FACULTAD DE MATEMÁTICAS

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Trabajo Fin de Grado

Técnicas Multivariantes para el Análisis de Datos Ómicos

Sandra Muñoz Armayones

Dirigido por:
Dña. Inmaculada Barranco Chamorro

2016

Resumen

El constante aumento en la generación de datos ómicos y el desarrollo de tecnologías que permiten su análisis han hecho que el interés por estudiar simultáneamente datos procedentes de distintas técnicas ómicas sea cada vez mayor, con el propósito de conocer las relaciones subyacentes entre ellos.

Para ello se requieren herramientas matemáticas que puedan adaptarse al análisis de una gran cantidad de datos, reduciendo a su vez la complejidad de los mismos y facilitando así su interpretación. Esto nos lleva a considerar métodos de proyección, que serán descritos en el trabajo además de aplicarse en diferentes conjuntos de datos reales.

En la introducción al trabajo, se han presentado las principales técnicas ómicas así como la tecnología de microarrays, además de explicar qué se entiende por integración de datos. Planteamos la necesidad de utilizar técnicas de Análisis Multivariante para analizar el tipo de datos que nos ocupa.

En el primer capítulo del trabajo se explica con detalle el Análisis de Componentes Principales (ACP), una técnica capaz de crear un pequeño conjunto de variables que resuman la información de las originales y que permitan posteriores análisis más profundos de los datos.
Comenzamos introduciendo la notación, además de definir componentes principales y componentes principales muestrales. Se presentan ejemplos para explicar el círculo de correlación y la utilidad del ACP para solucionar problemas de colinealidad en la regresión lineal múltiple. Se explican tres criterios para escoger las CP significativas.
Por último, se ha realizado una aplicación a un conjunto de datos de expresión génica, en la que utilizamos ACP como una técnica exploratoria de datos.

En el segundo capítulo vemos la Descomposición en Valores Singulares (DVS). Este será un paso intermedio en varias de las técnicas estadísticas expuestas a lo largo del trabajo y nos permite descomponer matrices rectangulares como producto de otras. Se detallan propiedades de la descomposición, la utilidad de la técnica para aproximar matrices y su representación gráfica como un biplot.

En el tercer capítulo nos centramos en el Análisis de Correspondencias (AC). Esta técnica es aplicable tan sólo a variables categóricas y define unos índices. A estos índices se les denomina coordenadas principales y estándar y se obtienen a partir de la descomposición del estadístico

chi-cuadrado, χ^2. Se realiza una aplicación a datos de expresión génica utilizando el paquete made4 de R.

El cuarto capítulo trata sobre Análisis de Coinercia (ACoi). Se presenta un coeficiente que permite medir la correlación entre conjuntos de datos donde se trabaja con las mismas muestras. Con el propósito de realizar un análisis integrado, consideramos dos conjuntos de datos de expresión génica. Se realiza un Análisis de Correspondencias como paso previo al Análisis de Coinercia, el cuál nos permite cuantificar y visualizar la relación existente entre ambos conjuntos.

La última técnica que trataremos se describe en el quinto capítulo y se conoce como Análisis de Correlación Canónica, que explora las relaciones de dependencia entre conjuntos de variables.Se ha resuelto detalladamente el problema que supone encontrar los dos primeros vectores de correlación canónica haciendo uso de la Descomposición en Valores Singulares. A partir de esto se han definido las variables de correlación canónica y se han planteado contrastes de significación para elegir las más relevantes.
Se ilustra con un estudio de datos de expresión génica en grupos de ratones sometidos a distintas dietas.

Para finalizar, se han detallado todos los paquetes de R empleados, describiendo cada paquete así como todas las funciones y argumentos que hemos usado.

Abstract

Constant growth in omics data generation and development of technology that allows this data analysis has resulted in an increasing interest in studying different kinds of omics techniques, in order to know the mutual interactions between these data sets.

