE,xlim=c(-0.7,0.3),cframe=1)
# si se desea grabar la gráfica en formato xfig para editarla
#dev.print(device = xfig,file="cienciasACScarreras12.fig")
```

4.2.2. 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.

4.2.3. Representación simultánea

Los dos ACP están relacionados, debido a que los perfiles fila y columna se derivan de la misma matriz $\mathbf{F}$ 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 (§**??**) permiten la representación

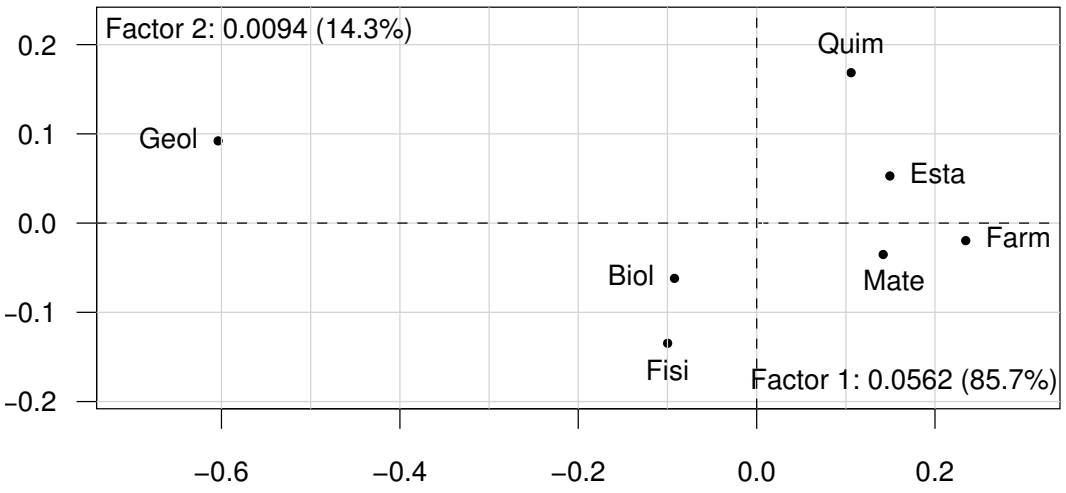


Figura 4.1: Primer plano factorial de los perfiles de carreras según estratos. Geología tiene el perfil más diferente del promedio y de las demás carreras, Biología se parece más al perfil promedio, Matemáticas, Estadísitica y Farmacia tienen perfiles parecidos. El primer eje retiene el 85.7 % de la inercia.

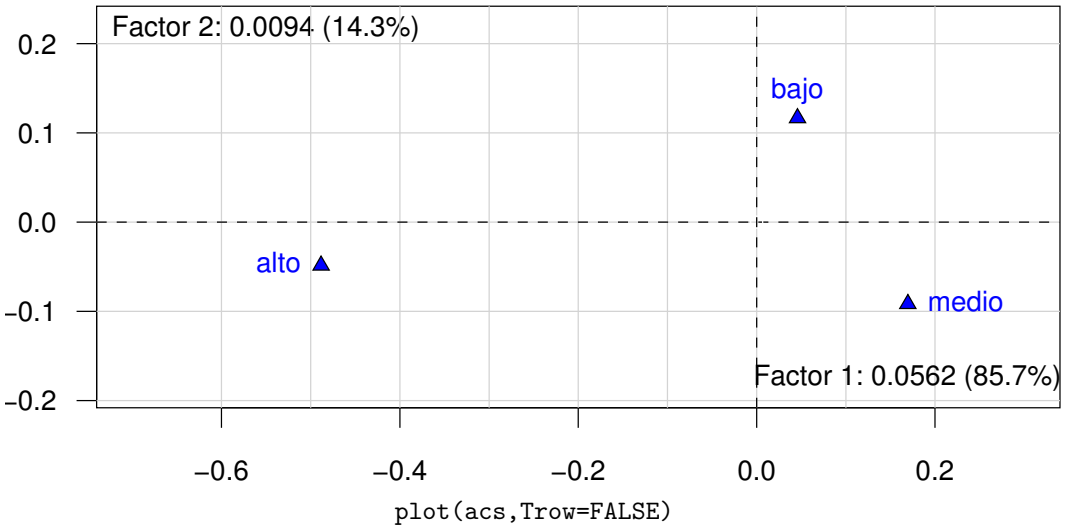


Figura 4.2: Primer plano factorial de los perfiles de estratos según carreras. El primer eje opone el estrato alto con el medio y el segundo eje opone, sobre todo, bajo con medio. El estrato alto es el que más se diferencia del promedio.

simultánea de los mapas factoriales. La deducción de las relaciones de transición es más fácil, cuando se ve el ACS como un ACP, que 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.

4.3. El ACS como un ACP(X,M,D)

El ACS de la tabla **F** también se obtiene mediante el ACP de la tabla **X** cuyo término general está dado por (4.5), usando $\mathbf{D} = \mathbf{D}_n = diag(f_{i\cdot})$, 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.

$$x_{ij} = \frac{f_{ij} - f_{i\cdot} f_{\cdot j}}{f_{i\cdot} f_{\cdot j}} \tag{4.5}$$

Todas las fórmulas del ACS se pueden derivar de las fórmulas correspondientes al ACP generalizado (ver tabla 3.1). 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:

$$d^2(i,l) = \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{lj}}{f_{l\cdot}} \right)^2; \quad d^2(j,k) = \sum_{i=1}^{n} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ik}}{f_{\cdot k}} \right)^2 \tag{4.6}$$

La expresiones de (4.6) son las mismas distancias ji-cuadrado entre los perfiles fila (4.1) y columna, derivados de sus respectivos ACP. Estas distancias tiene dos propiedades muy importantes para la interpretación de las salidas del ACS: equivalencia distribucional y relaciones cuasibaricéntricas.

4.3.1. 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 Estadística, Matemáticas y Farmacia; o las carreras Biología y Física (figura 4.3).

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.

4.3.2. Relaciones cuasibaricéntricas

En el ACS las relaciones de transición son:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{p} \frac{f_{ij}}{f_{i\cdot}} G_s(j) \tag{4.7}$$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{n} \frac{f_{ij}}{f_{\cdot j}} F_s(i) \tag{4.8}$$

Las relaciones cuasibaricéntricas (4.7 y 4.8) además de hacer posible la representación simultánea permiten su interpretación. Observemos por ejemplo (4.7), 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 (4.7) es un promedio ponderado de las coordenadas de las columnas; en general cada fila tiene un promedio diferente porque las ponderaciones cambian. En el primer plano, por ejemplo, el punto promedio se ubica dentro del polígono que une a los puntos columna, pero por las dilataciones $\frac{1}{\sqrt{\lambda_s}}; s = 1, 2$, el punto puede ubicarse afuera. De forma simétrica la sumatoria de la fórmula (4.8) 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; en un plano factorial este promedio se ubica dentro del polígono que une a todos los puntos fila; sin embargo por las dilataciones el punto columna 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.

Las fórmulas cuasibaricéntricas 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 4.3, Geología es atraída por el estrato alto y viceversa; Geología está más alejada del origen

que 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 %) (ver tabla 4.2).

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 4.2);
- las coordenadas de los estratos sobre el primer eje son [0.0458 0.1695 -0.4884] (figura 4.3);
- el primer valor propio es 0.0562 (figura 4.3);
- la fórmula (4.7) 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 porque es la que está más alejada del origen (-0.4884). La diferencia de la coordenada calculada -0.6315 con el valor -0.6037 del programa (figura 4.3) se debe a los errores de redondeo; el cálculo en R con más cifras significativas es:

```
1/sqrt(ca$eig[1])*sum(c(0.4,0.2,0.4)*ca$co[,1])
[1] -0.603727
```

4.3.3. 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.

Contribución absoluta

Las contribución de un perfil a la varianza del eje (inercia proyectada), depende del peso y de la coordenada al cuadrado:

$$Ca_s(i) = \frac{f_{i\cdot}(F_s(i))^2}{\lambda_s} \tag{4.9}$$

En el ejemplo, en la nube de carreras, la dirección del primer eje se debe sobretodo a Geología y Farmacia (81.7 % de contribución) y la del segundo eje a Química y Física (78.4 % de contribución).

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 $\mathbb{R}^p$:

$$Cos_s^2(i) = \frac{F_s^2(i)}{d^2(i,\mathbf{g})} \tag{4.10}$$

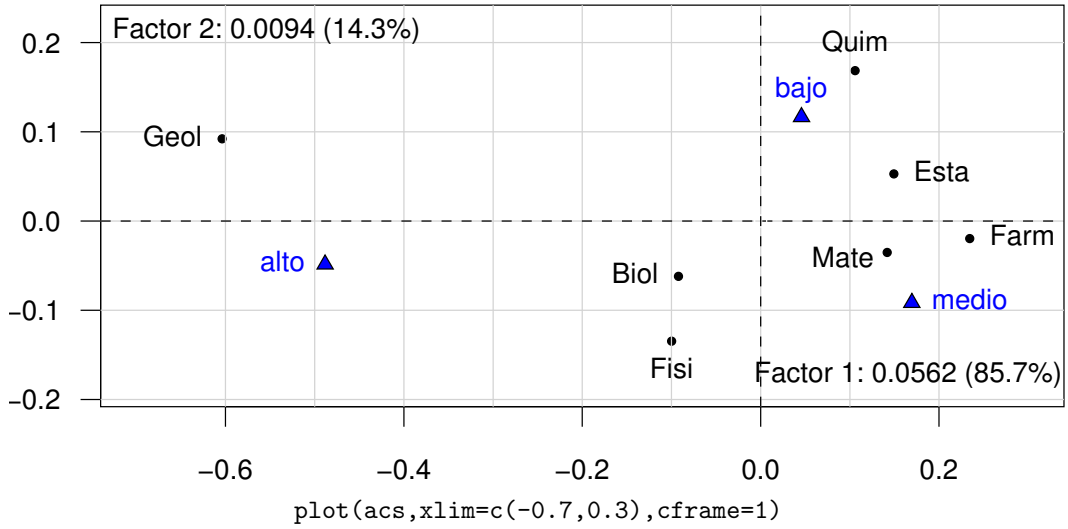
es decir su coordenada al cuadrado sobre la longitud al cuadrado en el espacio completo, que es la norma al cuadrado del vector centrado, o la distancia al cuadrado entre el punto y el centro de gravedad.

El nombre de contribución relativa se da porque el coseno cuadrado cuadrado es, también, un cociente de contribuciones a la inercia. Multiplicando numerador y denominador de (4.10) 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.



Coordenadas y ayudas para la interpretación

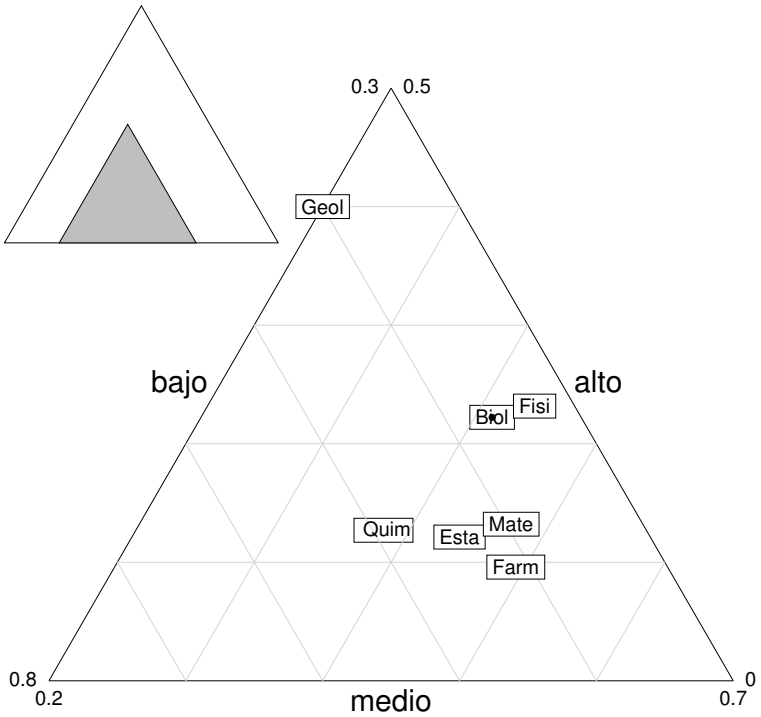
Carrera	Coordenadas		Contribuciones		Cosenos^2		Contribución
	Eje 1	Eje 2	Eje1	Eje 2	Eje1	Eje 2	Inercia en $\mathbb{R}^2$
Biología	-0.0922	-0.0620	2.1	5.8	68.9	31.1	2.7
Estadística	0.1494	0.0528	5.9	4.4	88.9	11.1	5.7
Farmacia	0.2345	-0.0196	16.1	0.7	99.3	0.7	13.8
Física	-0.0998	-0.1347	3.3	35.6	35.4	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 4.3: 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.

Diagrama triangular cuando una variable tiene tres categorías

En este ejemplo se puede utilizar un diagrama triangular para describir la composición de las carreras según los estratos de los estudiantes admitidos: utilizando la función `triangle.plot{ade4}`: `triangle.plot(K,label=rownames(K),clab=1)`, se obtiene la figura 4.4. Cada lado del triángulo equilátero representa un estrato con un extremo de valor cero y el otro uno, en el diagrama completo; sin embargo la función gráfica un triángulo interior

(sombreado) para aprovechar mejor el espacio; en donde estrato bajo va de 0.3 a 0.8, medio de 0.2 a 0.7 y alto de 0.0 a 0.5. En este gráfico se puede leer, por ejemplo, la composición de Geología según estratos: bajo 0.4, medio 0.2, alto 0.4, que son los valores del perfil que aparecen en la tabla 4.2. ¿Cuál es el perfil de Estadística, que se puede leer en este diagrama?.




```
triangle.plot(K,label=rownames(K),clab=1)
```

Figura 4.4: Diagrama triangular de composición de las carreras de Ciencias según los estratos de sus admitidos 2013-I. Para Farmacia se lee, aproximadamente: bajo 0.4, medio 0.5 y alto 0.1. En la tabla 4.2 se pueden ver los valores más precisos.

4.4. 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}`. El objetivo principal del análisis es comparar los perfiles departamentales según la calidad educativa de los colegios. La tabla tiene 29 departamentos, incluyendo 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 12 columnas, que es la clasificación combinada de la jornada del colegio: completa, mañana y tarde y su categoría de rendimiento: inferior, medio, bajo y alto. Los perfiles de los departamentos se muestran en la figura 4.5.

Código R. Para cargar datos y obtener la figura 4.5:

```
library(FactoClass)
data(icfes08)
azul<-c("white","cyan","blue","darkblue")
marron<-c("burlywood","coral","chocolate","brown")
verde<-c("green","darkolivegreen3","darkcyan","darkgreen")
mycolors<-c(azul,verde,marron)
par(mai=c(0,1,0,0)) # márgenes de la gráfica
plotct(icfes08,"row",col=mycolors)
# para grabar la gráfica en formato xfig y editarla
#dev.print(device = xfig,file="TCicfesPerpilesDeptos.fig")
```

Con el ACS se logra ordenar estos perfiles y ponerlos en correspondencia con las categorías de las columnas de la TC. En la figura 4.6, arriba están los valores propios y abajo el primer plano factorial mostrando tanto los departamentos como las categorías columna, se muestra también Chocó como suplementario. El histograma de valores sugiere analizar dos ejes, pero no se puede abandonar el tercero antes de ver qué muestra. La inercia total del ACS es 0.265 (`sum(acs$eig)`), el promedio de inercia para 11 valores propios es 0.024

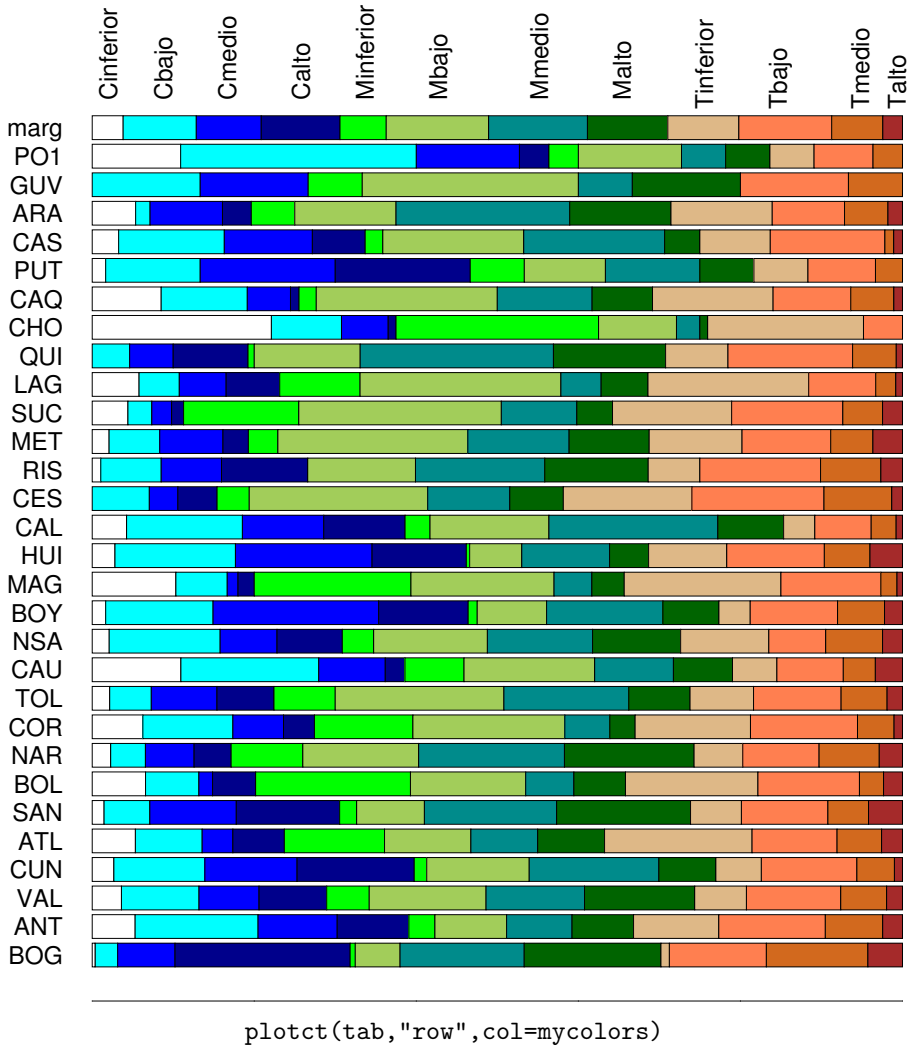


Figura 4.5: Perfiles de los departamentos según *jornadas* × *rendimiento*

y el tercer valor propio está cerca a este valor. Los dos primeros valores propios acumulan el 75.6 % de la inercia y los tres primeros el 84.2 %.

Código R. Para obtener la figura 4.6:

```
# Choco no activa
tab<-icfes08[-23,]
acs<-dudi.coa(tab,scannf=FALSE,nf=3)
barplot(acs$eig) #histograma valores propios
# para exportar a xfig y editarla
#dev.print(device = xfig,file="ACSicfesValP.fig")
valp<-t(inertia.dudi(acs)$TOT) #valores propios
# para obtener formato tabular para LaTeX
# xtable(valp,digits=rep(3,12))
```

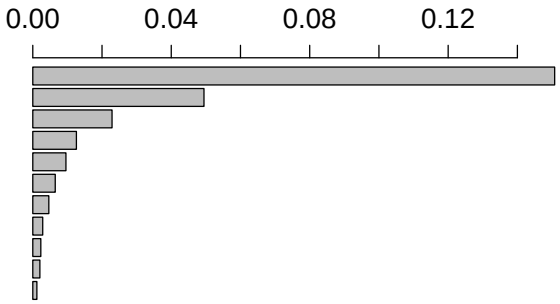
```
plot(acs,cframe=1,xlim=c(-1.2,0.8)) # primer plano
# proyección de Chocó como ilustrativa
Fchoco<-suprow(acs,icfes08[23,])$lisup
points(Fchoco,col="brown"); text(Fchoco,"CHO",col="brown",pos=1,cex
    =0.8)
#dev.print(device = xfig,file="ACSicfesP12.fig")
plot(acs,2,3,cframe=1.1)#,xlim=c(-1.2,0.8)) # plano 2-3
#dev.print(device = xfig,file="ACSicfesP23.fig")
```

Tabla 4.4: Coordenadas y ayudas para la interpretación de los departamentos

		Coordenadas			Contribuciones			Cosenos^2			Cont.
Depto	Peso	Eje 1	Eje 2	Eje 3	Eje1	Eje 2	Eje 3	Eje1	Eje 2	Eje3	Inercia
BOG	14.21	0.63	-0.27	-0.12	37.76	21.14	8.82	81.15	14.90	2.88	26.49
ANT	11.80	-0.03	0.24	-0.18	0.10	13.37	15.80	1.19	54.67	29.90	4.57
VAL	9.66	0.02	-0.01	0.07	0.02	0.01	2.10	1.41	0.27	19.02	0.95
CUN	7.08	0.20	0.19	0.07	1.87	5.24	1.48	34.03	31.36	4.11	3.12
ATL	5.79	-0.42	-0.16	-0.17	6.76	3.01	7.43	68.70	10.05	11.46	5.60
SAN	4.75	0.32	-0.07	0.04	3.23	0.46	0.29	65.50	3.04	0.88	2.81
BOL	4.68	-0.66	-0.28	-0.17	13.73	7.61	5.63	77.07	14.01	4.80	10.14
NAR	3.48	-0.00	-0.24	0.19	0.00	4.16	5.59	0.00	36.39	22.66	2.13
COR	2.86	-0.54	0.04	-0.02	5.47	0.09	0.07	92.81	0.52	0.18	3.35
TOL	3.68	-0.11	-0.09	0.26	0.28	0.56	11.19	10.45	7.00	64.49	1.50
CAU	3.27	-0.31	0.30	-0.04	2.07	6.02	0.24	28.47	27.17	0.49	4.14
NSA	2.83	-0.01	0.09	0.05	0.00	0.47	0.32	0.09	12.48	3.88	0.71
BOY	3.60	0.29	0.39	0.02	2.02	11.23	0.04	26.98	49.32	0.08	4.25
MAG	2.98	-0.89	-0.21	-0.13	15.64	2.68	2.19	90.18	5.08	1.92	9.87
HUI	2.48	0.17	0.38	-0.15	0.49	7.41	2.40	11.58	57.24	8.57	2.42
CAL	2.58	0.10	0.25	0.19	0.17	3.37	3.97	5.25	34.71	18.92	1.82
CES	2.26	-0.20	-0.10	0.23	0.61	0.45	5.35	15.10	3.65	20.02	2.31
RIS	1.87	0.26	-0.02	0.14	0.83	0.02	1.69	59.07	0.37	18.20	0.80
MET	1.91	-0.14	-0.02	0.32	0.24	0.02	8.46	10.93	0.29	57.70	1.27
SUC	2.03	-0.59	-0.27	0.21	4.69	3.02	3.77	69.85	14.77	8.52	3.82
LAG	1.20	-0.55	-0.10	0.12	2.43	0.24	0.70	67.70	2.20	2.99	2.04
QUI	1.29	0.25	-0.12	0.31	0.52	0.39	5.49	23.68	5.78	37.63	1.26
CAQ	0.93	-0.34	0.16	0.21	0.73	0.50	1.72	37.39	8.43	13.37	1.11
PUT	0.60	0.12	0.21	-0.07	0.06	0.54	0.11	6.03	19.29	1.87	0.52
CAS	0.91	-0.06	0.30	0.24	0.02	1.63	2.37	2.02	47.10	31.61	0.65
ARA	0.56	-0.07	-0.12	0.21	0.02	0.15	1.08	2.27	6.44	21.42	0.44
GUV	0.15	-0.08	0.18	0.39	0.01	0.10	0.97	1.40	6.36	29.87	0.28
PO1	0.55	-0.26	0.74	-0.18	0.25	6.09	0.74	8.60	70.03	3.96	1.62

El primer eje ordena las categorías de rendimiento de menor (izquierda) a mayor (derecha) y por lo tanto Bogotá es la región con mejor rendimiento y Magdalena el departamento 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.



	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

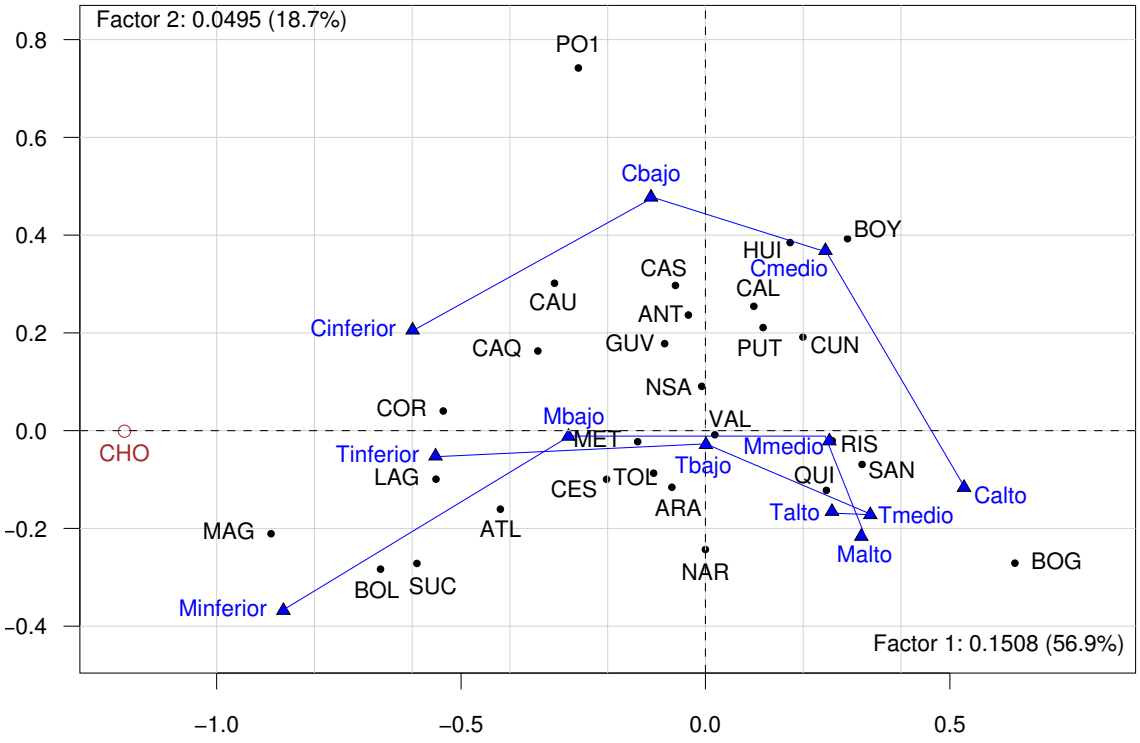


Figura 4.6: Primer plano factorial del ACS de la TC departamentos x categorías de rendimiento. Arriba valores propios con su histograma.

El tercer eje (ver figura 4.7) destaca al lado positivo las categorías baja y media de la jornada de la mañana, asociada a los departamentos Meta, Quindío, Tolima y Arauca.

En las tablas 4.4 y 4.5 se muestran las ayudas para la interpretación: las contribuciones a la inercia de los ejes sirven para detectar los puntos más relevantes en cada eje y el coseno

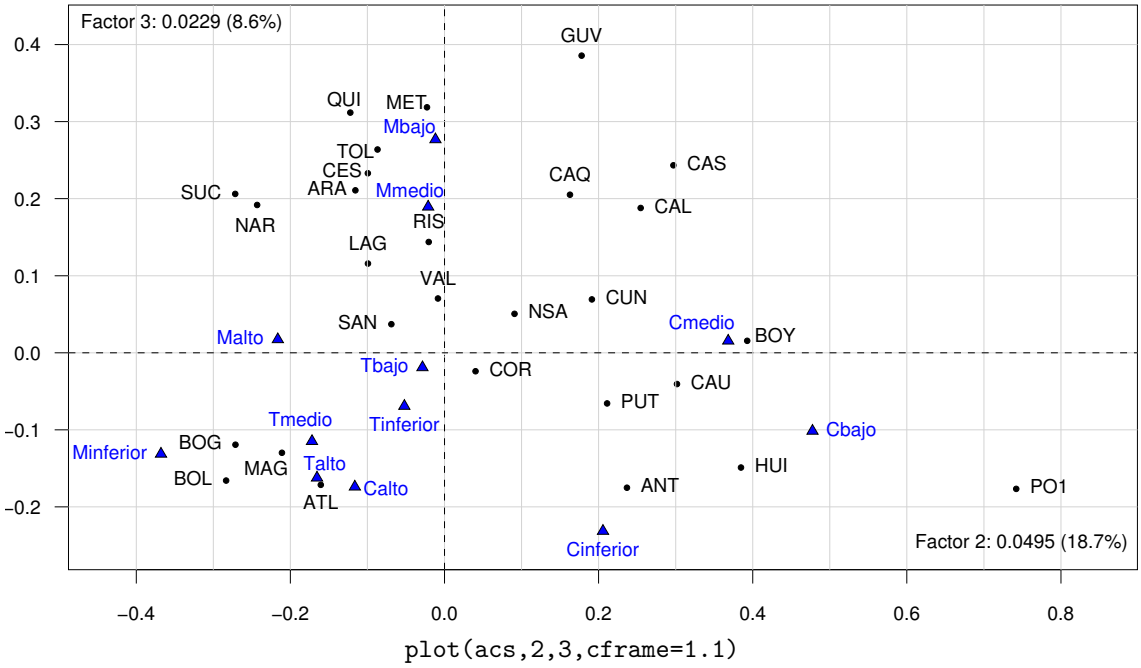


Figura 4.7: Plano factorial 2-3 del ACS de la TC deptos. x categorías. La segunda bisectriz ordena los departamentos por la jornada con relación al promedio: arriba a la derecha los que tienen mayor proporción en jornada de la mañana, hacia el centro otra jornada y en la parte derecha inferior los de mayor proporción de jornada completa.

cuadrado para ver las calidades de las proyecciones. El coseno cuadrado sobre un plano es la suma de los cosenos cuadrados de los ejes.

Código R. Para obtener las tablas 4.4 y 4.5:

```
ayu<-inertia.dudi(acs,T,T)
salayu<-data.frame(peso=acs$lw*100,acs$li,con=ayu$row.abs/100,
                                   cos=abs(ayu$row.rel)/100)
xtable(salayu,digits=rep(2,12))
salayu<-data.frame(peso=acs$cw*100,acs$co,con=ayu$col.abs/100,
                                   cos=abs(ayu$col.rel)/100)
xtable(salayu,digits=rep(2,12))
```

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 (bajo y medio) de los extremos. El ordenamiento de las categorías se traslada a los departamentos mostrando

Tabla 4.5: Coordenadas y ayudas para la interpretación de las columnas

		Coordenadas			Contribuciones			Cosenos^2			Cont.
Categoria	Peso	Eje 1	Eje 2	Eje 3	Eje1	Eje 2	Eje 3	Eje1	Eje 2	Eje3	Inercia
Cinferior	3.63	-0.60	0.21	-0.23	8.64	3.11	8.48	61.72	7.27	9.20	7.97
Cbajo	9.04	-0.11	0.48	-0.10	0.75	41.68	4.07	4.62	84.72	3.83	9.19
Cmedio	8.04	0.25	0.37	0.02	3.21	22.02	0.08	24.68	55.52	0.10	7.41
Calto	9.82	0.53	-0.12	-0.17	18.21	2.69	12.97	78.21	3.79	8.46	13.26
Minferior	5.50	-0.86	-0.37	-0.13	27.19	15.05	4.14	77.08	13.99	1.78	20.08
Mbajo	12.67	-0.28	-0.01	0.28	6.60	0.04	42.40	45.98	0.08	44.87	8.17
Mmedio	12.29	0.25	-0.02	0.19	5.24	0.11	19.26	51.79	0.36	28.91	5.76
Malto	10.01	0.32	-0.22	0.02	6.75	9.48	0.13	52.61	24.23	0.16	7.30
Tinferior	8.66	-0.55	-0.05	-0.07	17.52	0.48	1.81	79.15	0.70	1.24	12.60
Tbajo	11.53	0.00	-0.03	-0.02	0.00	0.19	0.18	0.00	2.92	1.30	1.22
Tmedio	6.36	0.34	-0.17	-0.11	4.80	3.80	3.66	55.13	14.34	6.39	4.95
Talto	2.46	0.26	-0.17	-0.16	1.09	1.36	2.82	29.69	12.13	11.66	2.09

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 departamentos del sur del país, luego los de la región Andina, y sobresalen Risaralda, Quíndio y Santander, hasta llegar a Bogotá la de mayor rendimiento.

Se utilizan las coordenadas sobre el primer eje para presentar los perfiles ordenando los departamentos, para que los perfiles parecidos queden vecinos en la figura 4.8. 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 este perfil se sitúa en el origen de la representación (coordenadas (0,0) en los planos) y la lejanía del centro de un punto, indica que es más diferente de este perfil.

Código R. Para obtener la figura 4.8:

```
ordep<-order(acs$li[,1])
par(mai=c(0,1,0,0))
plotct(tab[ordep,],"row",col=mycolors)
#dev.print(device = xfig,file="ACSicfesPerfilesOrdenados.fig")
```

Finalmente se presenta de nuevo el primer plano factorial, proyectando los centros de gravedad, de los grupos de departamentos y de las jornadas (figura 4.9), se observa que estos centros están próximos al centro del mapa, indicando que las diferencias no se deben a estos factores, sino que son regionales o debidas a otras características de los departamentos. Se observa, como es de esperarse, que la jornada de la tarde tiene menos porcentaje de colegios con buen rendimiento.

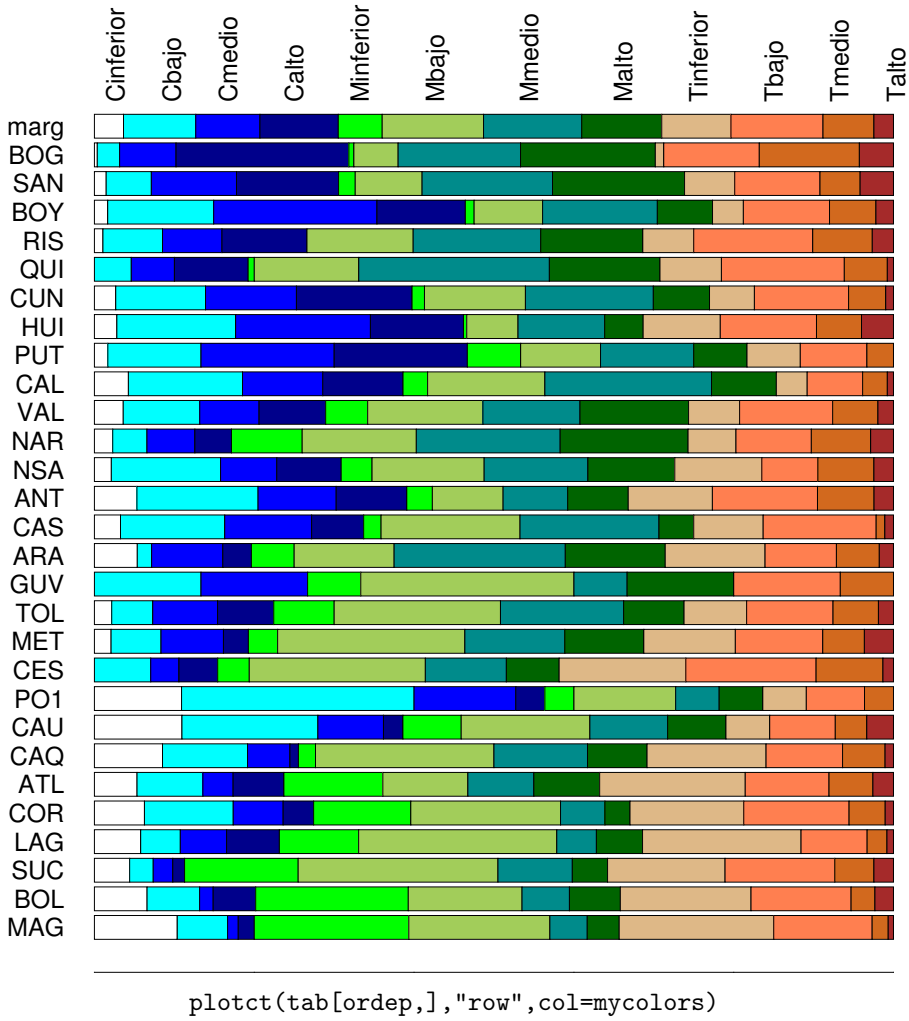


Figura 4.8: Perfiles de los departamentos ordenados por las coordenadas sobre el primer eje del ACS.

Código R. Para obtener la figura 4.9:

```
plot(acs,cframe=1,xlim=c(-1.2,0.8)) # primer plano
# proyección de Chocó como ilustrativa
Fchoco<-suprow(acs,icfes08[23,])$lisup
points(Fchoco,col="brown"); text(Fchoco,"CHO",col="brown",pos=1,cex
    =0.8)
rbl<-c(7,8,7,6)
rblf<-factor(rep(c("P5","P4","P3","P2"),rbl))
s.class(acs$li,rblf,acs$lw,col=c("brown","orange","darkgreen","
    darkblue"),
            cstar=0,cellipse=0,add.plot=TRUE)
jorf<-factor(rep(c("co","ma","ta"),each=4))
s.class(acs$co,jorf,acs$cw,col=c("blue","green","magenta"),cstar=0,
```

```
            cellipse=0,add.plot=TRUE)
#dev.print(device = xfig,file="ACSicfesSup12.fig")
```

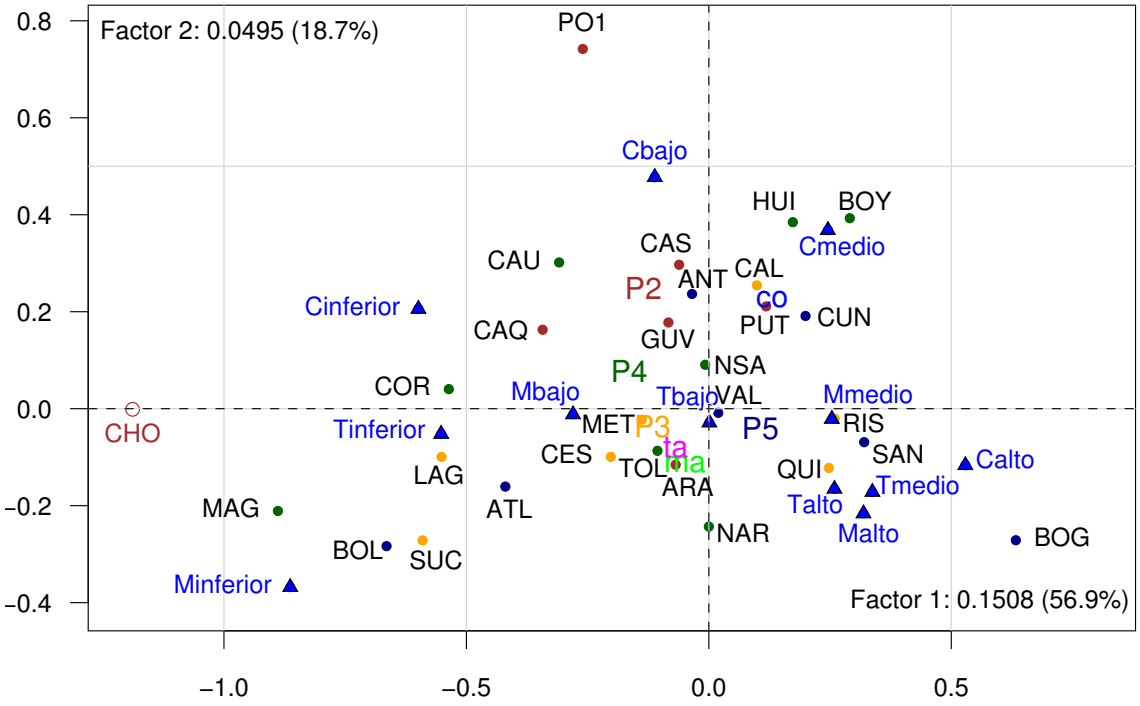


Figura 4.9: Primer plano factorial del ACS mostrando los centros de gravedad de grupos de departamentos (P2 a P5) y jornadas (co, ma, ta).

4.5. 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.
3. Demostrar que los valores propios en un AC 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* × *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 él.

7. Lea el perfil de Farmacia en el diagrama triangular de la figura 4.4.

4.6. Talleres de ACS

4.6.1. 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.

Trabajo

Realizar el ACS de la TC utilizando los estratos del 1 al 6 como columnas (frecuencias) activas y la columna *sin estrato* como ilustrativa.

Preguntas

1. Comente la repartición de las manzanas según estratos (histograma de la distribución de las manzanas en los 6 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 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. Utilizando el primer plano factorial proponga una partición de las localidades en cinco clases.

17. Proponga una partición en cinco clases utilizando la gráfica de perfiles entre localidades.

18. Proponga un partición "final" 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.

19. Construya una TC más pequeña formando algunos grupos de localidades de perfiles muy parecidos: Usme-CiudadBolívar, etc. (las dos filas se colapsan en una sumándolas en la TC). Realice un ACS de esta tabla, compare los resultados de la tabla completa. Comente la propiedad de equivalencia distribucional.

20. 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 6 estratos. Analice el ejercicio (¿quién atrae a quién y porqué?).

21. Describa la distribución geográfica de los habitantes de Bogotá según su nivel socioeconómico, utilizando el estrato de la manzana donde vive cada uno como indicador de ese nivel.

4.6.2. ACS *adjetivos* × *colores*

Este ejemplo se encuentra en español en Fine (1996) y en inglés en Jambu (1983), de donde se tomó la tabla para el ejemplo del paquete `FactoClass`.

Objetivo

Una agencia de publicidad encarga un estudio sobre las asociaciones entre colores y adjetivos, para mejorar la adaptación de los colores de 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). Los datos están en el archivo *colores.txt*.

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 1081 de la tabla de contingencia?.
3. ¿Cuántos ejes retiene para el análisis ¿Por qué?.
4. Teniendo en cuenta los seis primeros ejes identifique en qué planos están mejor representados cada uno de los 11 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 una categoría de otros.
7. Para un adjetivo cualquiera compruebe numéricamente la fórmula de transición (cuasibaricentro de las coordenadas de los 11 colores ponderadas por el perfil del respectivo adjetivo).
8. Escriba la conclusión del análisis (¿qué adjetivos se asocian más a cada color?.)

Capítulo 5

Análisis de correspondencias múltiples (ACM)

Con este método se describen tablas de "individuos" por variables cualitativas, ya sean nominales u ordinales. Con el ACM se abordan los siguientes objetivos:

1. Comparar los individuos, generalmente anónimos, para detectar patrones que emergen de los datos.

2. Comparar las categorías de las variables y detectar grupos de ellas.

3. Explorar relaciones entre las variables a través de sus categorías.

4. Describir correspondencia entre individuos y variables.

5. Cuantificar las variables cualitativas y reducir de dimensión. El ACM es un método que sirve para reemplazar las variables cualitativas por las coordenadas factoriales, que son variables continuas y de esa manera se pueden utilizar métodos estadísticos que funcionan bien con variables continuas. En ese sentido el ACM se puede considerar un método de pretratamiento, por ejemplo para: regresión, discriminación, agrupamiento, etc.

El ACM es una extensión del ACS con propiedades muy particulares, que se abordan en este capítulo.

5.1. Ejemplo: descripción de admitidos según algunas variables sociodemográficas

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.

Código R. Para obtener $\mathbf{Y}$ de `admi{FactoClass}`, $\mathbf{Z}$ y producir la tabla 5.1:

```
library(FactoClass)
data(admi)
Y<-admi[,c(8,11,9,10)]
# Para tabla del texto
Z<-acm.disjonctif(Y); data.frame(Y,Z)[seq(0,nrow(Y),25),]#
   registros múltiplos de 25
xtable(data.frame(Y,Z)[seq(0,nrow(Y),25),],digits=rep(0,17))
```

5.2. Transformaciones de la tabla de datos y notación

En esencia se utiliza la misma notación de Lebart et al. (2006), que se va presentando a medida que aparecen los distintos elementos.

5.2.1. 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 de 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 5.1 se muestra un extracto de la tabla $\mathbf{Y}$ del ejemplo.

5.2.2. Tabla disyuntiva completa (TDC)

La 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 5.1 se muestra un extracto de la TDC para el ejemplo de los admitidos.

Tabla 5.1: Extracto de las tablas: de código condensado $\mathbf{Y}$ y disyuntiva completa $\mathbf{Z}$

	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

```
Z<-acm.disjonctif(Y); data.frame(Y,Z)[seq(0,445,25),]#registros múltiplos de 25
```

La TDC $\mathbf{Z}$ es una yuxtaposición de s tablas, donde 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. En la tabla 5.1 se puede ver la estructura de la matriz $\mathbf{Z} = [\mathbf{Z}_1 \ \mathbf{Z}_2 \ \mathbf{Z}_3 \ \mathbf{Z}_4]$, para el ejemplo. Una variable cualitativa q divide al conjunto de los n individuos en p_q grupos, donde p_q es el número de categorías de la variable q. 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). Como hay s submatrices $\mathbf{Z}_q$, la suma de cada fila de $\mathbf{Z}$ es s, es decir que 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: $n_j = z_{.j}$ 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 tabla $\mathbf{B}$ (tabla 5.2).

5.2.3. Tabla de Burt

Se denomina tabla de Burt o tabla de contingencia múltiple 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 (5.2) cruza las 12 categorías de las 4 variables de los admitidos a Ciencias. Algunas lecturas son: en los admitidos hay 128 mujeres y 317 hombres; 46 mujeres tienen 16 años o menos, 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 $\mathbf{Z_1'Z_3} = \begin{pmatrix} 46 & 59 & 23 \\ 133 & 126 & 58 \end{pmatrix}$.

Tabla 5.2: 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))
```

5.3. El ACM como un AC de la TDC

El ACM es el AC de la tabla disyuntiva completa $\mathbf{Z}$, también es el AC de la tabla de Burt $\mathbf{B}$, pero en este último caso se pierde la información de los individuos.

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 $f_{\cdot 1}=\frac{128}{1780}=7.19\,\%$.

5.3.1. Nube de individuos

Los n individuos conforman la nube N_n en $\mathbb{R}^p$, cuyas propiedades se muestran a continuación.

Coordenadas, pesos

Los perfiles de los individuos son las filas de la tabla $\frac{1}{s}\mathbf{Z}$, es decir que 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 de la tabla 5.1 (25) es: $\{0.25\ 0\ 0\ 0.25\ 0\ 0\ 0\ 0.25\ 0\ 0\ 0\ 0.25\}$. La métrica en este espacio de los individuos tiene término general $m_j=\frac{1780}{n_j}$.

Centro de gravedad

La coordenada j del centro de gravedad g_p es $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{s}z_{ij}=\frac{n_j}{ns}$, que es la marginal columna de $\mathbf{F}=\frac{1}{ns}\mathbf{Z}$.

Código R. 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+	Ebajo	Emedio	Ealto	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

Distancia entre individuos

La distancia al cuadrado entre dos individuos es:

$$d^2(i,l) = ns \sum_{j=1}^{p} \frac{1}{n_j} \left(\frac{1}{s} [z_{ij} - z_{lj}] \right)^2 = \frac{n}{s} \sum_{j=1}^{p} \frac{1}{n_j} (z_{ij} - z_{lj})^2 \tag{5.1}$$

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.

Por ejemplo la distancia al cuadrado entre los individuos 50 y 100 de la tabla 5.1 es $d^2(i50, i100) = \frac{445}{4} \left(\frac{1}{179} + \frac{1}{185} \right) = 1.22$. La única diferencia entre ellos es que el primero es de estrato bajo y el segundo es de estrato medio.

La distancia (5.1) se puede expresar como una distancia euclidiana canónica, introduciendo la métrica en las coordenadas:

$$d^2(i,l) = \frac{n}{s} \sum_{j=1}^{p} \frac{1}{n_j} (z_{ij} - z_{lj})^2 = \sum_{j=1}^{p} \left(\frac{\sqrt{n} z_{ij}}{\sqrt{s n_j}} - \frac{\sqrt{n} z_{lj}}{\sqrt{s n_j}} \right)^2 \tag{5.2}$$

Código R. Para calcular las distancias entre individuos usando la función `dist` y obtener la tabla 5.3:

```
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(Dis,1)
xtable(as.matrix(Dis),digits=rep(1,18))
```

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 5.1. Lo mismo sucede para las parejas 75 y 275; 250 y 325; 350 y 375.

Tabla 5.3: Distancia entre algunos admitidos asociada al ACM

	25	50	75	100	125	150	175	200	225	250	275	300	325	350	375	400	425
25	0.0	2.6	2.0	2.3	0.0	2.1	2.5	1.1	2.2	2.2	2.0	2.1	2.2	2.1	2.1	2.3	2.5
50	2.6	0.0	1.6	1.1	2.6	2.0	2.3	2.3	2.5	2.2	1.6	2.1	2.2	2.1	2.1	1.1	2.2
75	2.0	1.6	0.0	2.0	2.0	1.7	1.9	1.7	2.3	1.4	0.0	1.8	1.4	1.7	1.7	2.0	1.9
100	2.3	1.1	2.0	0.0	2.3	2.3	2.3	2.6	2.5	2.1	2.0	2.4	2.1	1.8	1.8	1.6	2.2
125	0.0	2.6	2.0	2.3	0.0	2.1	2.5	1.1	2.2	2.2	2.0	2.1	2.2	2.1	2.1	2.3	2.5
150	2.1	2.0	1.7	2.3	2.1	0.0	2.3	1.8	2.2	2.2	1.7	2.2	2.2	2.1	2.1	1.7	1.8
175	2.5	2.3	1.9	2.3	2.5	2.3	0.0	2.5	1.9	1.3	1.9	1.9	1.3	1.4	1.4	2.5	1.4
200	1.1	2.3	1.7	2.6	1.1	1.8	2.5	0.0	2.2	2.2	1.7	1.7	2.2	2.4	2.4	2.0	2.5
225	2.2	2.5	2.3	2.5	2.2	2.2	1.9	2.2	0.0	1.8	2.3	2.0	1.8	2.4	2.4	2.8	1.2
250	2.2	2.2	1.4	2.1	2.2	2.2	1.3	2.2	1.8	0.0	1.4	2.3	0.0	1.9	1.9	2.4	1.3
275	2.0	1.6	0.0	2.0	2.0	1.7	1.9	1.7	2.3	1.4	0.0	1.8	1.4	1.7	1.7	2.0	1.9
300	2.1	2.1	1.8	2.4	2.1	2.2	1.9	1.7	2.0	2.3	1.8	0.0	2.3	1.7	1.7	2.4	2.4
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	0.0	1.9	1.9	2.4	1.3
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	0.0	0.0	2.4	2.0
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	0.0	2.4	2.0
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	0.0	2.5
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	0.0

Inercia de la nube de perfiles fila

La inercia de la nube de puntos es

$$\frac{1}{n}\sum_{i=1}^{n} d^2(i,\mathbf{g}_p) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{ns}{n_j}(\frac{z_{ij}}{s}-\frac{n_j}{ns})^2 = \frac{1}{s}\sum_{j=1}^{p}\frac{1}{n_j}\sum_{i=1}^{n}(z_{ij}^2 - 2z_{ij}\frac{n_j}{n}+\frac{n_j^2}{n^2}) =$$

$$\frac{1}{s}(p-2s+s) = \frac{p}{s}-1$$

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 $12/4-1=2$.

Ejes y subespacios vectoriales

Las proyecciones de la nube de individuos se hacen mediante el $ACP(\frac{1}{s}\mathbf{Z}, ns\mathbf{D}_p^{-1}, \frac{1}{n}\mathbf{I}_n)$. La matriz de inercia es:

$$\frac{1}{s}\mathbf{Z}'\frac{1}{n}\mathbf{I}_n\frac{1}{s}\mathbf{Z}ns\mathbf{D}_p^{-1} = \frac{1}{s}\mathbf{Z}'\mathbf{Z}\mathbf{D}_p^{-1} = \frac{1}{s}\mathbf{B}\mathbf{D}_p^{-1}.$$

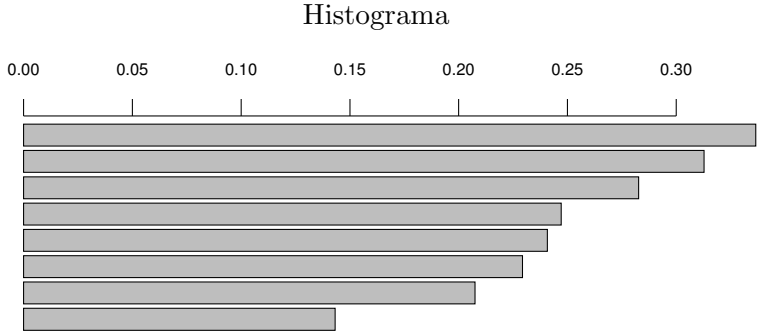
El rango de esta matriz es igual al rango de $\mathbf{Z}$, que es $p-s$, porque por cada variable hay una columna que es linealmente dependiente, es decir que una columna se puede obtener como la diferencia entre el vector de n unos y la suma de las demás columnas asociadas

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 5.1.

Código R. Para obtener la figura 5.1:

```
acm<-dudi.acm(Y,scannf=FALSE,nf=3)
dev.new()
barplot(acm$eig,cex.axis=0.6)
# ajustar la gráfica para tener la apariencia adecuada del
    histograma y
# grabarla  para editarla con el programa xfig
dev.print(device = xfig,file="ACMadmiValP.fig")
eiglst<-data.frame(vp=acm$eig,porce=acm$eig*100/sum(acm$eig),
                   acupor=cumsum(acm$eig)*100/sum(acm$eig))
eiglst
xtable(eiglst,digits=c(1,3,1,1)) #tabla en formato LaTeX
```



	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 5.1: 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.

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 infor-

mación. El porcentaje de inercia no es un criterio apropiado en el caso del ACM. Del histograma de la figura 5.1 se puede concluir que tres ejes son suficientes y, sin embargo retienen menos del 50 % de la inercia (46.6 %).

Código R. Para obtener el primer plano factorial, figura 5.2, con `plot.dudi{FactoClass}`, que tiene como entrada el objeto `dudi` que se obtiene con la función `dudi.acm{ade4}` y la tabla de ayudas en entorno *tabular* de LaTeX:

```
plot(acm,Tcol=FALSE,roweti=as.character(selin[-c(125,275,325,375)/
    25]), cframe=0.9,cex.row=0.6)
points(acm$li[selin,],col="darkgreen")
text(acm$li[c(125,275,325,375),],labels=c(125,275,325,375),col="
    darkgreen",cex=0.6,pos=c(4,1,1,1))
# dev.print(device = xfig,file="ACMadmiEstuP12.fig")
# coordenadas y ayudas
ineracm<-inertia.dudi(acm,T,T)
lstestu<-data.frame(coor=acm$li[selin,],contr=ineracm$row.abs[selin
    ,]/100,cos2=abs(ineracm$row.rel[selin,]/100))
lstestu
xtable(lstestu,digits=rep(2,11))
```

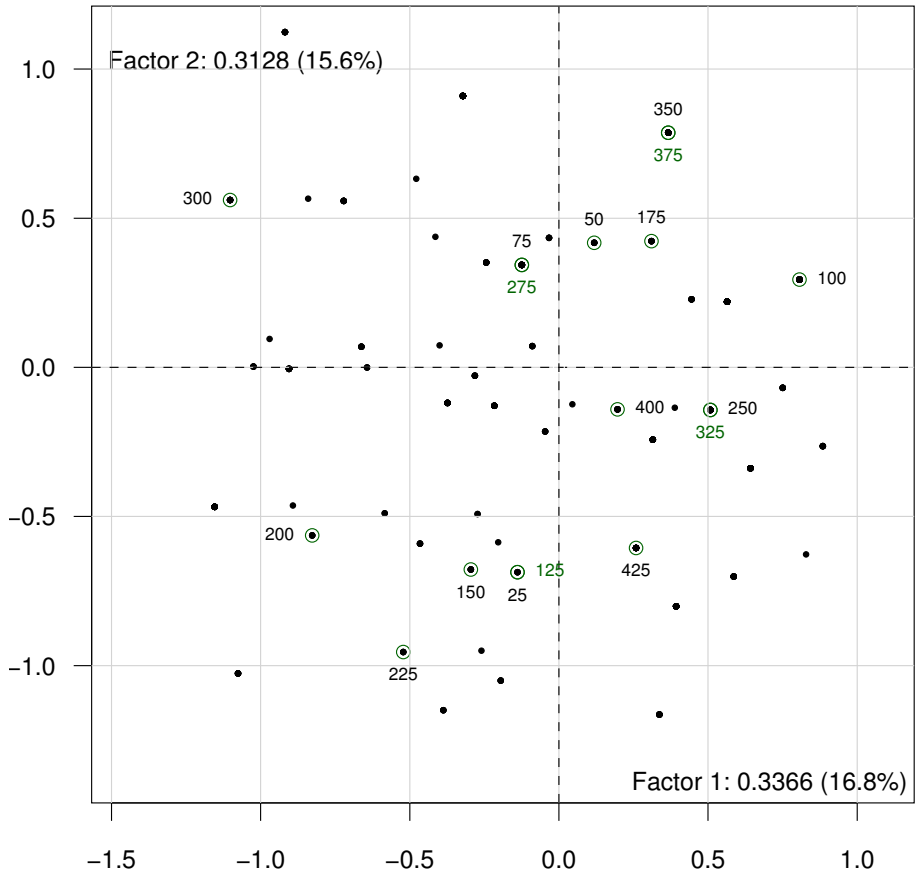
En la figura 5.2, se identifican los admitidos que están en la tabla 5.1. En el plano factorial no hay 445 puntos porque los admitidos que asuman las mismas categorías en las cuatro variables quedan superpuestos: en la tabla hay cuatro pares de estudiantes superpuestos, los individuos adicionales se etiquetaron con color verde.

5.3.2. Nube de categorías

En el ACM se pone más atención en 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.

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 $\frac{n_j}{ns}$.



Coordenadas y ayudas para la interpretación de los puntos de la tabla 5.1, hay 4 puntos superpuestos (etiqueta verde)

	Coordenadas			Contribuciones %			Cosenos2 %			Contri.
Est.	F1	F2	F3	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3	inercia %
25	-0.14	-0.69	0.08	0.01	0.34	0.00	0.84	20.70	0.27	0.26
50	0.12	0.42	0.57	0.01	0.13	0.26	0.61	7.54	14.00	0.26
75	-0.12	0.34	-0.38	0.01	0.08	0.11	1.58	12.02	14.52	0.11
100	0.81	0.29	0.63	0.43	0.06	0.32	28.34	3.78	17.38	0.26
125	-0.14	-0.69	0.08	0.01	0.34	0.00	0.84	20.70	0.27	0.26
150	-0.30	-0.68	0.51	0.06	0.33	0.21	4.87	25.64	14.70	0.20
175	0.31	0.42	-0.32	0.06	0.13	0.08	4.39	8.16	4.69	0.25
200	-0.83	-0.56	0.02	0.46	0.23	0.00	29.74	13.81	0.01	0.26
225	-0.52	-0.95	-0.38	0.18	0.65	0.11	9.64	32.23	5.08	0.32
250	0.51	-0.14	-0.99	0.17	0.01	0.77	14.87	1.18	56.27	0.19
275	-0.12	0.34	-0.38	0.01	0.08	0.11	1.58	12.02	14.52	0.11
300	-1.10	0.56	0.35	0.81	0.23	0.09	54.18	14.03	5.30	0.25
325	0.51	-0.14	-0.99	0.17	0.01	0.77	14.87	1.18	56.27	0.19
350	0.37	0.79	0.35	0.09	0.44	0.10	9.44	43.48	8.69	0.16
375	0.37	0.79	0.35	0.09	0.44	0.10	9.44	43.48	8.69	0.16
400	0.20	-0.14	0.91	0.03	0.01	0.65	1.36	0.70	29.04	0.32
425	0.26	-0.61	-0.43	0.04	0.26	0.15	3.30	18.12	9.33	0.23

Figura 5.2: Estudiantes sobre el primer plano factorial del ACM.

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 $\mathbf{g}(i)$:

$$\mathbf{g}(i) = \sum_{j=1}^{p} \frac{n_j}{ns} \frac{z_{ij}}{n_j} = \sum_{j=1}^{p} \frac{1}{ns} z_{ij} = \frac{1}{ns} s = \frac{1}{n}$$

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:

$$d^2(j,k) = \sum_{i=1}^{n} n \left(\frac{z_{ij}}{n_j} - \frac{z_{ik}}{n_k} \right)^2 \tag{5.3}$$

Para calcular la distancia en el ejemplo admitidos con la función `dist{stats}` hay que introducir n en el paréntesis para tener las coordenadas de una distancia euclidiana canónica. En la tabla 5.4 se muestran las distancias entre las categorías del ACM del ejemplo admitidos.

Código R. Para calcular las distancias entre categorías y obtener la tabla 5.4:

```
Xcat<-sqrt(n)*solve(Dp) %*% t(Z) #coord eucli canónicas
rownames(Xcat)<-substr(colnames(Z),6,nchar(colnames(Z)))
Discat<-dist(Xcat) # distancias
round(Discat,1)
xtable(as.matrix(Discat),digits=rep(1,13))
```

Tabla 5.4: Distancia entre las categorías activas asociadas al ACM del ejemplo admitidos

	F	M	a16m	a17	a18	a19M	bajo	medio	alto	Bogo	Cund	Otro
F	0.0	2.2	2.1	2.1	3.0	2.6	2.0	1.9	2.6	1.7	3.7	2.4
M	2.2	0.0	1.9	1.4	2.7	1.9	1.3	1.4	2.2	0.9	3.3	2.0
a16m	2.1	1.9	0.0	2.5	3.4	2.9	2.1	2.1	2.6	1.9	3.7	2.3
a17	2.1	1.4	2.5	0.0	3.2	2.7	1.8	1.7	2.4	1.4	3.4	2.3
a18	3.0	2.7	3.4	3.2	0.0	3.5	2.9	2.8	3.4	2.6	4.3	3.4
a19M	2.6	1.9	2.9	2.7	3.5	0.0	2.0	2.2	3.0	1.9	3.8	2.8
bajo	2.0	1.3	2.1	1.8	2.9	2.0	0.0	2.2	2.8	1.5	3.4	2.0
medio	1.9	1.4	2.1	1.7	2.8	2.2	2.2	0.0	2.8	1.2	3.6	2.4
alto	2.6	2.2	2.6	2.4	3.4	3.0	2.8	2.8	0.0	2.2	4.0	3.0
Bogo	1.7	0.9	1.9	1.4	2.6	1.9	1.5	1.2	2.2	0.0	3.6	2.5
Cund	3.7	3.3	3.7	3.4	4.3	3.8	3.4	3.6	4.0	3.6	0.0	4.0
Otro	2.4	2.0	2.3	2.3	3.4	2.8	2.0	2.4	3.0	2.5	4.0	0.0

La interpretación de la distancia entre categorías como está dada en (5.3) no es directa, porque 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. Haciendo una tabla de contingencia para las dos categorías se obtiene:

		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 que 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.