Mathematical tools to analyse large data sets are required. They must be able to reduce complexity and make the interpretation of these data easier. So we consider projection methods which will be described along this work. As illustrations, applications to different real data sets are included.

In the Introduction, the main omics techniques are presented. Microarray technology and data integration are explained. The need for using multivariate analysis techniques is also contemplated.

In Chapter 1, we focus on Principal Components Analysis (PCA). This technique is able to create a little set of variables which summarize information and permit deeper analysis of data. We introduce the appropriate notation, and define principal and sample principal components. Examples to explain the correlation circle are given. We show how useful this method can be to deal with highly correlated variables in linear regression. Three different options to choose important components are described. Finally, we apply PCA to explore microarray gene expression data.

In Chapter 2, we study Singular Value Decomposition (SVD). This is a common tool in multivariate analysis used to decompose a rectangular matrix as product of other matrices. We highlight properties and biplot representation of this technique.

In Chapter 3, Correspondence Analysis (CA) is presented as a technique applicable to categorical variables. Indexes, called principal and standard coordinates, are obtained from the decomposition of a χ^2 statistic. An application is carried out by using *made4* package of R.

Chapter 4 is devoted to Coinertia Analysis (CIA). This technique allows us to obtain a coefficient which explains the existing correlation between two data sets containing the same samples. An application in which we perform an integrated analysis is given. Quantification and visualization of the relationships between the two data sets, under consideration, is possible thanks to Correspondence Analysis, which is a previous step in order to apply CIA.

In Chapter 5, Canonical Correlation Analysis (CCA) is proposed as a technique to explore

dependence between variable sets. A method to find the two first canonical correlation vectors is studied in detail. Canonical correlation variables are presented, and significance tests are proposed to choose the most relevant ones. CCA is applied to a nutritional study in mice.

Finally, an Appendix is given with the R packages and functions used in this work.

Índice general

Introducción y Contexto

De la biología a la ómica

Se conoce como Ómica al estudio de las diferentes componentes que participan y/o regulan procesos biológicos complejos.

Las *técnicas ómicas* se basan en el análisis de una gran cantidad de datos. La principales técnicas ómicas son:

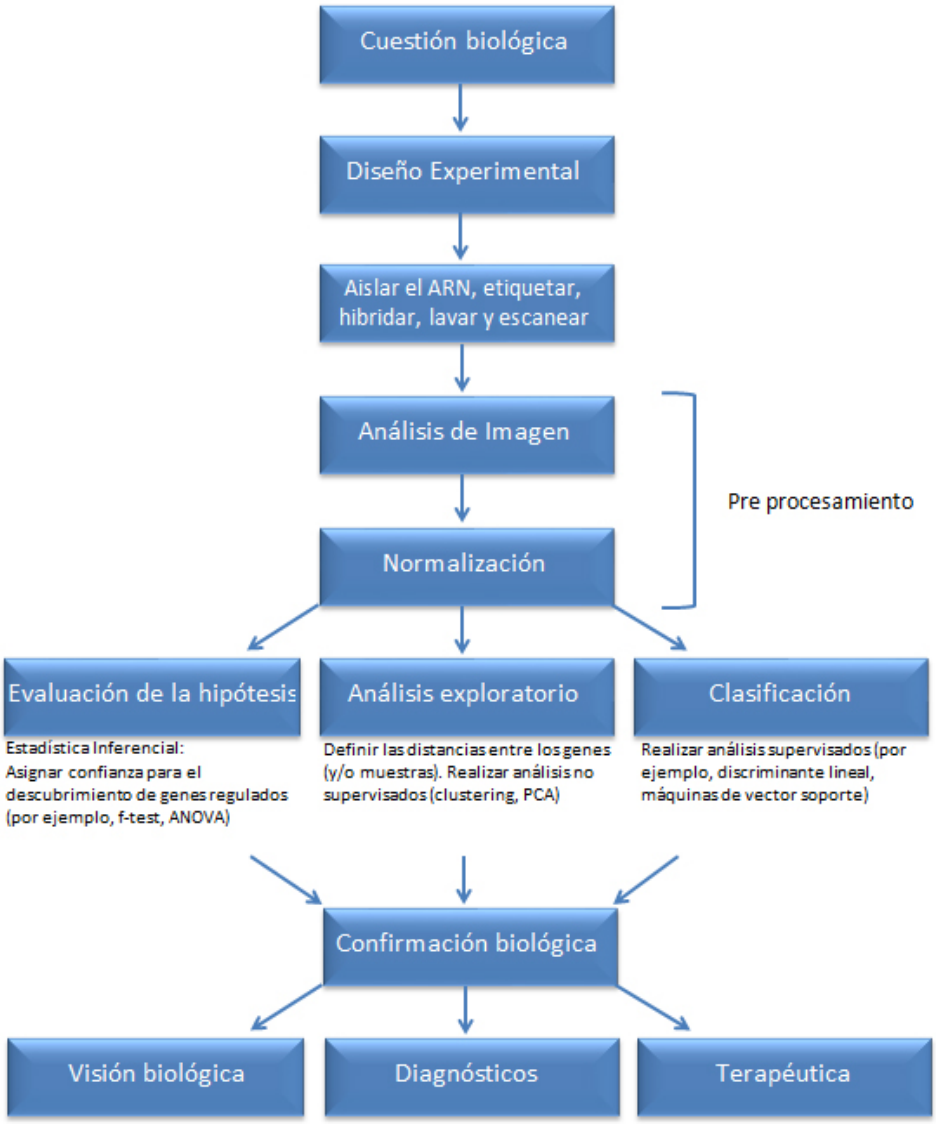
- Genómica: Se encarga de secuenciar, montar y analizar la estructura y función de los genomas; siendo un genoma el conjunto completo de ADN dentro de una sola célula en un organismo.
- Transcriptómica: Mide los niveles de ARN de una población celular.
- Proteómica: Estudia la estructura y función de las proteínas.
- Inómica: Estudia el perfilado de elementos, las interacciones entre ellos y su regulación bioquímica.
- Metabolómica: Mide niveles de metabolitos, siendo éstas las moléculas que intervienen en las reacciones de las células de un organismo.
- Fenómica. Evalúa rasgos. Relaciona genética, epigenética (cambios en el fenotipo o expresión génica) y factores ambientales.

Nos centraremos en la tecnología de microarrays (o chips de ADN) para el estudio de la transcriptómica.

Esta tecnología se basa en el Dogma Central de la Biología Molecular, una hipótesis formulada en 1953 para referirse a los procesos llevados a cabo en la transmisión y expresión de la herencia genética. Estos procesos se conocen como replicación (ADN se duplica), transcripción (ADN se sintetiza) y por último; síntesis de proteínas, que se lleva a cabo en los ribosomas a partir del ADN transcrito.

Describiremos el procedimiento para obtener datos mediante este tipo de tecnología con detalle.

Podemos describir el proceso analítico de los microarrays ayudándonos del siguiente esquema:



Como se observa, tenemos que partir de una cuestión biológica y orientar el experimento para encontrar respuesta. Una vez que se ha diseñado el experimento, podemos proceder a medir los niveles de ARN.

El ácido ribonucleico que contiene la información del ADN original y la traspasa al ribosoma, donde tienen lugar la síntesis de las proteínas, recibe el nombre de ARNm o ARN mensajero.

Se representan en un chip las secuencias biológicas, de manera que se pueda cuantificar el nivel de transcripción en una matriz numérica. En cada una de las celdas del chip se almacenan copias de un segmento de ARNm, de manera que todas las celdas tienen el mismo número de copias

pero en cada una de ellas aparece una secuencia distinta. El número de ARNm de cada segmento será variable.