El desarrollo del cuadrado en (5.3) da:

$$d^2(j,k)=\sum_{i=1}^{n} n\left(\frac{z_{ij}^2}{n_j^2}-2\frac{z_{ij}}{n_j}\frac{z_{ik}}{n_k}+\frac{z_{ik}^2}{n_k^2}\right)=n\left(\frac{1}{n_j}+\frac{1}{n_k}-2\sum_{i=1}^{n}\frac{z_{ij}}{n_j}\frac{z_{ik}}{n_k}\right) \tag{5.4}$$

El último término del paréntesis en (5.4) solo suma cuando los dos individuos asumen la misma categoría, es decir para a individuos, entonces:

$$d^2(j,k)=n\left(\frac{n_k+n_j-2a}{n_jn_k}\right)=n\left(\frac{a+c+a+b-2a}{n_jn_k}\right)=\frac{n}{n_jn_k}(b+c) \tag{5.5}$$

La fórmula (5.5) muestra que en la distancia de dos categorías solo suman los individuos que asumen una y solo una de las dos categorías. Además las categorías de baja frecuencia se alejan más de las demás.

Se muestran a continuación dos ejemplos de cálculo de distancias entre categorías usando "la calculadora R":

1.

Código R. Entre Femenino y Masculino: son dos categorías de la misma variable con solo dos categorías, 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.

Código R. 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:

$$d^2(j,\mathbf{g}_n) = n\sum_{i=1}^{n}\left(\frac{z_{ij}}{n_j} - \frac{1}{n}\right)^2 = n\sum_{i=1}^{n}\left(\frac{z_{ij}^2}{n_j^2} - 2\frac{z_{ij}}{n_j}\frac{1}{n} + \frac{1}{n^2}\right) = \frac{n}{n_j} - 1 \tag{5.6}$$

Se observa en (5.6) que las categorías de menos frecuencia son las más alejadas del origen.

Inercia de la nube de categorías

Es interesante obtener la inercia calculando la contribución de una categoría a la 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:

$$I(N_p) = \sum_{q=1}^{s}\sum_{j\in J_q}\frac{n_j}{ns}d^2(j,\mathbf{g}_n) = \sum_{q=1}^{s}\sum_{j\in J_q}\frac{n_j}{ns}\left(\frac{n}{n_j} - 1\right) = \sum_{q=1}^{s}\sum_{j\in J_q}\frac{1}{s}\left(1 - \frac{n_j}{n}\right) \tag{5.7}$$

Donde J_q es el conjunto de categorías que pertenecen la variable q. A partir de (5.7) se observan o derivan las contribuciones a la inercia de una categoría, una variable y, de nuevo, 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.

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 subsección 5.3.3. El primer plano factorial de las categorías, obtenido en el ACM del ejemplo *admitidos*, se presenta en la figura 5.3.

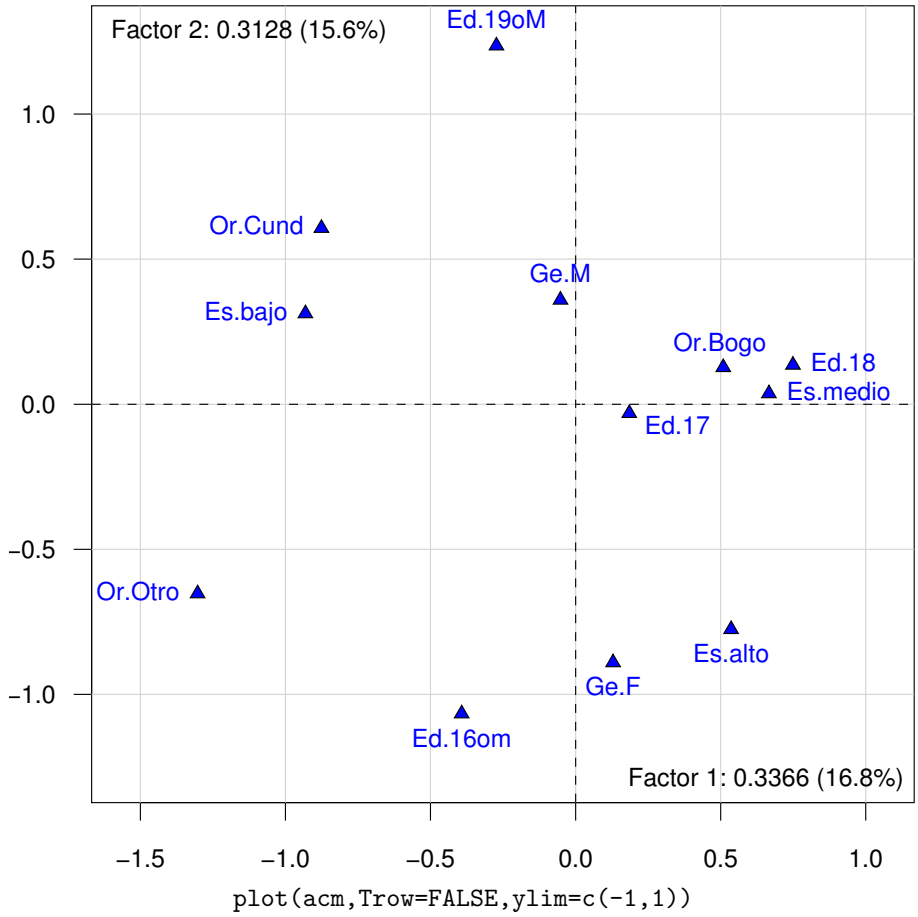
5.3.3. El ACM como un ACP

Es conveniente ver el ACM como un solo $ACP(\mathbf{X},\mathbf{M},\mathbf{D})$, 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 (Escofier & Pagès 1992), que no se abordan en este texto.

Haciendo específica la sección 4.3 al caso del ACM, las matrices $\mathbf{X},\mathbf{M},\mathbf{D}$ son:

- $\mathbf{X}=n\mathbf{I}_n\frac{1}{ns}\mathbf{Z}ns\mathbf{D}=n\mathbf{Z}\mathbf{D}_p^{-1}$, término general: $x_{ij}=\frac{n}{n_j}z_{ij}$.
- $\mathbf{M}=\frac{1}{ns}\mathbf{D}_p$, término general: $m_j=\frac{n_j}{ns}$.
- $\mathbf{D}=\frac{1}{n}\mathbf{I}_n$, término general: $d_i=\frac{1}{n}$.

La $\mathbf{M}$-distancia al cuadrado entre dos individuos i y l y la $\mathbf{D}$-distancia al cuadrado entre



Coordenadas y ayudas para la interpretación de las categorías para los tres primeros ejes

Cate	Peso	Coordenadas			Contribuciones %			Cosenos2 %			Cinercia
	%	G1	G2	G3	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3	%
Ge.F	7.19	0.13	-0.89	0.51	0.36	18.21	6.66	0.67	31.97	10.58	8.90
Ge.M	17.81	-0.05	0.36	-0.21	0.14	7.35	2.69	0.67	31.97	10.58	3.60
Ed.16om	6.63	-0.39	-1.07	0.29	3.04	24.09	1.99	5.56	41.02	3.07	9.19
Ed.17	9.61	0.19	-0.03	-0.88	0.98	0.03	26.54	2.14	0.06	48.74	7.70
Ed.18	3.15	0.75	0.14	1.13	5.24	0.18	14.21	8.08	0.26	18.38	10.93
Ed.19oM	5.62	-0.27	1.24	0.53	1.24	27.43	5.67	2.16	44.26	8.27	9.69
Es.bajo	10.06	-0.93	0.31	0.18	25.90	3.15	1.17	58.34	6.58	2.21	7.47
Es.medio	10.39	0.67	0.04	0.31	13.71	0.05	3.61	31.59	0.10	6.99	7.30
Es.alto	4.55	0.54	-0.78	-1.12	3.88	8.75	20.06	6.39	13.38	27.73	10.22
Or.Bogo	17.47	0.51	0.13	0.11	13.45	0.91	0.70	60.15	3.76	2.65	3.76
Or.Cund	2.13	-0.88	0.61	-1.44	4.87	2.51	15.72	7.16	3.44	19.43	11.43
Or.Otro	5.39	-1.30	-0.65	0.23	27.18	7.34	0.97	46.66	11.71	1.39	9.80

Figura 5.3: Primer plano factorial del ACM de admitidos, mostrando las categorías.

dos categorías j y k de $\mathbf{X}$ son:

$$d^2(i,l) = \sum_{j=1}^{p} \frac{n_j}{ns} \left(\frac{n}{n_j} z_{ij} - \frac{n}{n_j} z_{lj} \right)^2 = \frac{n}{s} \sum_{j=1}^{p} \frac{1}{n_j} (z_{ij} - z_{lj})^2 \tag{5.8}$$

$$d^2(j,k) = \sum_{i=1}^{n} \frac{1}{n} \left(\frac{n}{n_j} z_{ij} - \frac{n}{n_k} z_{ik} \right)^2 = n \sum_{i=1}^{n} \left(\frac{z_{ij}}{n_j} - \frac{z_{ik}}{n_k} \right)^2 \tag{5.9}$$

Que corresponden a las fórmulas (5.1) y (5.3), respectivamente.

5.3.4. Relaciones cuasibaricéntricas

En el ACM las relaciones de transición son:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{s} \sum_{j=1}^{p} z_{ij} G_s(j) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{s} \sum_{j \in J_i} G_s(j) \tag{5.10}$$

donde J_i es el conjunto de categorías que son 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.

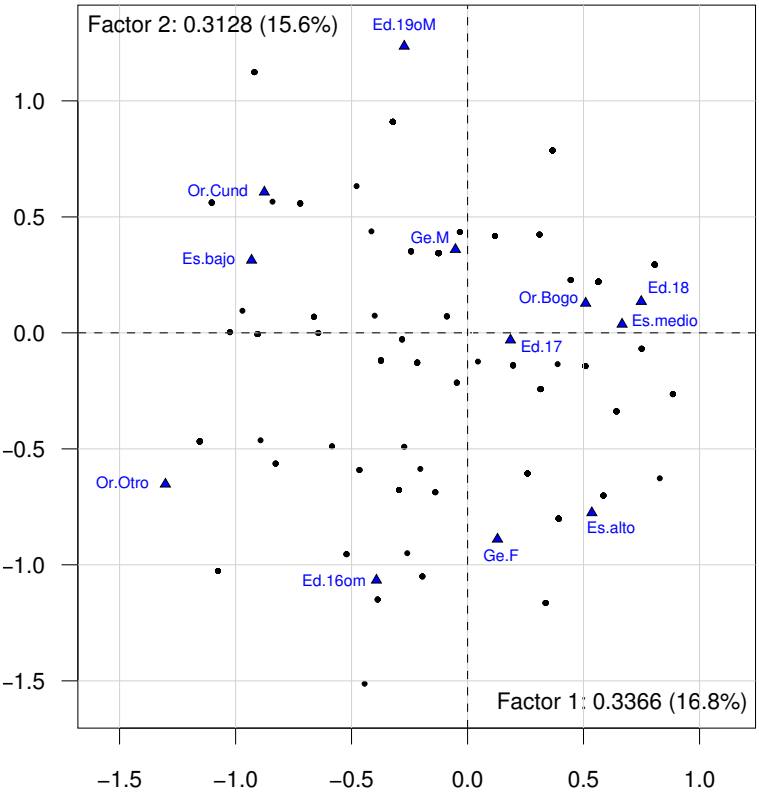
El primer individuo de la tabla 5.1 asume las categorías Ge.F (0.1 es la coordenada sobre el primer eje, que se lee en la tabla incluida en la figura 5.3, Ed.17 (0.2), Es.medio (0.7) y Or.Otro (-1.3), entonces el promedio aritmético es (0.1+0.2+0.7-1.3)/4 = -0.08; la dilatación es $1/\sqrt{0.337} = 1.72$ 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 5.2.

La fórmula es análoga para la coordenada de una categoría:

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{n_j} \sum_{i=1}^{n} z_{ij} F_s(i) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{n_j} \sum_{i \in I_j} F_s(i) \tag{5.11}$$

donde I_j es el conjunto de individuos que asumen la categoría j. Entonces, la 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 permiten la representación simultánea y su interpretación. Para el ejemplo de admitidos a la Facultad de Ciencias el primer plano de esa representación se muestra en la figura 5.4.




```
plot(acm,Trow=FALSE,cframe=1.05,cex.col=0.6);points(acm$li,pch=20,cex=0.6)
```

Figura 5.4: Primer plano factorial del ACM de admitidos, mostrando individuos y categorías.

Es importante pensar que una categoría representa al grupo de individuos que la asumen, de hecho cuando dos categorías de variables diferentes aparecen cerca es porque hay unos cuantos individuos que asumen simultáneamente ambas categorías.

La figura 5.5 muestra los grupos de estudiantes según su origen y las categorías del ACM, donde se puede observar el centro de gravedad verdadero con respecto a los tres puntos de las coordenadas de las categorías de origen, que se han desplazado por el inverso de los valores propios asociados a los dos ejes. Los individuos en ACM generalmente son anónimos pero se manifiestan a través de las categorías que asumen.

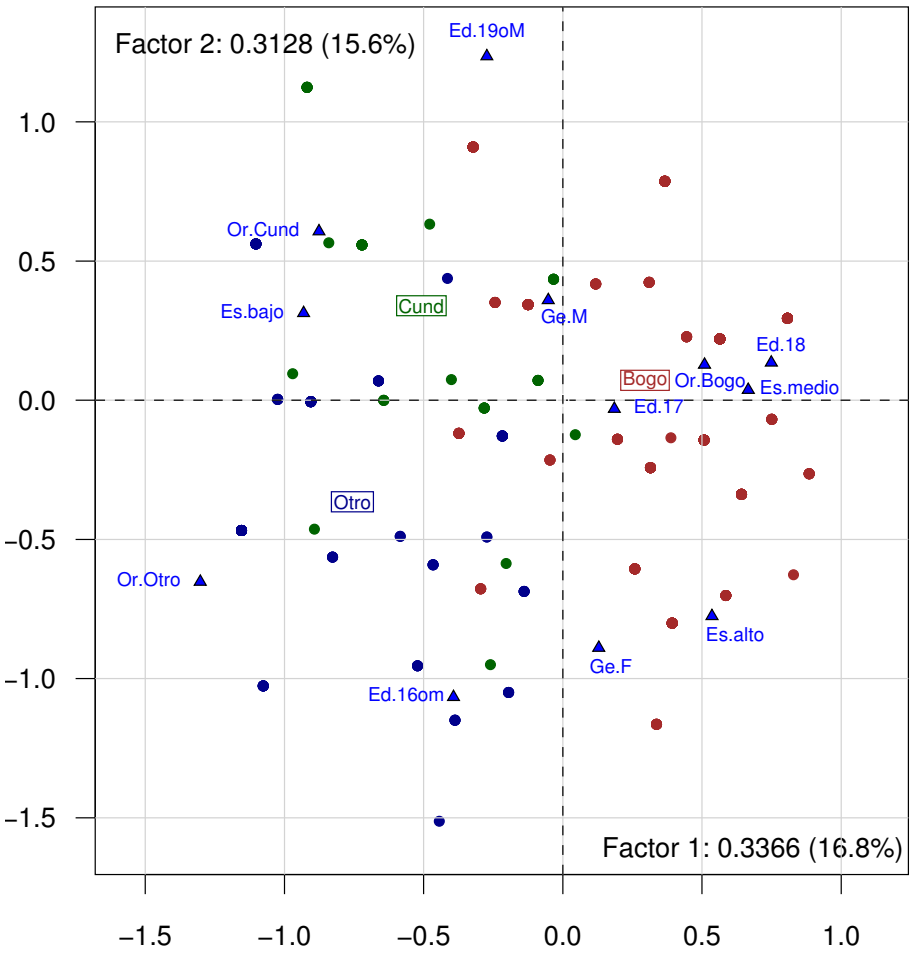


Figura 5.5: Primer plano factorial del ACM de admitidos mostrando los individuos según su origen.

5.3.5. 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 4.3.3). Se adiciona la contribución absoluta de las variables, obtenida como la suma de las contribuciones de sus categorías.

Razón de correlación

Para una variable cualitativa se puede calcular la razón correlación con respecto al 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 que 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:

$$Varianza\ intra(q) = \lambda_s \sum_{i \in J_q} \frac{n_j}{n} G_s^2(j)$$

y la razón de correlación es:

$$\eta_s^2(q) = \sum_{j \in J_q} \frac{n_j}{n} G_s^2(j) \tag{5.12}$$

que se puede expresar como función de la contribución absoluta de las categorías como:

$$\eta_s^2(q) = \lambda_s s \sum_{j \in J_q} Ca_s(j) \tag{5.13}$$

$Ca_s(j)$ es la contribución absoluta de la categoría j sobre el eje s. Estos valores se encuentran en el objeto de salida de la función `dudi.acm{ade4}` , en la tabla `$cr`. Con estas razones de correlación se pueden obtener los planos factoriales para las variables. Por ejemplo para la variable *Origen*, la suma de las contribuciones de las tres categorías (figura 5.3) sobre el primer eje es 13.45+4.87+27.18 = 45.5 %, el primer valor propio es 0.3366 y el número de variables 4, entonces la razón de correlación es $\eta_1^2(Origen) = 4 * 0.3366 * 0.455 = 0.613$, valor que se puede leer en la figura 5.6.

La figura 5.6 muestra la relación de las variables Estrato y Origen con el eje 1 y Edad y Género con el eje 2.

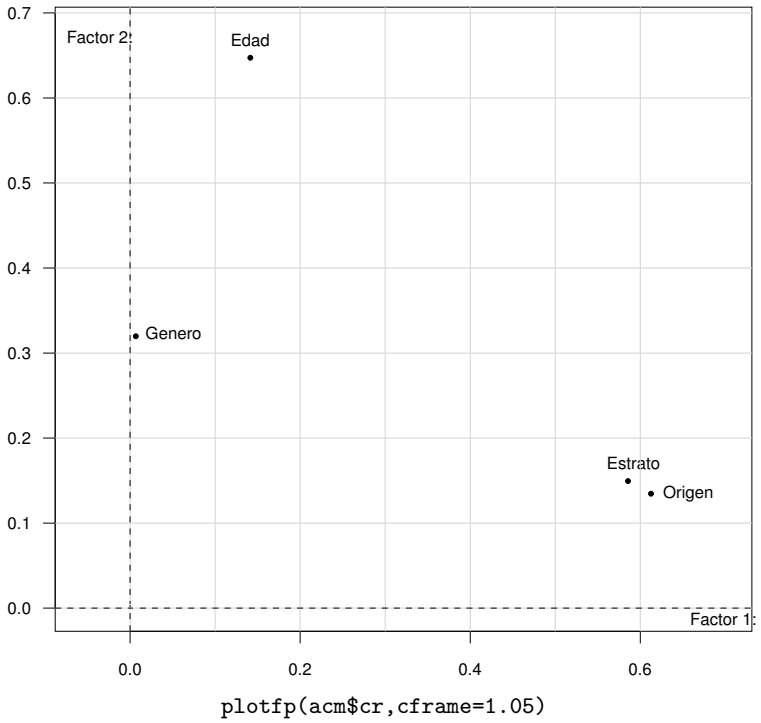


Figura 5.6: Relaciones de correlación de las variables sobre el primer plano factorial del ACM de admitidos. Buena parte de la inercia del primer eje es entre estratos y también, entre origen; y la del segundo eje entre edades. Estrato y origen se muestran relacionados.

5.3.6. Elementos suplementarios

De la misma manera que en el 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 (5.10), es decir que 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 (5.11), 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. No son pruebas de hipótesis estadísticas pero se construyen siguiendo el procedimiento para contrastar una prueba.

Una categoría j es asumida por los n_j individuos del conjunto I_j, entonces su coordenada, sobre un eje s es el promedio aritmético de las coordenadas de esos individuos sobre el eje, multiplicada por el inverso de la raíz cuadrada del valor propio λ_s. Si se supone que los n_j individuos se extraen al azar de los n individuos, la media de las coordenadas es 0 y su varianza es $\left(\frac{n-n_j}{n-1}\right)\frac{\lambda_s}{n_j}$. Entonces la varianza de la categoría j se obtiene multiplicando la varianza anterior por $\frac{1}{\lambda_s}$: $\frac{1}{\lambda_s}\left(\frac{n-n_j}{n-1}\right)\frac{\lambda_s}{n_j} = \frac{n-n_j}{n_j(n-1)}$ y el valor calculado de la estadística T para la verdadera coordenada es:

$$t_s(j) = \sqrt{\frac{n_j(n-1)}{n-n_j}}G_s(j) \tag{5.14}$$

Como ejemplo se muestra el cálculo del valor test para *Química* sobre el primer eje:

```
summary(tab$car)
## Bio Est Far Fis Geo Mat Qui
## 63  66  73  82  45  53  63
```

$$t_1(Qui) = \sqrt{\frac{63(445-1)}{445-63}}(-0.259) = -2.216$$

La diferencia con el valor de la figura 5.7 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.