Este ARNm extraído de las muestras se identifica con unos marcadores fluorescentes. Luego se hibrida, de manera que cadenas de ARNm complementarias se combinan.
Tras esto se lava el microarray para eliminar aquellas muestras que no se hayan hibridado y se escanea. En el escaneo se ilumina con un láser que revela el color del marcador fluorescente en función de la cantidad de ARNm que haya en cada celda del chip. El resultado final se denomina nivel de expresión.

Posteriormente, se lleva a cabo un análisis de imagen; donde se convierte la cantidad de secuencias hibridadas (medidas por fluorescencia) en una intensidad de luz (número), para luego normalizar los datos. Esto forma parte de lo que se conoce como preprocesamiento.

En el proceso, existen varios elementos que pueden interferir en la medida de la expresión génica. Estos son:

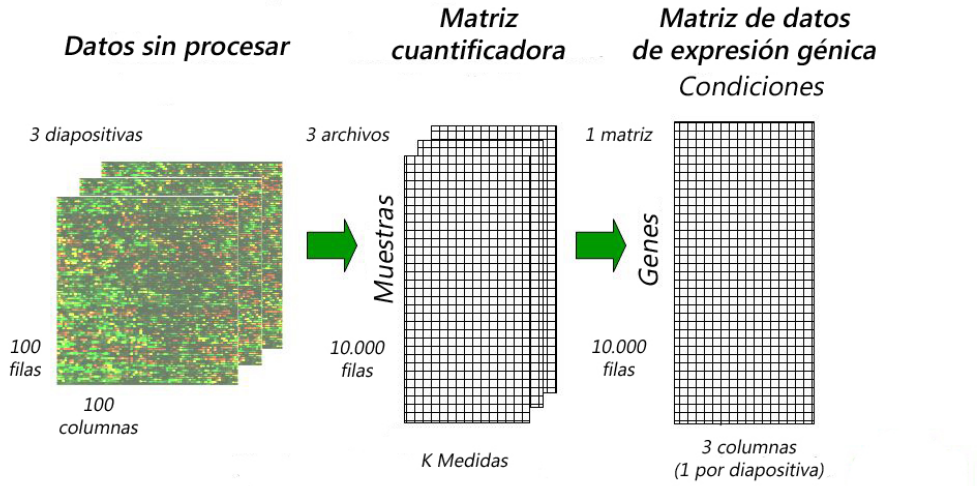
- La fluorescencia, debida al rendimiento del escáner o a distinta eficiencia de las etiquetas.
- La densidad de impresión.
- El experimento biológico, por la pureza de las muestras o la manipulación de las mismas.

Por ello se hace el preprocesamiento, donde se busca disminuir tendencias erróneas, así como la varianza de los datos. Dentro del preprocesamiento se llevan a cabo varios pasos: cuantificación de la imagen, exploración de los datos, corrección de fondo; normalización; sumarización y, por último, determinación de la calidad.

En esta etapa del análisis se pasa de microarrays a una matriz numérica analizable, conocida como matriz de expresión; donde las columnas son las condiciones y las filas los genes. Esta matriz cuantifica la abundancia de la variable i en la muestra j. Los valores de expresión para cada muestra son independientes, mientras que para cada variable no.

De cada muestra se obtiene información que la describe. Este tipo de información se conoce como metadatos o variables fenotípicas.

Una vez que hemos llegado a esta etapa, sólo queda hacer el análisis estadístico correspondiente en función del objetivo que persigamos y luego interpretarlo.



El análisis de alto nivel que se realiza antes de la verificación biológica depende del estudio que se desee llevar a cabo. Existen tres tipos: comparación de clases, descubrimiento de clase o predicción de clase.
Es por ello por lo que se usan diferentes técnicas o bien de estadística descriptiva, o bien de estadística inferencial.

Los test de hipótesis son útiles para los estudios de comparación de clases, mientras que la parte de la estadística descriptiva que define distancias entre genes y muestras se usa para el descubrimiento de clases. El modelo lineal discriminante y las máquinas de vector soporte son efectivas para la predicción de clase.

En los problemas de “Análisis de grupos de genes” se mide la asociación entre grupos de variables y una variable fenotípica de interés.

Los resultados de la mayoría de los análisis ómicos consisten en una o más listas de identificadores. Estos identificadores son nombres o números únicos que necesitan estar:

- asociados con una o más bases de datos (anotación)
- interpretados en un contexto biológico (anotación funcional)

Se conoce como “GO” (Gene Ontology) a la base de datos de anotaciones creada para proveer un vocabulario apropiado para describir genes y atributos del producto génico, en otras palabras,

del material bioquímico que resulta de la expresión de un gen.

Esta base de datos está organizada por función molecular, proceso biológico y componente celular.

Análisis interactivo e integración de datos

Existe una verdadera necesidad de analizar e integrar datos para cualquier investigación biológica.

La primera década del siglo XXI fue significativa, pues comenzaron a generarse una gran cantidad de datos biológicos debido a la creciente disponibilidad de secuencias de genomas además del desarrollo de tecnologías de alto rendimiento.

Todas estas tecnologías se caracterizan porque miden un sólo tipo de información en muchas variables, simultáneamente. En muchos casos, éstas pertenecen a un tipo específico de tecnología ómica; como por ejemplo la transcriptómica o la proteómica.

En un principio, cada una de las aproximaciones anteriores se utilizaba por separado ya que las tecnologías eran muy costosas al estar poco avanzadas. No obstante, a medida que fueron mejorando y siendo más asequibles iba creciendo el interés por considerar simultáneamente datos de distinto tipo, a fin de entender mejor los procesos biológicos que intervienen en un mismo problema.

Mientras más datos para trabajar y más posibilidades de obtenerlos había, mayor era el interés de orientar esta parte de la biología hacia una modelización y análisis de organismos como un conjunto.

Para el mismo estudio se emplearían datos de distinta clase (expresión, proteínas, metabolitos,...), y por tanto, se necesitaban métodos y herramientas para analizarlos de manera conjunta.
Es por ello por lo que se han desarrollado muchos métodos en los últimos años que permiten alcanzar nuestro objetivo. Existen, o bien métodos basados en machine-learning (aprendizaje automático), redes bayesianas, máquinas de vector soporte y métodos basados en gráficos; o bien métodos de estadística multivariante, que lleva usándose mucho tiempo para combinar y visualizar datos multivariantes.

El objetivo de la integración de datos como la entendemos es combinar diferentes recursos de datos que contribuyan a una mejor comprensión del fenómeno general de estudio. Pero el concepto “integración de datos” no tiene siempre el mismo significado.

Por una parte, el término puede usarse para describir herramientas y métodos que combinan y analizan múltiples recursos de datos; un significado puramente informático. Esto conduce a que

el usuario en ocasiones no sepa en concreto de qué base de datos procede la información con que trabaja, pues esta está dispersa entre distintas bases.

Por otra parte, se pueden combinar estudios relacionados a fin de obtener una conclusión de más peso. Esto se describiría como una aproximación mixta que se centra en combinar estudios uniendo los datos (meta-análisis) a la vez que los reprocesa y reanaliza (integración de los datos).

Sin embargo, pueden considerarse distintos tipos de datos (medidos o no en los mismos individuos) y tratar de combinarlos de manera que ayuden a interpretar los procesos biológicos que interfieren. Éste es el tipo de estudio de datos en el que nos centraremos.

Hay muchos tipos de análisis de datos ómicos en función del problema biológico (mapa genético, clasificación, extracción de características,...) o estadístico, el tipo de datos (similar o heterogéneo) y la etapa de integración (comparación de datos, normalización, filtrado de calidad,...).

El método que usemos para nuestro análisis tiene que:

- Reducir dimensión eficientemente
- Representar simultáneamente muestras y variables de cada conjunto de datos
- Reducir el problema que surge al haber muchas variables y pocos individuos
- Integrar datos suplementarios en un espacio común con los datos originales.