Código R. Para proyectar carrera del admitido como variable suplementaria:

```
supcar<-supqual(acm,admi[,1])
#dev.new()
plot(acm,Trow=FALSE,cframe=1,ylim=c(-1.2,1.5))
points(supcar$coor,col="darkgreen")
text(supcar$coor,attributes(admi[,1])$levels,col="darkgreen",pos=1,
   cex=0.8)
#dev.print(device = xfig,file="ACMadmiCarreraSup.fig")
xtable(cbind(ncat=supcar$ncat,d2=supcar$dis2,supcar$coor,supcar$tv,
   supcar$cos2),digits = rep(3,12))
```

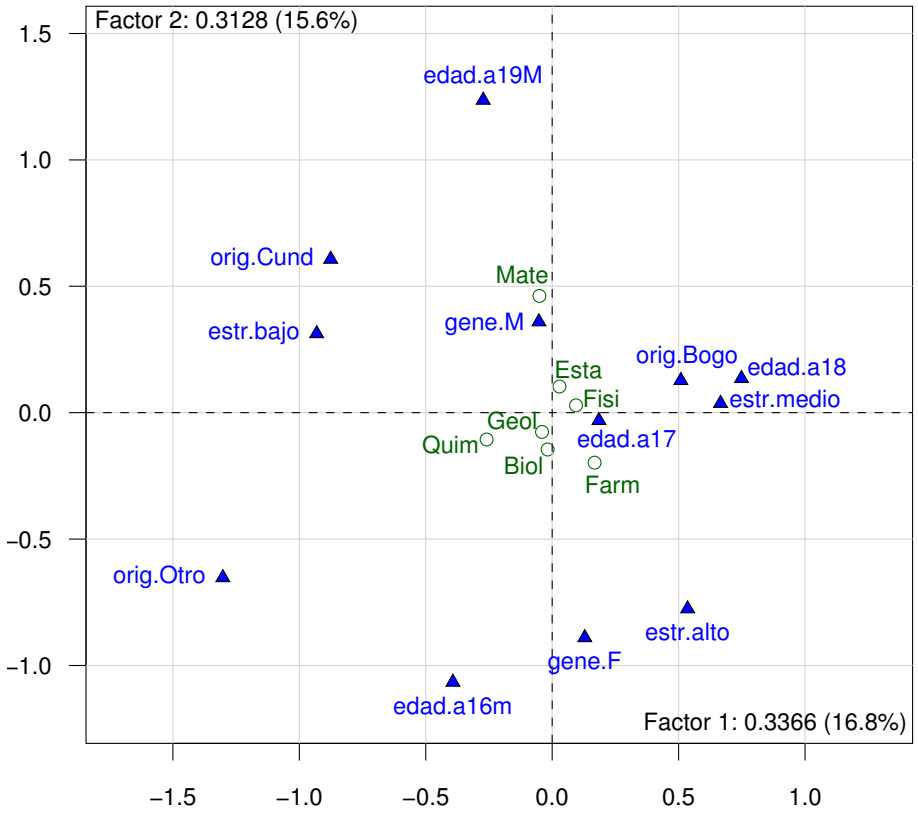
Los valores test de la figura 5.7 indican que solo es legítimo interpretar como diferente de cero las coordenadas de: Química (-) sobre el primer eje, Matemáticas (+) sobre el segundo eje; y Geología (-) opuesto a Farmacia (+) sobre el tercer eje. Matemáticas tiene en comparación al promedio, mayor proporción de admitidos de 19 o más años, Geología de 17 años y de 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 1.8.

A veces se calculan los valores test para las variables activas, para interpretarlos en aquellas que tienen baja contribución a los ejes, porque su comportamiento sería como ilustrativas.

A una variable cualitativa suplementaria también se le puede calcular la relación de correlación con la primera igualdad de la fórmula (5.12) e incluirla en la gráfica de las variables. En el ejemplo se proyecta la variable Carrera, que se presenta en la figura 5.8. Su posición, cerca al origen, muestra poca relación con las cuatro variables sociodemográficas.

Variables continuas

Las variables continuas se pueden proyectar en los planos simultáneos de individuos y categorías, utilizando como coordenadas los coeficientes de correlación entre la variable y el factor. Se interpretan como en un ACP.



Coordenadas y ayudas para la interpretación

Carrera	Admitidos	dis $\tan^2$	Coordenadas			Valores test			Cosenos cuadrados		
			Eje1	Eje2	Eje3	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3
Biol	63	6.063	-0.018	-0.146	-0.068	-0.156	-1.245	-0.579	0.000	0.003	0.001
Esta	66	5.742	0.029	0.104	-0.018	0.255	0.912	-0.160	0.000	0.002	0.000
Farm	73	5.096	0.168	-0.198	0.392	1.566	-1.845	3.664	0.006	0.008	0.030
Fisi	82	4.427	0.095	0.029	-0.127	0.948	0.292	-1.267	0.002	0.000	0.004
Geol	45	8.889	-0.040	-0.076	-0.635	-0.284	-0.540	-4.490	0.000	0.001	0.045
Mate	53	7.396	-0.050	0.462	0.112	-0.387	3.579	0.869	0.000	0.029	0.002
Quim	63	6.063	-0.259	-0.106	0.156	-2.217	-0.907	1.335	0.011	0.002	0.004

Figura 5.7: Primer plano factorial mostrando las carreras como categorías suplementarias y ayudas para la interpretación.

5.3.7. Retorno a los datos

En la aplicación del ACM y los otros métodos factoriales siempre conviene regresar a los datos para observar lo que los mapas factoriales revelan, y a veces mostrarlos de otra forma. Eso es lo que se ha hecho en el ejemplo, al mostrar los perfiles de las carreras según las cuatro variables activas, y al presentar los diagramas de caja y bigotes del examen según estrato y origen.

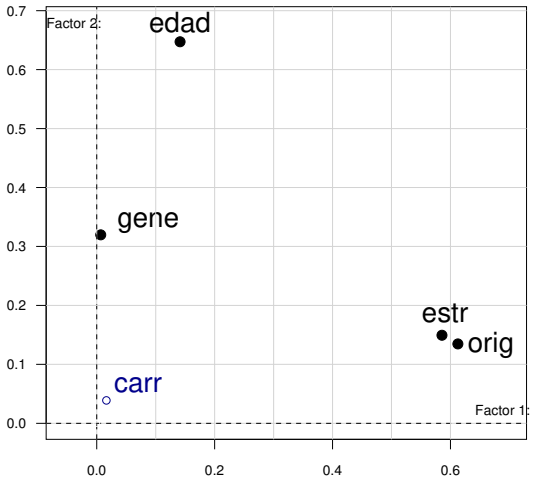


Figura 5.8: Variables activas e ilustrativas sobre el primer plano factorial del ACM de los admitidos.

5.4. Comparación del ACM con otros AC de la misma tabla

5.4.1. 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: $\lambda_{\mathbf{B}} = \lambda_Z^2$
- Las coordenadas factoriales del AC de $\mathbf{B}$, $G_{\mathbf{B}}$, son homotecias (contraídas por $\sqrt{\lambda_Z}$) de las de las categorías del ACM: $G_{\mathbf{B}} = \sqrt{\lambda_Z} G$

5.4.2. 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 las dos variables directamente, es decir de la tabla de n individuos por las dos variables. Lebart et al. (2006, p. 203) analiza ese caso y aquí se presentan 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 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 y para los vectores de 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:

- Valores propios: $\lambda_{\mathbf{Z}} = \dfrac{1+\sqrt{\lambda}}{2}$
- Ejes factoriales de las coordenadas, si $p_2 \leq p_1$:
  - $p_2 - 1$ factores del tipo $\begin{pmatrix} F \\ G \end{pmatrix}$, correspondiente a los valores propios $\dfrac{1+\sqrt{\lambda}}{2}$
  - $p_2 - 1$ factores del tipo $\begin{pmatrix} F \\ -G \end{pmatrix}$, correspondiente a los valores propios $\dfrac{1-\sqrt{\lambda}}{2}$
  - $p_1 - p_2$ factores del tipo $\begin{pmatrix} \mathbf{0} \\ \mathbf{a} \end{pmatrix}$, correspondiente a los valores propios $\dfrac{1}{2}$

    donde los valores $\mathbf{a}$ complementan la base en $\mathbb{R}^p$.

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.

5.4.3. El criterio de Benzécri para seleccionar el número de ejes en el ACM

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:

$$\tau(\lambda_Z) = \left(\frac{s}{s-1}\right)^2 \left(\lambda_Z - \frac{1}{s}\right)^2 \quad para \quad \lambda_Z > \frac{1}{s} \tag{5.15}$$

Obsérvese que en el caso del ACM con dos variables $\tau(\lambda_Z) = \lambda$, donde λ representa el correspondiente valor propio de la tabla de contingencia que cruza las dos variables y $\lambda_{\mathbf{Z}} = \dfrac{1+\sqrt{\lambda}}{2}$.

El histograma de las tasas de inercia (5.15) se puede usar, en lugar del histograma de valores propios, para decidir el número de ejes a retener en un ACM.

5.5. Ejemplo de aplicación de ACM

Una de las principales aplicaciones del ACM es en el análisis de encuestas y de archivos de tipo administrativo o de transacciones. A partir de la información de clientes, de empleados, de proveedores, etc., se pueden construir gráficos que permitan orientar las decisiones de las entidades y compañías. Se presenta un ejemplo simplificado utilizando la Encuesta de Consumo Cultural del Dane de 2014.

Frecuencia de lectura de niños colombianos entre 5 y 11 años en Colombia

El objetivo del análisis, en el ejemplo, es describir la frecuencia de lectura de los niños entre 5 y 11 años presentes en la muestra y explorar su relación con algunas variables sociodemográficas. La tabla construida para este ejercicio tiene 3476 niños de todo el país. Se toman como activas las variables de frecuencia de lectura y otras que se consideran muy relacionadas: *Leer* el niño sabe leer; asistencia a: *Teat* teatro, *Conc* conciertos, *Cine*; lectura de: *Libr* libros, *Revi* revistas; *Peri* periódicos.

Código R. Para leer datos, definir variables activas e ilustrativas y obtener sus distribuciones de frecuencias.

```
load("ninios5a11.Rda")
# variables activas
Y <- ninios[,c(3,6,13,19:22)]
summary(Y)
# variables suplementarias
Ys<-ninios[,c(1,2,4,5,29,30,32,35)]
summary(Ys)
```

Las frecuencias de las categorías de las variables activas son:

```
> summary(Y)
 Leer        Teat       Tite        Libr         Revi         Peri         Cine
 si:2823   sema:  89   Si:1079   diar: 519   diar:  83   vdse: 284   mens: 310
 no: 653   mens: 141   No:2397   vdse: 844   vdse: 279   sema: 256   trim: 332
           trim: 209             sema: 391   sema: 274   mens: 154   seme: 260
           seme: 170             mens: 159   mens: 198   no  :2782   anua: 259
           anua: 257             trim:  91   anua:  56               no  :2315
           no  :2610             anua:  68   no  :2586
                                 no  :1404
```

Se tienen entonces $s = 7$ variables activas con un total de $p = 32$ categorías. Se puede observar que las frecuencias de no realizar las actividades descritas es muy alta, siendo inferior la de leer libros.

A continuación se presentan las variables que se utilizan como suplementarias, con las distribuciones de frecuencias:

```
> summary(Ys)
  Etnia        Pare       Escu        Nivel         Sexo       EdadG        Regi       Estr
 mest:1938   hijo:2544   si:3369   ning:  89   masc:1750   a5 :470   Atla:517   0o1:1448
 blan: 916   niet: 730   no: 107   pree: 619   feme:1726   a6 :471   Orie:752   2  :1350
 otro: 508   otro: 202             prim:2561               a7 :504   Cent:592   3oM: 678
 nr  : 114                         secu: 207               a8 :497   Paci:538
                                                           a9 :488   Bogo:545
                                                           a10:521   OrAm:532
                                                           a11:525
```

Para realizar el ACM usamos las 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.

Código R. Para obtener las gráficas y tabla la figura 5.9.

```
library(FactoClass)
acm<-dudi.acm(Y,scannf = FALSE,nf=3)
barplot(acm$eig,las=3)
#dev.print(device = xfig)
eigtab<-data.frame(valp=acm$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 Benzécri
s<-7; 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)*10
barplot(ptau,las=3)
#dev.print(device = xfig)
```

Número de ejes a interpretar

La primera decisión a tomar es el número de ejes a interpretar. No hay recetas pero si 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, sería el número de ejes a analizar. 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 los histogramas: de valores propios, del criterio de Benzécri; y los valores propios y porcentajes de inercia en la figura 5.9. Los gráficos sugieren analizar tres ejes. El criterio de Benzécri se construye considerando los 11 valores propios, que tienen un valor superior a $1/7 = 0.1429$.

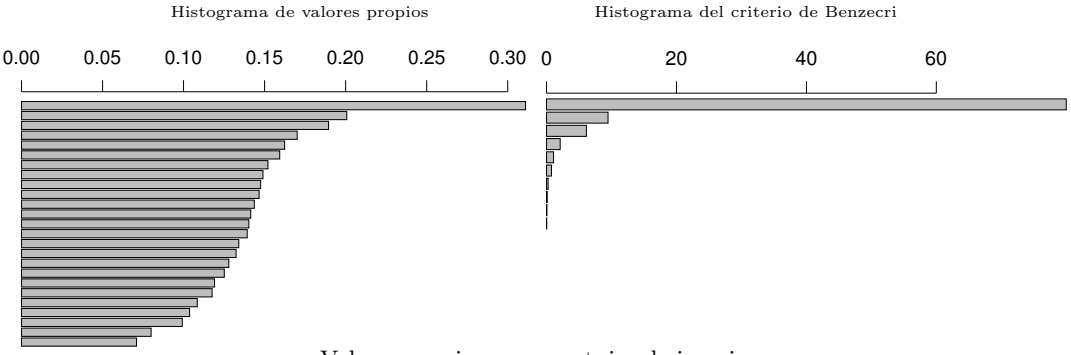
Primer eje factorial

Una guía para saber en qué categorías poner atención en cada uno de los ejes es ver aquellos que superen el porcentaje promedio: en este ejemplo $100/32 = 3.125$. Las coordenadas y ayudas para la interpretación están en la tabla 5.5.

El primer eje factorial contrapone las dos categorías le lectura: si (+), no (-); ir a títeres: si (+), leer libros: diaramente (+), no (-); leer revistas: varias veces a la semana y semanalmente (+), no(-); leer periódicos: varios días a la semana y semanalmente (+), no (-). Es decir que separa los que no leen, al lado negativo, de los que leen con más frecuencia al lado positivo.

Código R. Para obtener las coordenadas y ayudas para la interpretación (tabla 5.5).

```
ayu<-inertia.dudi(acm,,T)
```



Valores propios y porcentajes de inercia

eje	valp	porc	pacu	eje	valp	porc	pacu	eje	valp	porc	pacu
1	0.311	8.7	8.7	9	0.148	4.1	46.0	17	0.128	3.6	76.9
2	0.201	5.6	14.3	10	0.147	4.1	50.1	18	0.125	3.5	80.4
3	0.189	5.3	19.6	11	0.144	4.0	54.1	19	0.119	3.3	83.8
4	0.170	4.8	24.4	12	0.141	4.0	58.1	20	0.118	3.3	87.1
5	0.162	4.5	28.9	13	0.140	3.9	62.0	21	0.108	3.0	90.1
6	0.159	4.5	33.4	14	0.139	3.9	65.9	22	0.104	2.9	93.0
7	0.152	4.3	37.7	15	0.134	3.8	69.6	23	0.099	2.8	95.8
8	0.149	4.2	41.8	16	0.132	3.7	73.3	24	0.080	2.2	98.0

Figura 5.9: Histogramas de valores propios y del criterio de Benzecri y tabla de valores propios del ACM de frecuencia de lectura de niños.

```
xtable(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))
```

Segundo eje factorial

El segundo eje contrapone en la frecuencia de lectura de libros: diaria (+) con frecuencia semanal y menores (-); de revistas: diaria (+) con las demás frecuencias (-); de periódicos: varios días a la semana (+) con semanal y mensual (-). Es decir que este eje contrapone la frecuencia de lecturas más alta de las demás.

Tercer eje factorial

El tercer eje destaca en el lado positivo al grupo de niños que no saben leer (-) y opone en las frecuencias de asistencia a teatro mensual (-) de la no asistencia; de la asistencia a títeres (-) a la no asistencia (+); de lectura de libros varios días a la semana (+) con lectura anual (-). Destaca también la lectura de revistas y periódicos varios días a la semana (+) y asistencia a cine mensual (-).

Tabla 5.5: 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
Leer.si	11.6	0.293	-0.102	0.159	3.2	0.6	1.6	37.2	4.5	10.9	0.8
Leer.no	2.7	-1.268	0.441	-0.688	13.9	2.6	6.7	37.2	4.5	10.9	3.2
Teat.sema	0.4	0.768	0.544	-0.724	0.7	0.5	1.0	1.6	0.8	1.4	3.9
Teat.mens	0.6	0.663	0.754	-1.577	0.8	1.6	7.6	1.9	2.4	10.5	3.8
Teat.trim	0.9	0.953	0.107	-0.708	2.5	0.1	2.3	5.8	0.1	3.2	3.8
Teat.seme	0.7	0.887	0.039	-0.397	1.8	0.0	0.6	4.0	0.0	0.8	3.8
Teat.anua	1.1	0.697	0.031	-0.443	1.6	0.0	1.1	3.9	0.0	1.6	3.7
Teat.no	10.7	-0.265	-0.074	0.236	2.4	0.3	3.1	21.1	1.6	16.8	1.0
Tite.Si	4.4	0.520	0.203	-0.868	3.9	0.9	17.6	12.2	1.9	33.9	2.8
Tite.No	9.9	-0.234	-0.091	0.391	1.7	0.4	7.9	12.2	1.9	33.9	1.2
Libr.diar	2.1	0.893	0.903	-0.353	5.5	8.7	1.4	14.0	14.3	2.2	3.4
Libr.vdse	3.5	0.495	0.190	0.687	2.7	0.6	8.7	7.9	1.2	15.2	3.0
Libr.sema	1.6	0.351	-1.039	0.159	0.6	8.6	0.2	1.6	13.7	0.3	3.5
Libr.mens	0.7	0.571	-1.121	-0.254	0.7	4.1	0.2	1.6	6.0	0.3	3.8
Libr.trim	0.4	0.438	-1.781	0.517	0.2	5.9	0.5	0.5	8.5	0.7	3.9
Libr.anua	0.3	0.319	-1.673	-0.329	0.1	3.9	0.2	0.2	5.6	0.2	3.9
Libr.no	5.8	-0.834	0.165	-0.316	12.9	0.8	3.0	47.1	1.8	6.8	2.4
Revi.diar	0.3	1.525	2.185	-0.986	2.5	8.1	1.8	5.7	11.7	2.4	3.9
Revi.vdse	1.1	1.169	1.280	1.300	5.0	9.4	10.2	11.9	14.3	14.8	3.7
Revi.sema	1.1	1.051	-0.888	0.085	4.0	4.4	0.0	9.4	6.8	0.1	3.7
Revi.mens	0.8	1.065	-1.326	-0.409	3.0	7.1	0.7	6.8	10.6	1.0	3.8
Revi.anua	0.2	0.740	-2.517	-0.909	0.4	7.3	1.0	0.9	10.4	1.4	3.9
Revi.no	10.6	-0.384	0.042	-0.067	5.0	0.1	0.2	42.8	0.5	1.3	1.0
Peri.vdse	1.2	1.337	1.316	1.233	6.7	10.1	9.4	15.9	15.4	13.5	3.7
Peri.sema	1.1	1.300	-0.684	-0.241	5.7	2.5	0.3	13.4	3.7	0.5	3.7
Peri.mens	0.6	0.989	-1.586	-0.517	2.0	7.9	0.9	4.5	11.7	1.2	3.8
Peri.no	11.4	-0.311	0.016	-0.075	3.5	0.0	0.3	38.7	0.1	2.3	0.8
Cine.mens	1.3	0.851	0.588	-1.009	3.0	2.2	6.8	7.1	3.4	10.0	3.6
Cine.trim	1.4	0.556	-0.027	-0.350	1.4	0.0	0.9	3.3	0.0	1.3	3.6
Cine.seme	1.1	0.171	-0.437	-0.289	0.1	1.0	0.5	0.2	1.5	0.7	3.7
Cine.anua	1.1	0.344	-0.232	-0.200	0.4	0.3	0.2	0.9	0.4	0.3	3.7
Cine.no	9.5	-0.251	0.000	0.240	1.9	0.0	2.9	12.6	0.0	11.5	1.3

Primer plano factorial

En la figura 5.10 se muestra el primer plano factorial con categorías activas. A la izquierda se ubican los niños que no leen, abajo a la derecha los que leen con una frecuencia semanal o menor y arriba a la derecha los que leen diariamente o varios días a la semana. Los niños que más leen tienden a ir con más frecuencia a teatro y cine.

Variables suplementarias

En la figura 5.11 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

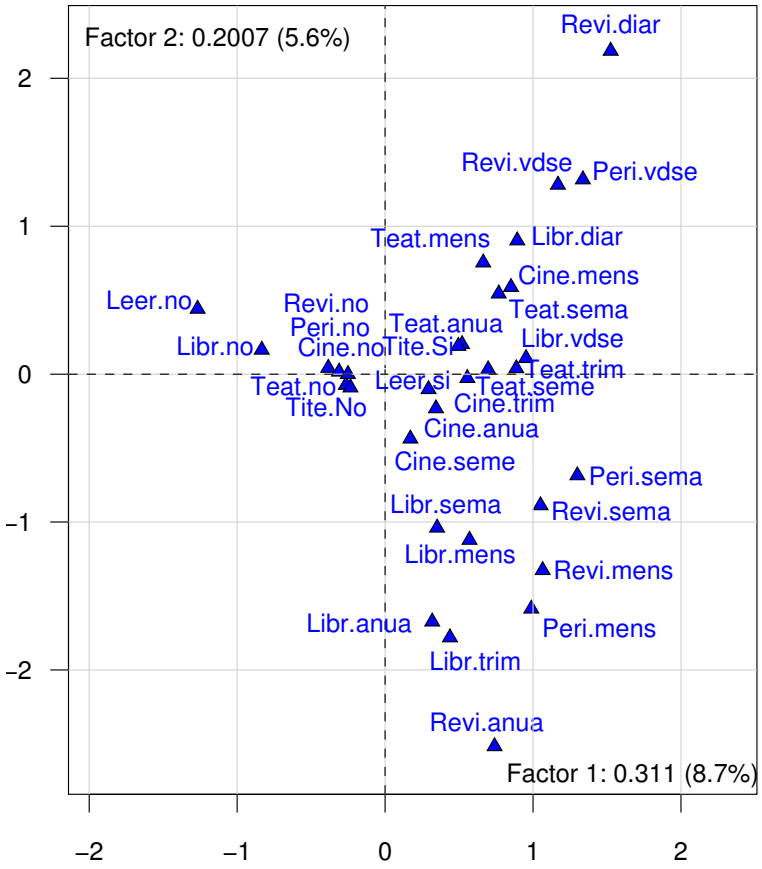


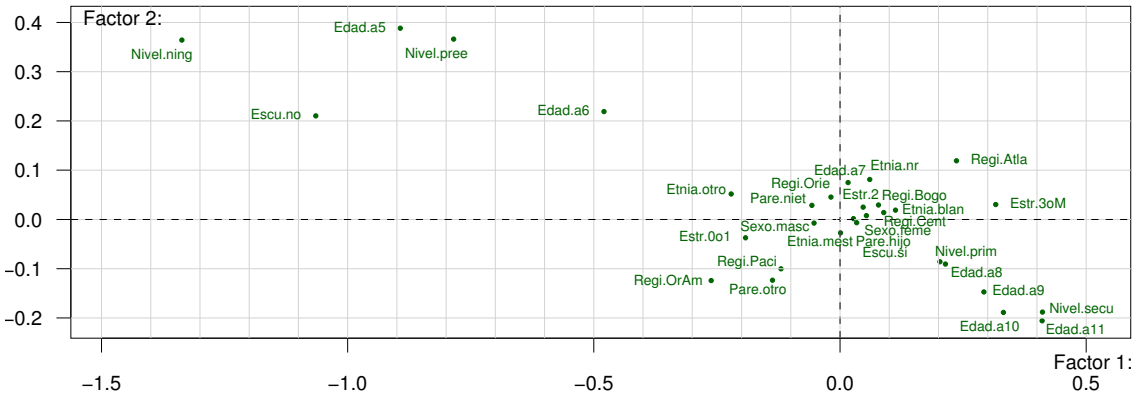
Figura 5.10: Primer plano factorial del ACM de frecuencia de lectura de niños, mostrando las categorías activas.

del centro de la figura 5.10, se observa que hay un ordenamiento de los niños según su edad y nivel de escolaridad, más o menos en el sentido de la primera bisectriz, explicando en parte la no lectura de la lectura intermedia, pero no explicando los niveles de lectura más frecuentes. También explican las frecuencias de lectura: el estrato (0o1, 2 y 3oM) y la región: con mayor lectura en la Atlántica y menor en la Oriental y Pacífica.

En el eje 3 se muestra que parte de la frecuencia de asistencia a teatro y a cine y a la lectura de revistas se explica por la región Atlántica y el estrato 3 o más.

Relación entre variables y ejes

En la figura 5.12 se muestran el primer plano factorial de las variables cualitativas activas (arriba) y suplementarias (abajo), obtenidas a partir de las razones de correlación entre



Coordenadas y ayudas para la interpretación

Cate	%	dis	Coordenadas			Valores test			Cosenos cuadrados		
goria		$\tan^2$	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3	Eje1	Eje2	Eje3
Etnia.mest	55.8	0.794	0.001	-0.027	0.043	0.074	-1.802	2.819	0.000	0.001	0.002
Etnia.blan	26.4	2.795	0.113	0.019	-0.199	3.975	0.665	-7.022	0.005	0.000	0.014
Etnia.otro	14.6	5.843	-0.221	0.052	0.184	-5.392	1.261	4.488	0.008	0.000	0.006
Etnia.nr	3.3	29.491	0.060	0.081	0.056	0.656	0.880	0.608	0.000	0.000	0.000
Pare.hijo	73.2	0.366	0.027	0.002	0.003	2.652	0.155	0.249	0.002	0.000	0.000
Pare.niet	21.0	3.762	-0.057	0.029	-0.040	-1.732	0.870	-1.226	0.001	0.000	0.000
Pare.otro	5.8	16.208	-0.137	-0.124	0.114	-2.006	-1.808	1.663	0.001	0.001	0.001
Escu.si	96.9	0.032	0.034	-0.007	0.008	11.185	-2.209	2.705	0.036	0.001	0.002
Escu.no	3.1	31.486	-1.065	0.210	-0.257	-11.185	2.209	-2.705	0.036	0.001	0.002
Nivel.ning	2.6	38.056	-1.337	0.364	-0.524	-12.774	3.482	-5.003	0.047	0.003	0.007
Nivel.pree	17.8	4.616	-0.785	0.366	-0.541	-21.533	10.049	-14.850	0.133	0.029	0.063
Nivel.prim	73.7	0.357	0.203	-0.086	0.123	20.009	-8.479	12.092	0.115	0.021	0.042
Nivel.secu	6.0	15.792	0.411	-0.188	0.326	6.101	-2.791	4.843	0.011	0.002	0.007
Sexo.masc	50.3	0.986	-0.053	-0.008	-0.012	-3.134	-0.446	-0.717	0.003	0.000	0.000
Sexo.feme	49.7	1.014	0.054	0.008	0.012	3.134	0.446	0.717	0.003	0.000	0.000
Edad.a5	13.5	6.396	-0.893	0.388	-0.549	-20.820	9.054	-12.790	0.125	0.024	0.047
Edad.a6	13.6	6.380	-0.479	0.219	-0.276	-11.189	5.115	-6.453	0.036	0.008	0.012
Edad.a7	14.5	5.897	0.016	0.075	0.047	0.395	1.814	1.146	0.000	0.001	0.000
Edad.a8	14.3	5.994	0.214	-0.091	0.110	5.154	-2.182	2.652	0.008	0.001	0.002
Edad.a9	14.0	6.123	0.292	-0.147	0.124	6.959	-3.507	2.959	0.014	0.004	0.003
Edad.a10	15.0	5.672	0.332	-0.189	0.235	8.217	-4.677	5.809	0.019	0.006	0.010
Edad.a11	15.1	5.621	0.411	-0.206	0.241	10.207	-5.122	6.001	0.030	0.008	0.010
Regi.Atla	14.9	5.723	0.237	0.119	-0.242	5.830	2.936	-5.963	0.010	0.002	0.010
Regi.Orie	21.6	3.622	-0.018	0.045	0.287	-0.572	1.407	8.898	0.000	0.001	0.023
Regi.Cent	17.0	4.872	0.089	0.014	-0.076	2.376	0.376	-2.023	0.002	0.000	0.001
Regi.Paci	15.5	5.461	-0.120	-0.100	-0.109	-3.022	-2.532	-2.760	0.003	0.002	0.002
Regi.Bogo	15.7	5.378	0.078	0.029	0.037	1.987	0.743	0.938	0.001	0.000	0.000
Regi.OrAm	15.3	5.534	-0.262	-0.124	-0.014	-6.560	-3.109	-0.345	0.012	0.003	0.000
Estr.0o1	41.7	1.401	-0.192	-0.037	0.164	-9.562	-1.862	8.191	0.026	0.001	0.019
Estr.2	38.8	1.575	0.047	0.025	-0.017	2.212	1.170	-0.795	0.001	0.000	0.000
Estr.3oM	19.5	4.127	0.316	0.030	-0.317	9.176	0.878	-9.213	0.024	0.000	0.024

Figura 5.11: Proyección de categorías suplementarias sobre el primer plano factorial y ayudas para su interpretación

variables y ejes. El eje uno tiene alguna relación con todas las variables pero mucho mayor con frecuencias de lectura de libros, revistas y periódicos, variables que están también relacionadas con el eje 2.

Las variables cualitativas suplementarias edad y nivel de escolaridad son las que presentan alguna relación importante con el eje 1 y se debe a que a más años y escolaridad los niños tienden a leer con más frecuencia.

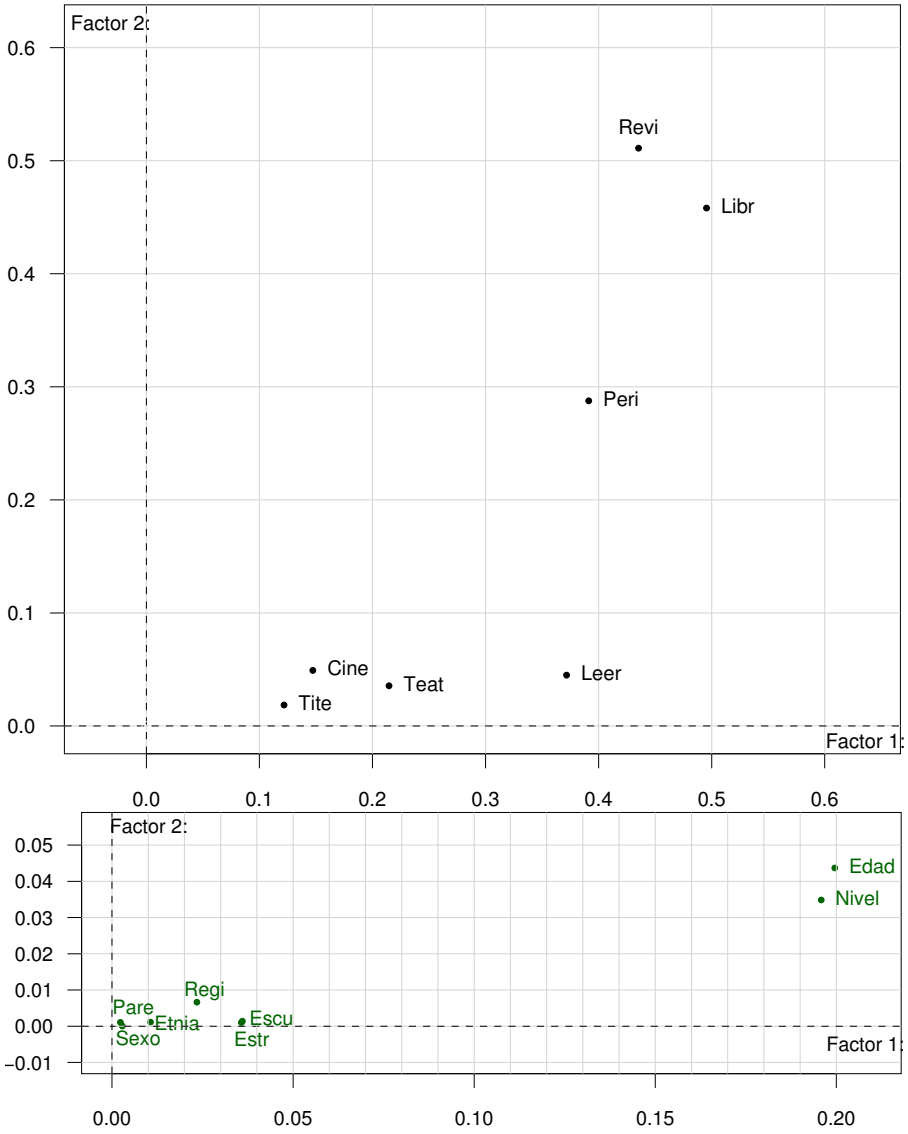


Figura 5.12: Primer plano factorial mostrando las variables activas e ilustrativas.

Resumen del ACM de frecuencias de lectura de niños

La mayoría de niños colombianos de 5 a 11 años, presentes en la muestra de Consumo Cultural, realizada por el Dane en 2014, no leen o lo hacen con poca frecuencia. Los niños

se ordenan por la frecuencia de lectura de libros, revistas y periódicos, los que leen con mayor frecuencia también tienden a asistir a cine y a teatro. En el primer plano factorial se pueden establecer tres grupos: no leen (a la izquierda), los que leen con frecuencia semanal o menor (abajo a la derecha) y los que leen con frecuencia diaria o varios días a la semana (arriba a la derecha).

La frecuencia de lectura se explica, en parte, por la escolaridad y la edad de los niños; también por el estrato socioeconómico. En la región Atlántica se incrementa la proporción de niños que leen con mayor frecuencia y en las regiones Orinoquía-Amazonía y Pacífica se incrementa la proporción de niños que no leen.

Retorno a los datos

Como retorno a los datos se muestran los perfiles de frecuencia de lectura de libros para regiones, estratos y escolaridad en la figura 5.13. Las regiones quedan ordenadas por niveles de lectura así: Atlántica, Bogotá, Oriental, Central, Pacífica y Orinoquía-Amazonía. Las frecuencias de lectura aumentan con el estrato y la escolaridad.

5.6. 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 $d^2(j,k) = \dfrac{n}{n_j n_k}(b+c)$ donde b (respectivamente, c) es el número de individuos que asumen la categoría j (k) pero no la k (j), es decir que $(b+c)$ 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.

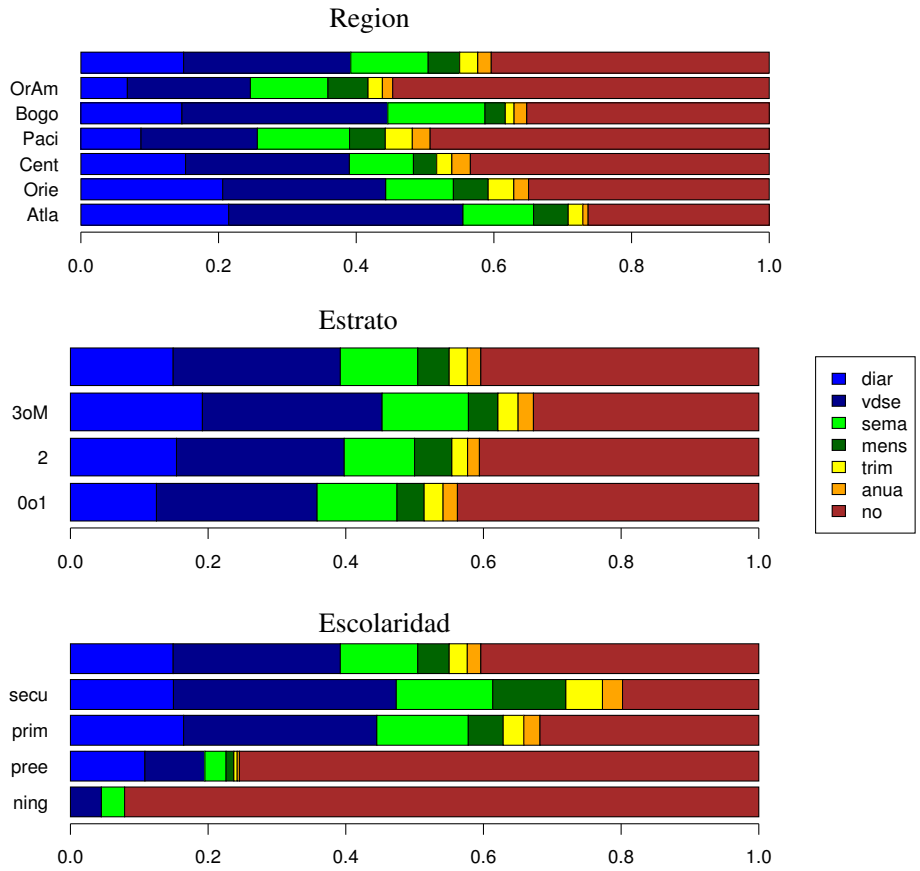


Figura 5.13: Perfiles de regiones, estratos y escolaridad según frecuencia de lectura de libros.

6. Obtenga e interprete las inercias: de una categoría, de una variable y la inercia 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 (5.12).

5.7. Talleres de ACM

5.7.1. Taller ACM: razas de perros

Este taller se encuentra en Fine (1996) y es un ejemplo pequeño para entender los conceptos del ACM.

Objetivo

El objetivo del estudio es seleccionar las razas de perros de acuerdo con la función para la que se utilizan: compañia, caza o utilidad (salvamento, defensa, perro para ciego o perro de policía, etc.). Para cada una de las 27 razas estudiadas se registran seis variables que miden las cualidades físicas o psíquicas de 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

Datos

Los datos se encuentran en el archivo `perros.txt`, también se encuentran en el paquete `FactoClass` como `BreedsDogs`. Realice el análisis utilizando el programa de su preferencia 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) Identifica 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 Boxer 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 afectuosos?

   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 las mismas?

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 observada 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.

5.7.2. Taller: comparación de análisis de correspondencias

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 cada una de las fórmulas dadas en Lebart, Piron & Morineau (2006) para la comparación se cumplen. Escriba la tabla 5.2 reemplazando las fórmulas por los resultados numéricos del ejercicio. Haga un resumen de la comparación. Responda las preguntas siguientes:

Para el ACS de la TC

1. Las coordenadas del centro de gravedad de la nube de los perfiles fila son:

   _____

2. La nube está soportada en: _____dimensiones.

3. La inercia asociada al ACS es: _____

   ¿Qué significa? _____

4. El porcentaje de inercia retenido por el primer plano es: _____%.

   Para hacer un ACM equivalente al ACS

5. La TDC tiene _____filas y _____columnas.

6. En el ACM de la TDC la nube de "individuos" tiene _____puntos y la nube de categorías tiene _____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 al la TDC tiene _____filas y _____columnas y su total es:_____
10. Según el criterio de Benzecri en el ACM se deben tener en cuenta _____ejes y el primer plano retiene el _____% de la inercia.

ACM y ACS 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.

Criterio de Benzécri para los porcentaje de inercia

Realice el histograma del Benzécri para el ACM de *razas de perros* y utilícelo en la decisión de cuántos ejes interpretar.

Capítulo 6

Métodos de clasificación

Los métodos de clasificación que se abordan en este documento, son los denominados en inglés: *Cluster Analysis*, que se pueden traducir como métodos de agrupamiento y 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 que emerge de los datos.

En primera instancia, se conocen dos tipos de métodos:

1. Los que permiten obtener una partición directa mediante un algoritmo, entre los que 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 los van uniendo en pasos sucesivos hasta llegar a un solo grupo o clase de n individuos.

Los algoritmos de clasificación requieren de 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 et al. 2006). Para el estudio general de los métodos de agrupamiento se recomienda el texto de Everitt et al. (2011).

6.1. Métodos para obtener una partición directa

En estos algorítmos se da 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*, el cual 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, la inercia se puede descomponer en inercia entre-clases e inercia intra-clases, lo que se muestra en la sección siguiente.

6.1.1. Descomposición de la inercia asociada a una partición

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 intra-clases.

$$Inercia(N_n) = \sum_{i=1}^{n} p_i d^2(i, \mathbf{g}) = \sum_{k=1}^{K} p_k d^2(\mathbf{g}_k, \mathbf{g}) + \sum_{k=1}^{K} \sum_{i \in I_k} p_i d^2(i, \mathbf{g}_k) \tag{6.1}$$

donde:

p_i**:** peso del individuo i, $\sum_{i=1}^{n} p_i = 1$

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, $\mathbf{g}_k = \sum_{i \in I_k} p_i \mathbf{x}_i$

$d^2(.,.)$ es la distancia Euclidiana canónica.

El primer término de la fórmula (6.1) es la inercia entre-clases y el segundo término la inercia intra-clases. En el cálculo de la inercia ha intervenido la distancia Euclidiana canónica, entonces la medida de disimilitud entre individuos ya está seleccionada.

El método *K-means*, como se ve en la sección siguiente, busca una partición en K clases que tenga inercia intra-clases mínima.

6.1.2. Un método de agregación alrededor de centros de gravedad 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, para entender la lógica e interpretación del *K-means*; los algoritmos implementados en los programas estadísticos son un poco más complejos, ya que están optimizados por expertos en cálculo numérico.

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 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 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 $\{I_1^{m-1}, I_2^{m-1}, \ldots, I_k^{m-1}, \ldots, I_K^{m-1}\}$. 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 varianza intra-clases 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 intra-clases

Hay que mostrar que la inercia intra-clases de la partición $P^m = \{I_1^m, I_2^m, \ldots, I_k^m, \ldots, I_K^m\}$ es menor o igual a la inercia intra-clases 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

$$v(m) = \sum_{k=1}^{K} \sum_{i \in I_k^m} p_i d^2(i, C_k^m) \tag{6.2}$$

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 intra-clases en la etapa m es la cantidad

$$V(m) = \sum_{k=1}^{K} \sum_{i \in I_k^m} p_i d^2(i, C_k^{m+1}) \tag{6.3}$$

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

$$\mathbf{v}(m+1) = \sum_{k=1}^{K} \sum_{i \in I_k^{m+1}} p_i d^2(i, C_k^{m+1}) \tag{6.4}$$

$V(m)$ de (6.3) es menor o igual que $v(m)$ de (6.2) porque la inercia con respecto al centro de gravedad es siempre menor o igual a la inercia con respecto a cualquier otro punto.

$v(m+1)$ de (6.4) es menor o igual a $V(m)$ de (6.3) porque si al menos un individuo cambia de clase en la nueva partición es porque queda más cerca de otro nuevo centro.

Entonces $v(m+1) \leq V(m) \leq v(m) \leq V(m-1)$, es decir que la inercia intra-clases disminuye con cada paso del algoritmo.

Ejemplo Café

Como "ejemplo de juego" vamos a utilizar el ejemplo Café del capítulo 2, partiendo de las coordenadas sobre el primer plano factorial. Buscaremos dos clases ejecutando "a mano" el procedimiento *K-means* partiendo de los cafés: ExCl y O40C, como centros iniciales. En la tabla 6.1 se muestran los pasos del proceso; las coordenadas sobre el primer plano factorial son las dos primeras columnas de la tabla 6.1 (abajo), ordenadas por las coordenadas sobre el primer eje. En la figura 6.1 se puede visualizar el procedimiento.

Tabla 6.1: Clasificación "a mano" de los 10 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

	Coordenadas		Paso 0			Paso 1			Paso 2			Paso 3		
Cafés	F_1	F_2	C_1^0	C_2^0	P_0	C_1^1	C_2^1	P_1	C_1^2	C_2^2	P_2	C_1^3	C_2^3	P_3
ExOs	-2.47	-0.04	2.28	2.95	1	2.00	2.98	1	1.27	3.34	1	1.26	3.71	1
O20C	-1.29	0.67	2.38	1.58	2	1.68	1.73	1	1.14	2.17	1	0.91	2.55	1
O20M	-1.04	0.65	2.33	1.36	2	1.59	1.48	2	1.14	1.92	1	0.91	2.30	1
ExCl	-0.89	-1.68	0.00	3.17	1	0.82	2.63	1	1.27	2.64	1	1.48	2.87	1
C20C	-0.44	-0.82	0.97	2.21	1	0.23	1.66	1	0.90	1.71	1	0.98	1.98	1
O40C	0.15	1.31	3.17	0.00	2	2.34	0.78	2	2.28	1.22	2	2.08	1.51	2
O40M	0.98	1.01	3.28	0.88	2	2.51	0.68	2	2.69	0.71	2	2.54	0.80	2
C20M	1.14	-1.02	2.13	2.53	1	1.80	1.76	2	2.47	1.36	2	2.50	1.27	2
C40C	1.18	0.15	2.76	1.55	2	2.11	0.86	2	2.53	0.37	2	2.44	0.11	2
C40M	2.68	-0.22	3.86	2.96	2	3.41	2.38	2	3.96	1.91	2	3.91	1.52	2

Pasos del *K-medias* de los 10 cafés

Paso 0: $K = 2$; centros iniciales: $C_1^0 = [-0.89, -1.68]$, $C_2^0 = [0.15, 1.31]$. Las distancias de los 10 cafés a los dos centros iniciales y la asignación de los cafés a las dos clases están en la tabla 6.1; entonces la partición P_0 es:

$$I_1^0 = \{ExOs, ExCl, C20C, C20M\}; \quad I_2^0 = \{O20C, O20M, O40C, O40M, C40C, C40M\}$$

Paso 1: Los centros de gravedad de la partición P_0 son: $C_1^1 = [-0.66, -0.89]$ y

$C_2^1 = [0.44, 0.59]$, las nuevas distancias a los centros se ven en la tabla 6.1 y la partición P_1 es:

$$P_1 : I_1^1 = \{ExOs, O20C, ExCl, C20C\}; \quad I_2^1 = \{O20M, O40C, O40M, C20M, C40C, C40M\}$$

Paso 2: Los centros de gravedad de la partición P_1 son: $C_1^2 = [-1.27, -0.47]$ y $C_2^2 = [0.85, 0.31]$, las distancias a estos puntos están en la tabla 6.1 y la asignación de los

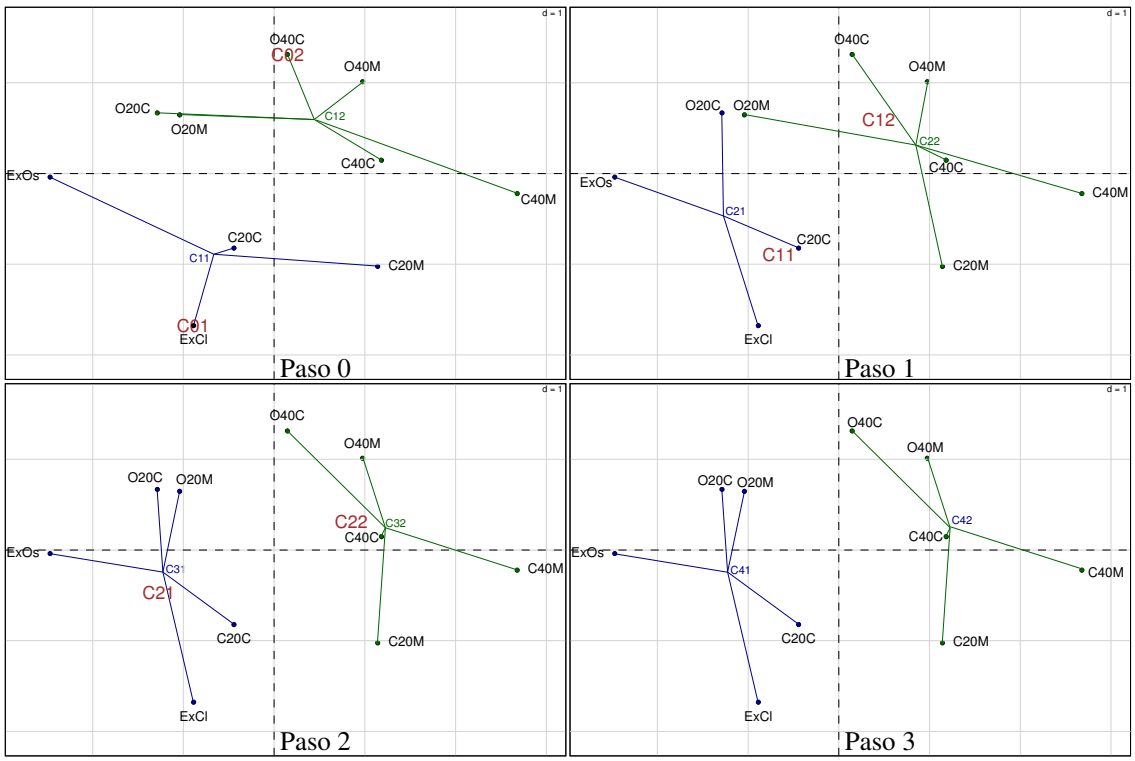


Figura 6.1: Ejemplo de clasificación con *K-means* del ejemplo Café a partir de las coordenadas factoriales sobre los ejes 1 y 2. Los rayos indican la pertenencia a cada clase al unir el centro de gravedad con los puntos. Con los centros iniciales C_1^0 y C_2^0 se construye la partición que se muestra 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.

cafés a las clases produce la partición:

$$P_2 : I_1^2 = \{ExOs, O20C, O20M, ExCl, C20C\}; \quad I_2^2 = \{O40C, O40M, C20M, C40C, C40M\}$$

Paso 3: Los centros de gravedad de la partición P_3 son: $C_1^3 = [-1.23, -0.24]$ y $C_2^3 = [1.23, 0.25]$, las distancias a estos puntos están en la tabla 6.1 y la asignación de los cafés a las clases produce la misma partición del paso 2 (última columna de la tabla 6.1).

Como la partición P_3 es igual a la partición P_2 , el algoritmo termina.

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 grupos y el algoritmo $K - means$ requiere que se le suministre el número de clases y los puntos iniciales.

2. En general, la inercia mínima que se obtiene depende de los puntos iniciales.

6.2. 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. 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 requieren de 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 que le da nombre a un método específico. Los más simples 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 **D**: d^0_{il}: grupo I^0_{il}.
   b) Calcular los índices de disimilitud entre I^0_{il} y los demás individuos.
   c) Eliminar las filas y columnas i y l e incluir la fila y columna I^0_{il}, para colocar las disimilitudes.
   d) Volver a 3*a* y repetir hasta tener un solo grupo de n individuos.

6.2.1. Índices de similitud, disimilitud y distancias entre individuos

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**x**E** que va hacia $\mathbb{R}^+ \cup \{0\}$

$$\begin{array}{rccc} s: & \mathbf{E}\text{x}\mathbf{E} & \longrightarrow & \mathbb{R}^+ \cup \{0\} \\ & (i,l) & \longmapsto & s(i,l) \end{array}$$

tal que:

$$\begin{array}{ll} s(i,l) = s(l,i) & \forall\, (i,l) \in \mathbf{E}\text{x}\mathbf{E} \\ s(i,i) = s(l,l) = s_{max} > s(i,l) & \forall\, i \in \mathbf{E} \end{array}$$

Las **medidas de disimilitud** ponen el énfasis sobre 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**x**E** que va hacia $\mathbb{R}^+ \cup \{0\}$.

$$\begin{array}{rccc} d: & \mathbf{E}\text{x}\mathbf{E} & \longrightarrow & \mathbb{R}^+ \cup \{0\} \\ & (i,l) & \longmapsto & d(i,l) \end{array}$$

tal que:

$$d(i,l) = d(l,i) \quad \forall\,(i,l) \in \mathbf{E}\mathrm{x}\mathbf{E}$$
$$d(i,i) = 0 \qquad \forall\, i \in \mathbf{E}$$

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:

$$\begin{aligned} &1.\quad d(i,l) = 0 \rightarrow i = l \\ &2.\quad d(i,l) \leq d(i,k) + d(l,k) \qquad \forall i,l,\mathrm{y}k \in \mathbf{E} \\ &3.\quad d(i,l) \leq \mathrm{máx}\{d(i,k), d(l,k)\} \quad \forall i,l,\mathrm{y}k \in \mathbf{E} \end{aligned} \tag{6.5}$$

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 la propiedad 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.

6.2.2. Índices de similitud para tablas binarias

Se calculan sobre una tabla de n individuos por p atributos de naturaleza binaria, donde la presencia de un 1 en una celda (i,j) indica que el individuo i tiene el atributo j y 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 y
- $u = b + c$ no coincidencias.

En la tabla 6.2 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.

6.2.3. Distancias para variables de intervalo

La distancias entre dos individuos i y l se calculan 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:

$$\begin{array}{rccc} d: & \mathbb{R}^p \,\mathrm{x}\, \mathbb{R}^p & \longrightarrow & \mathbb{R}^+ \cup \{0\} \\ & (i,l) & \longmapsto & d(i,l) \end{array}$$

Las distancias disponibles en la función `dist{stats}` se presentan en la tabla 6.3. La distancia euclidiana, y la distancia de Manhattan son casos particulares de la distancia de

Tabla 6.2: 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 and 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}`

Minkowski cuando $r = 2$ y $r = 1$ respectivamente. La distancia del máximo también es un caso particular de la distancia de Minkowski, ya que: $r \to \infty$:

$$\lim_{r\to\infty} \left(\sum_{j=1}^{p} |x_{ij} - x_{lj}|^r \right)^{1/r} = \max_j \{x_{ij} - x_{lj}\}$$

6.2.4. Criterios de agregación

Para completar un procedimiento de aglomeración jerárquica se requiere seleccionar una similitud, disimilitud o distancia entre grupos, que se denomina también criterio de agregación. Aquí solo se mencionan dos y en la sección 6.2.7 se introduce el criterio de agregación de Ward.

Tabla 6.3: Distancias para variables de intervalo

Nombre	Fórmula	`method`[1]
Euclideana	$d(i,l) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{lj})^2}$	`euclidian`
Manhattan o Cityblock	$\sum_{j=1}^{p} \lvert x_{ij} - x_{lj} \rvert$	`manhattan`
Minkowski	$\left(\sum_{j=1}^{p} \lvert x_{ij} - x_{lj} \rvert^r \right)^{1/r}$	`minkowski`
Del máximo o de Chebychev	$\max_j \{x_{ij} - x_{lj}\}$	`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}`.

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:

$$d(A,B) = \max\{d(i,l); i \in A; l \in B\}$$

El enlace completo tiende a producir grupos esféricos.

El procedimiento descrito en la página 172 es posible porque no se requiere retornar a la distancia entre individuos en cada paso de la clasificación, sino que es posible calcularla 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:

En el enlace simple: $d(AB, C) = \text{mín}\{d(A, C), d(B, C)\}$.

En el enlace completo: $d(AB, C) = \text{máx}\{d(A, C), d(B, C)\}$.

6.2.5. 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 6.2 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 que no hay diagonales. Por ejemplo la distancia ente a y b es 2, lo mismo que entre a y c y entre b y c.

El proceso de aglomeración procede así:

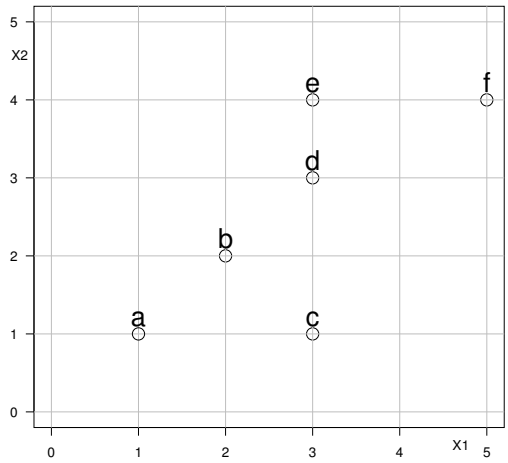
1. Se unen los puntos d y e a una distancia de 1. Se tiene ahora una partición en 5 clases.

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, 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.

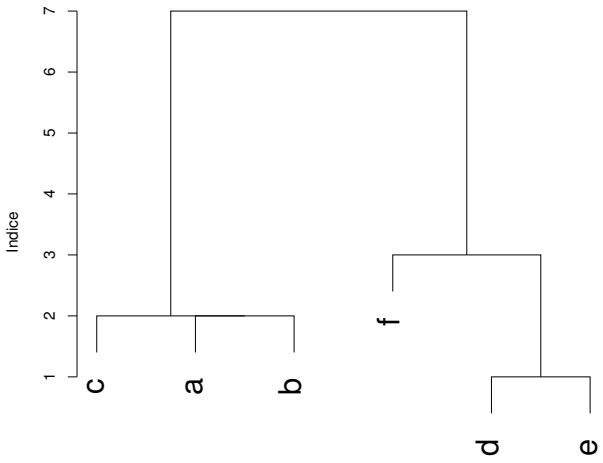
3. Se calculan las distancias de enlace completo entre el grupo ab y los demás puntos y grupos. Se une c al grupo ab a una distancia de 2. Ahora la partición es en 3 clases.

4. Se calculan las distancias de enlace completo entre el grupo abc con 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 clase.



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 asociada al árbol

	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 6.2: Ejemplo "de juego" de una clasificación jerárquica aglomerativa.

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 denominan: *nodos* a los puntos de unión; e *índices de nivel* a las distancias asociadas, que corresponden a las alturas del árbol.

6.2.6. Ultramétrica asociada a un árbol de clasificación

Una distancia Euclidiana d cumple la desigualdad triangular: sean a, b y c tres puntos en $\mathbb{R}^p$, entonces:

$$d(a,b) \leq d(a,c) + d(b,c) \tag{6.6}$$

Una distancia ultramétrica es más restrictiva y cumple la propiedad:

$$d(a,b) \leq max\{d(a,c), d(b,c)\} \tag{6.7}$$

Si se cumple (6.7) también se cumple (6.6); 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 6.2 se muestra un árbol y la ultramétrica asociada a él. 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 6.2 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
```

6.2.7. Método de Ward

Para lograr grupos que tengan inercia mínima intra-clases se debe utilizar una distancia Euclidiana y unir en cada paso del procedimiento los dos grupos que aumenten menos la inercia intra-clases, que corresponde al método de Ward (Ward 1963, Wishart 1969).

Distancia de Ward entre grupos e individuos

En la figura 6.3 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 la notaremos W, entonces hay que calcular $W(A,B)$, $W(A,C)$ y $W(B,C)$ y la mejor de ellas determinará los grupos a unir.

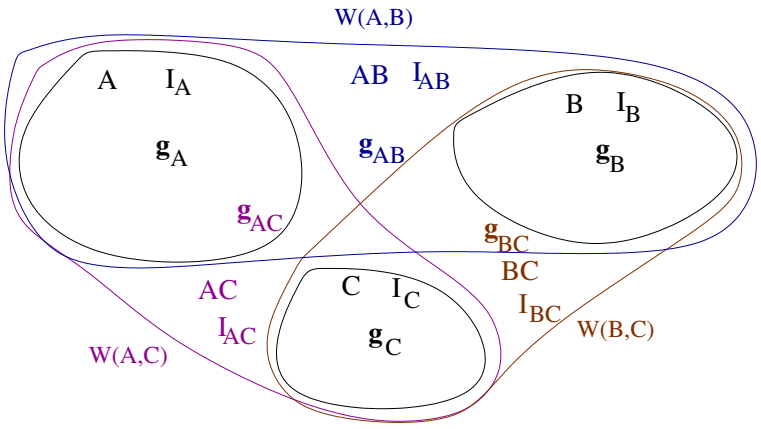


Figura 6.3: Esquema de tres grupos y sus posibles uniones en dos grupos, según el criterio de Ward.

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 respectivamente.

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, más 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:

$$Inercia - entre(A,B) = p_A d^2(\mathbf{g}_A, \mathbf{g}_{AB}) + p_B d^2(\mathbf{g}_B, \mathbf{g}_{AB})$$

que es la distancia de Ward entre los grupos A y B.

Reemplazando $\mathbf{g}_{AB} = \dfrac{1}{p_A + p_B}(p_A \mathbf{g}_A + p_B \mathbf{g}_A)$ en la fórmula anterior, se obtiene:

$$W(A,B) = \frac{p_A p_B}{p_A + p_B} d^2(\mathbf{g}_A, \mathbf{g}_B) \tag{6.8}$$

Este valor es el incremento de la inercia intra-grupos al unir los grupos A y B en uno solo. En particular para dos individuos i y l la distancia de Ward es:

$$W(i,l) = \frac{p_i p_l}{p_i + p_l} d^2(i,l) \tag{6.9}$$

Si los pesos son iguales a $1/n$ para los dos individuos, la anterior expresión se reduce a:

$$W(i,l) = \frac{1}{2n} d^2(i,l) \tag{6.10}$$

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 para calcular las distancias entre el grupo conformado en un paso y los demás solo se requiere tener 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):

$$d(AB, C) = \frac{(p_A + p_C)W(A, C) + (p_B + p_C)W(B, C) - p_C W(A, B)}{p_A + p_B + p_C} \tag{6.11}$$

Una forma de demostrar 6.11 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 para construir el árbol con el 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 (6.9).

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 (6.11).

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 intra-clases y viceversa

Antes de empezar las uniones toda la inercia corresponde a inercia entre-clases (cada individuo es una clase) y a medida que llevan a cabo las uniones, la inercia entre-clases va pasando a inercia intra-clases, de modo que al terminar, toda la inercia es intra-clases (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 contarlo, es decir se parte de una clase, con los n individuos, 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 formo 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, es decir que 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.

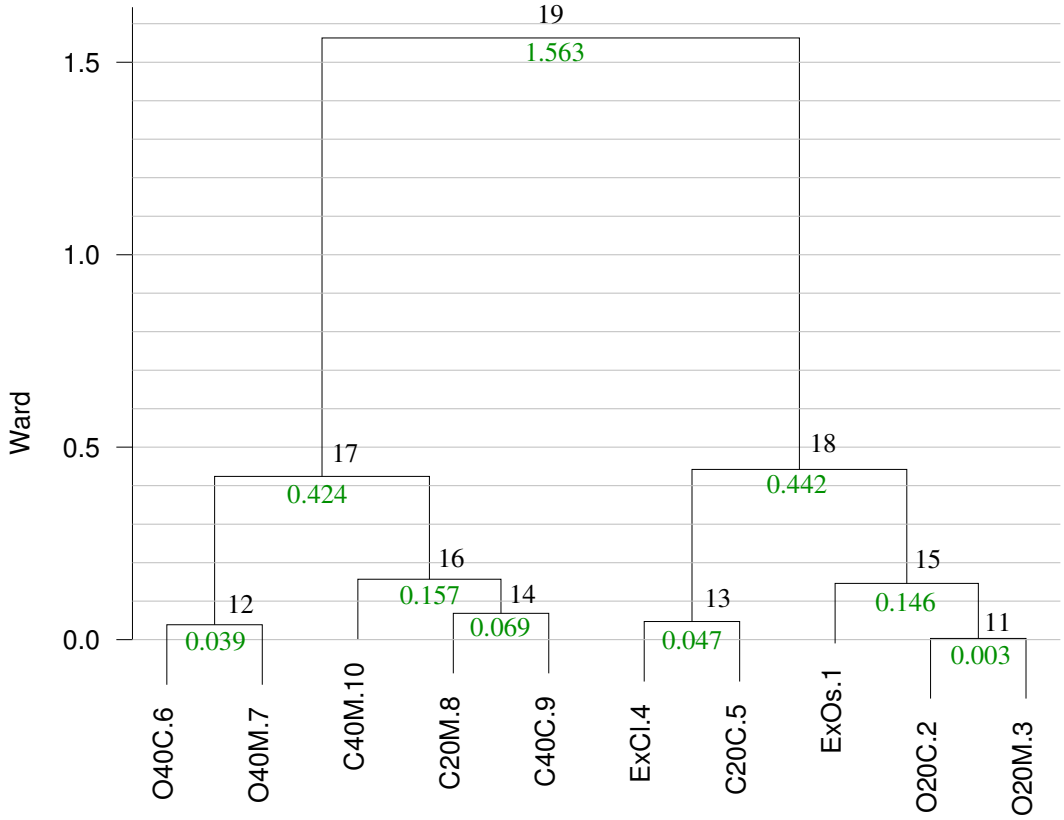
Algunos de los programas con implementación adecuada del método de Ward son: el de uso libre académico DtmVic (Lebart 2015) y el paquete FactoClass de R (Pardo & Del-Campo 2007).

Clasificación jerárquica con el método de Ward en el ejemplo Café

Se retoma el ejemplo de la sección 6.1.2 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 6.4, para mostrar los aspectos fundamentales del método.

1. Distancia de Ward entre cafés con la fórmula (6.10). Se muestran debajo del árbol en la figura 6.4.

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




```
> 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 6.4: 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.

dos cafés. En el árbol cada unión se llama nodo, los individuos a clasificar se llaman hojas o nodos terminales, son los nodos de 1 a 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 (6.11).

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.

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 6.4.

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, 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 intra-clases.

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 resulta ser la misma de la clasificación realizada con *K-means*, pero esto es un caso particular.

6.3. Combinación de métodos de clasificación

Desde el punto de vista de análisis de datos, los métodos a utilizar son el de Ward de aglomeración jerárquica y el *K-means*, porque buscan grupos que tengan inercia intra-grupos 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: requieren mayor recurso de cómputo y tiempo 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 que 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 por 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 intra-clases de la partición obtenida utilizando *K-means*, utilizando como puntos iniciales los centros de gravedad de la partición derivada de cortar el árbol.

6.4. Clasificación a partir de coordenadas factoriales

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 tablas de entrada a los métodos de clasificación son de la misma naturaleza, filas que representan individuos o categorías (grupos de individuos) por columnas que son las coordenadas factoriales del procedimiento previo. 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.

6.4.1. 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.

En un análisis de correspondencias las coordenadas factoriales se constituyen en nuevas variables que son continuas y con ellas se puede utilizar la combinación de métodos de la sección 6.3.

6.4.2. Función de filtro

Conectar un método factorial con la clasificación da la posibilidad de seleccionar el número de ejes a utilizar en la 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 que se registra 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.

En la figura 6.5 se muestra un esquema de la combinación, para el caso de la clasificación de individuos descritos por variables cualitativas.

6.5. 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 todas las variables presentes en la tabla de datos sobre las

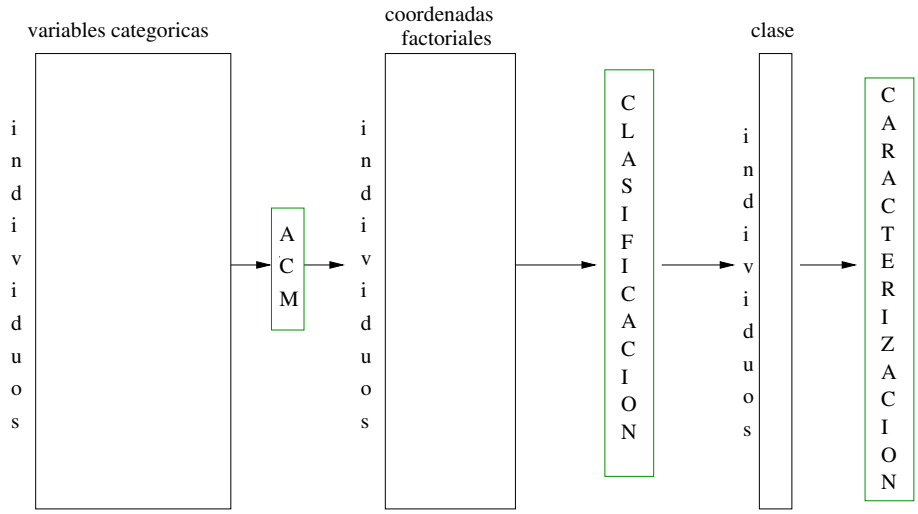


Figura 6.5: Esquema de una estrategia de clasificación con variables cualitativas.

que se desee.

El procedimiento de valores test presentado en la sección **??** es el que se utiliza para descubrir qué variables continuas o cualitativas son las que más caracterizan a las clases.

6.5.1. Descripción de las clases con 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 subsección 1.9.2.

6.5.2. Descripción de las clases 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 subsección 1.9.3.

6.6. Una estrategia de clasificación

La estrategia de clasificación que se ha propuesto, desde el punto de vista de la Estadistica 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 6.6 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.

6.7. Ejemplo de aplicación

A continuación se muestra la clasificación del ejemplo de admitidos a las carreras de la Facultad de Ciencias, utilizando `FactoClass`. Este análisis es complementario al realizado en el capítulo 5, 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 presenta con las decisiones tomadas:

```
fc<-FactoClass(Y,dudi.acm,admi[,1:7],scanFC=FALSE,nf=3,nfcl=6,k.clust=8).
```

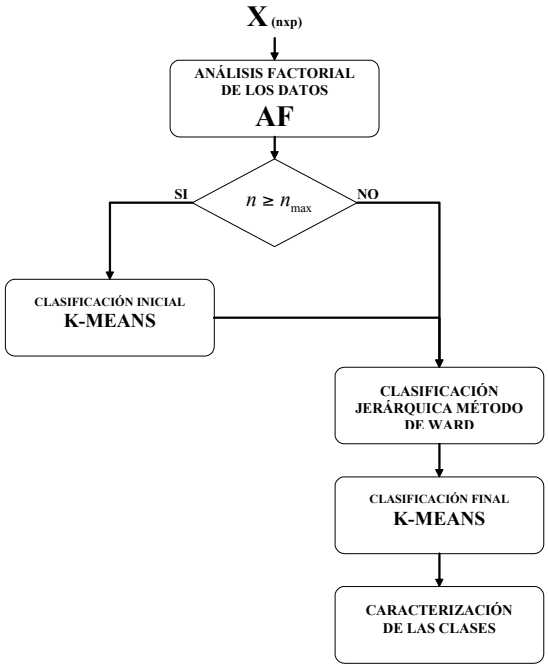


Figura 6.6: Esquema de la estrategia de clasificación.

A continuación se justifican las decisiones y se muestran los resultados.

1. Número de ejes para el ACM: en la página 132 se justifica la selección de ejes.

2. Número de ejes para la clasificación: en el histograma de valores propios, figura 5.1, luego de los tres primeros ejes se nota un salto en el eje 6, y se se decide utilizar 6 ejes para la clasificación.

3. Número de clases: en la figura 6.7 se observan cambios de inercia entre a intra notorios para justificar el árbol para 4, 7 y 8 clases. Se decidió 8 clases, la más fina de las tres.

   Código R. Para obtener la figura 6.7:

```
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))
```

4. Cambios en la consolidación: `xtable(fc$clus.summ[,1:4],digits=c(0,0,0,3,3))`

Las clases 4, 6 y 8 son inestables ya que tuvieron cambios en el proceso de consolidación, la 4 cedió individuos y las 6 y 8 los recibieron. La inercia intra-clases disminuyó de 0.490 a 0.464:

Clase	Tamaño antes	Tamaño des.	Inercia antes	Inercia des.
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
tot	445	445	0.490	0.464

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: la tabla 6.4 provee la información para describir las clases según las 4 variables cualitativas activas y la asociación con la carrera a la que ingresaron, que es ilustrativa en el análisis. La exploración de la relación de las clases con los resultados del examen de admisión según las áreas y el puntaje global, se incluye aquí, ya que solo 4 clases resultaron caracterizadas. Para cada clase se da el total de admitidos y su porcentaje.

   Cl1. 68 (15.3 %). De estrato alto, bogotanos (84 %), de 17 o menos años (65 %).

   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

   Cl2. 55 (12.4 %). De 17 años, estrato medio y bogotanos.

   Escogen Química en proporción menor al promedio (3.6 % vs. 14.2 %).

   Cl3. 54 (12.1 %). De 18 años, casi todos bogotanos.

   Aumenta la proporción de admitidos a Farmacia (25.9 % vs. 16.4 %).

   Cl4. 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 %).

   Cl5. 62 (13.9 %). De 17 años, con incremento en la proporción de estrato bajo (77.4 %), y de otro departamento (48.4 %).

El resultado en la componente textual es en promedio inferior al global:

	v.test	clase	Global
text	-2.1	11.1	11.4

Cl6. 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 %).

Cl7. 38 (8.5 %). Provienen de Cundinamarca, se incrementa la proporción de estrato bajo (57.9 %).

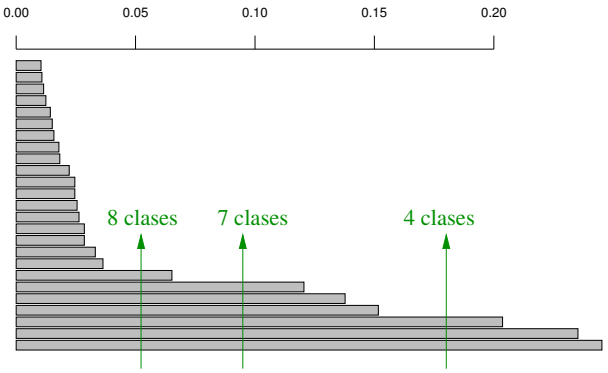
Cl8. 54 (12.1 %). Son, sobre todo, de otro departamento (79.6 %), estrato bajo (92.6 %) y de 16 años o menos (75.9 %).

En promedio tienen 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 6.8 se muestran las 8 clases sobre el primer plano factorial del ACM de los admitidos. Esto permite aprovechar el plano para la síntesis de la caracterización de las clases y compararlas. Por ejemplo, los centros de gravedad de las clases 1 y 4 se proyectan cerca en el plano, al observar la tabla 6.4 ambas clases tienen más porcentaje, que el promedio, de bogotanos y menos de cundinamarqueces. Las clases 2 y 3 se parecen por tener más bogotanos y por no tener admitidos de Cundinamarca.

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 6.7: Histograma índices de los últimos 25 nodos y últimos 12 nodos.

Tabla 6.4: Caracterización de las clases por las variables cualitativas activas y por la carrera a la que fueron admitidos

Clase 1

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
edad.a19M	-2.8	7.0	10.3	22.5	100
orig.Cund	-3.2	0.0	0.0	8.5	38
edad.a18	-4.1	0.0	0.0	12.6	56
estr.bajo	-8.6	0.0	0.0	40.2	179
estr.medio	-8.8	0.0	0.0	41.6	185

Clase 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
orig.Cund	-2.8	0.0	0.0	8.5	38
edad.a18	-3.6	0.0	0.0	12.6	56
estr.alto	-4.5	0.0	0.0	18.2	81
orig.Otro	-5.0	0.0	0.0	21.6	96
edad.a19M	-5.1	0.0	0.0	22.5	100
edad.a16m	-5.7	0.0	0.0	26.5	118
estr.bajo	-7.6	0.0	0.0	40.2	179

Clase 3

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
orig.Cund	-2.8	0.0	0.0	8.5	38
edad.a19M	-5.1	0.0	0.0	22.5	100
edad.a16m	-5.6	0.0	0.0	26.5	118
edad.a17	-7.2	0.0	0.0	38.4	171

Clase 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
carr.Fisi	-2.5	3.7	6.2	18.4	82
orig.Cund	-2.6	0.0	0.0	8.5	38
orig.Otro	-2.6	4.2	8.3	21.6	96
edad.a18	-3.3	0.0	0.0	12.6	56
estr.alto	-4.1	0.0	0.0	18.2	81
estr.bajo	-4.8	2.8	10.4	40.2	179
gene.M	-6.4	4.4	29.2	71.2	317
edad.a17	-6.8	0.0	0.0	38.4	171

Clase 5

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
orig.Cund	-3.0	0.0	0.0	8.5	38
orig.Bogo	-3.3	10.3	51.6	69.9	311
estr.medio	-3.4	7.6	22.6	41.6	185
edad.a18	-3.8	0.0	0.0	12.6	56
estr.alto	-4.8	0.0	0.0	18.2	81
edad.a19M	-5.5	0.0	0.0	22.5	100
edad.a16m	-6.1	0.0	0.0	26.5	118

Clase 6

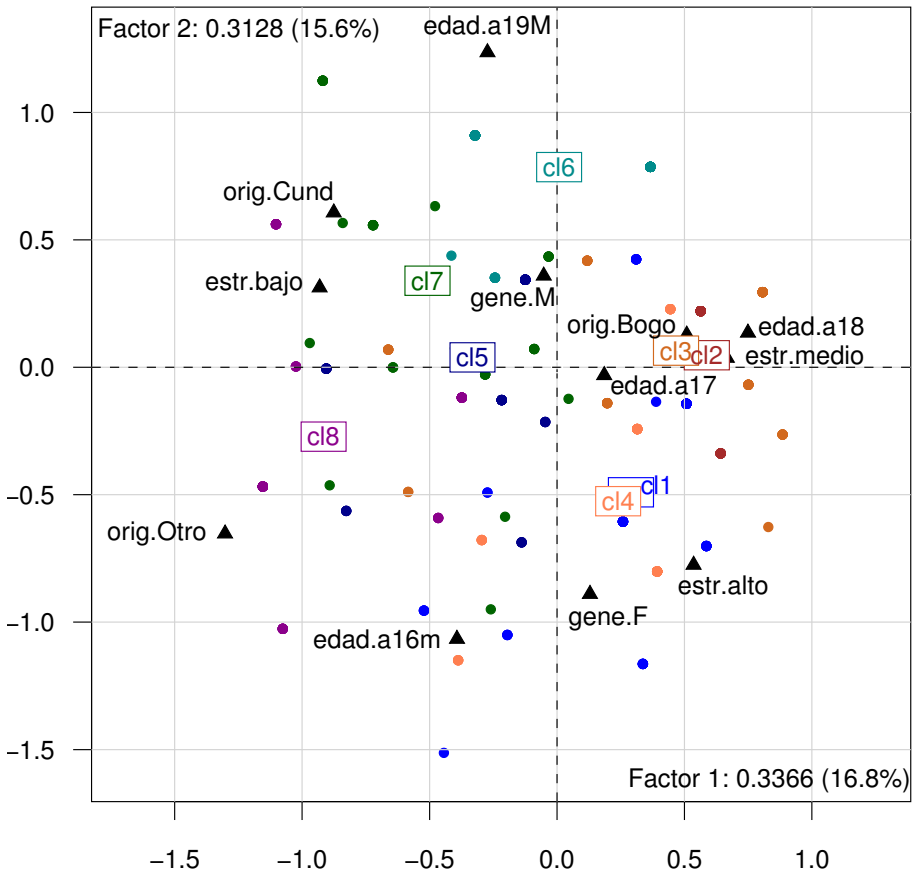
Categoria	Vtest	Cl/ca	ca/Cl	Global	n_{cat}
edad.a19M	15.4	66.0	100.0	22.5	100
orig.Bogo	6.4	20.9	98.5	69.9	311
gene.M	3.8	18.6	89.4	71.2	317
estr.bajo	2.2	19.0	51.5	40.2	179
carr.Mate	2.1	24.5	19.7	11.9	53
orig.Cund	-3.1	0.0	0.0	8.5	38
gene.F	-3.8	5.5	10.6	28.8	128
edad.a18	-4.0	0.0	0.0	12.6	56
orig.Otro	-5.0	1.0	1.5	21.6	96
estr.alto	-5.0	0.0	0.0	18.2	81
edad.a16m	-6.4	0.0	0.0	26.5	118
edad.a17	-8.1	0.0	0.0	38.4	171

Clase 7

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
orig.Otro	-4.0	0.0	0.0	21.6	96
orig.Bogo	-9.7	0.0	0.0	69.9	311

Clase 8

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
orig.Cund	-2.8	0.0	0.0	8.5	38
edad.a18	-3.5	0.0	0.0	12.6	56
estr.alto	-4.4	0.0	0.0	18.2	81
estr.medio	-5.9	2.2	7.4	41.6	185
edad.a17	-7.2	0.0	0.0	38.4	171
orig.Bogo	-8.0	3.5	20.4	69.9	311




```
mycolors<-colors()[c(26,32,52,57,73,74,81,84,88,100)]
plotFactoClass(fc,cstar=0,Trow=FALSE,cframe=1.05,col.row=mycolors,
    col.col="black")
```

Figura 6.8: Proyección de las clases sobre el primer plano factorial del ACM de admitidos.

6.8. Ejercicios

1. Muestre que en la distancia de Manhattan los puntos se unen mediante lineas paralelas a los ejes.
2. Verifique la propiedad 3) de una ultramétrica en la matriz de ultramétrica la figura 6.2.
3. Escriba la ultramétrica asociada al árbol del ejemplo Café (figura 6.4).
4. Calcule el coeficiente de correlación cofenética entre la ultramétrica y las distancias de Ward del ejemplo Café.
5. Demuestre que en una ultramétrica los triángulos son isósceles.
6. Demuestre que la propiedad 3) de una ultramétrica implica la propiedad 2): desigualdad triangular.
7. Derive la distancia de Ward entre dos grupos A y B (fórmula 6.8).
8. Demuestre la fórmula de recurrencia de Ward (6.11).
9. Muestre que en el método de Ward la inercia es igual a la suma de índices de nivel.
10. ¿Cuáles son las ventajas y desventajas del método $K-means$?
11. ¿Cuáles son las ventajas y desventajas de una clasificación jerárquica aglomerativa?

6.9. Talleres

6.9.1. Clasificación de razas de perros

El objetivo que se busca al realizar el ACM del ejemplo de Razas de Perros, presentado en la sección 6.9.1, 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 4 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.

6.9.2. Clasificación de las localidades de Bogotá

En la sección 4.6.1 se realiza el ACS de la TC $localidades \times estratos$, el objetivo del análisis se cumple mejor al complementar el ACS con la clasificación.

Realice la clasificación de las 19 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. El porcentaje de inercia explicado por la clasificación es: _____

6. La inercia intra-clases de la partición obtenida 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 5 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. ¿Qué significa si en el proceso de consolidación la inercia entre sobre la inercia total no cambia?.

16. ¿ Hay cambios en el proceso de consolidación? (Si/No): _____

17. El porcentaje de inercia explicado por la clasificación es: _____

18. Para cada clase escriba el estrato más asociado.

19. De las manzanas de estrato 4 el _____ % pertenecen a la clase 3.

20. Para cada estrato escriba la clase más asociada.

21. La clase 1 tiene el _____de sus manzanas en estrato 5.

22. Escriba los porcentajes de manzanas que hay en cada clase.

23. La clase más grande es la _____con el _____ % de las manzanas.

6.9.3. Clasificación de adjetivos según su perfil de colores asociados

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 4.6.2 y responda a las preguntas siguientes:

1. ¿Qué adjetivos se unen primero?
2. Relate las últimas 5 uniones en el proceso de clasificación, indicando los grupos que se unen en cada caso y el aumento de la inercia-intra clases.
3. ¿Cuánto 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 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 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 nueva variable categórica *clase* como ilustrativa.

Referencias

Agresti, A. (2002), *Analysis of Ordinal Categorical Data*, Wiley series in probability and statistics, 2 edn, Wiley Online Library.

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/*

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.

Canavos, G. (1988), *Probabilidad y Estadística. Aplicaciones y métodos*, McGraw-Hill.

Chessel, D., Dufour, A.-B. & Thioulouse, J. (2004), 'The ade4 package-I- One-table methods', *R News* **4**, 5–10.

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. (2014), *xtable: Export tables to LaTeX or HTML*. R package version 1.7-3. **URL:** *http://CRAN.R-project.org/package=xtable*

Dalgaard, P. (2008), *Introductory statistics with R*, Springer Science & Business Media.

Dalgaard, P. (2015), *ISwR: Introductory Statistics with R*. R package version 2.0-7. **URL:** *http://CRAN.R-project.org/package=ISwR*

DAPD (1997), *Población estratificación y aspectos socioeconómicos de Santa Fe de Bogotá*, Departamento Administrativo de Planeación Distrital.

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. Clas. Dewey 519.53/ E649c.

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.

Hamann, U. (1961), 'Merkmalsbestand und verwandtschafsbeziehungen der farinosae', *Willdenowia* **2**, 639–768.

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, *in* '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.

Lebart, L. (2015), '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 (2016), *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: *http://www.jstatsoft.org*

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 Subpartitions on Rowsand Columns', *Revista Colombiana de Estadística* **36**(1), 115–144.

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.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Rogers, D. & Tanimoto, T. (1960), 'A computer program for classifying plants', *Systematics Biology.* **132**, 1115–1118.

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.

Sutanthavibul, S., Smith, B. V. & King, P. (2016), *XFIG Version 3.2.6 August 2016 Users Manual*.
URL: *http://mcj.sourceforge.net/*

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 (2017), *LaTeX – A document preparation system*.
URL: *https://www.latex-project.org/*

Venables, W. N., Smith, D. M. & R Development Core Team (2015), *An introduction to R*, version 3.2.3 (2015-12-10) edn, R Core Team.

Ward, J. H. (1963), 'Hierarchical grouping to optimize an objective function', *Journal of the American Statistical Association* **58**(301), 236–244.

Wishart, D. (1969), 'An algorithm for hierarchical classifications', *Biometrics* pp. 165–170.

Índice temático

Índice temático