Técnicas Clásicas para el Análisis de Datos Ómicos

Los métodos clásicos para el análisis de datos multivariantes son Regresión Múltiple, que determina si existe relación de dependencia entre dos ó más variables; Análisis Discriminante, cuyo objetivo es describir diferencias significativas entre grupos sobre los que se observan p variables; y ANOVA o Análisis de la Varianza, que compara las medias de dos o más variables.

Estos métodos clásicos se aplican cuando el número de individuos es mucho mayor que el de variables.
Se asume:

- Variables independientes
- Más observaciones que variables
- Normalidad multivariante
- Existencia de una variable dependiente
- Pocos datos perdidos

En los problemas que nos plantearemos no se verifican la mayoría de asunciones necesarias para aplicar técnicas clásicas. En conjuntos de datos ómicos, por lo general; se miden muchas variables simultáneamente y hay muy pocas muestras analizadas, por lo que se necesitan otro tipo de técnicas para su análisis.

La mejor forma de solucionar el problema es reduciendo la dimensión de los datos mediante métodos de proyección, donde se evalúan todas las variables de forma conjunta proporcionando así modelos más estables, mostrando relaciones subyacentes (denominadas *variables latentes*) y además asegurando poca pérdida de información.

Estos métodos son:

- Análisis de Componentes Principales (ACP)
- Descomposición en Valores Singulares (DVS)
- Análisis de Correspondencias (AC)
- Análisis de Coinercia (ACoi)
- Análisis de Correlación Canónica (ACC)

Capítulo 1

Análisis de Componentes Principales

El objetivo del Análisis de Componentes Principales (ACP) es reducir la dimensionalidad de un conjunto de datos multivariante mediante la sustitución de las variables originales por transformaciones lineales de las mismas, obteniendo componentes incorreladas y con poca pérdida de información.

De este modo, el método permite hacer una representación gráfica que podamos interpretar y también adecuar los datos para realizar cualquier otro tipo de análisis.

Esta técnica, iniciada por Pearson en 1901 y desarrollada por Hotelling en 1933, no requiere supuesto de normalidad; sólo que exista el vector de medias y la matriz de varianzas y covarianzas del vector aleatorio objeto de estudio.

Mostraremos que el ACP es, además, útil para visualizar clusters en los datos e identificar outliers.

Expresaremos la variación en el conjunto de variables originales $\underline{X} = (X_1, \ldots, X_p)'$ en términos de un conjunto de variables incorreladas $\underline{Y} = (Y_1, \ldots, Y_p)'$, en orden decreciente de "importancia".

Las primeras componentes explicarán una gran proporción de la variabilidad total, lo que nos será útil para resumir los datos en una dimensión menor a la inicial.

1.1. Notación

Sea un vector aleatorio p-dimensional $\underline{X} = (X_1, \ldots, X_p)'$

Se supone que existe $\underline{\mu}$, vector de esperanzas de $\underline{X}$:

$$E(\underline{X}) = (\mu_1, \ldots, \mu_p)' = \underline{\mu}$$

Sea $\mathbf{\Sigma}$ la matriz de varianzas y covarianzas de las variables originales:

$$\mathbf{\Sigma} = Cov(\underline{X}) = \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \dots & \sigma_{pp} \end{pmatrix}$$

con

$$\sigma_{ij} = Cov(X_i, X_j) = E((X_i - \mu_i)(X_j - \mu_j)')$$

$$\sigma_{ii} = Cov(X_i, X_i) = Var(X_i) = E(X_i^2) - \mu_i^2$$

Entonces, $\underline{X} \sim (\underline{\mu}, \mathbf{\Sigma})$

Consideramos una m.a.s. de $\underline{X}$ de tamaño n. Construimos la matriz de datos muestrales siguiente:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix} = (\underline{x}_1, \dots, \underline{x}_n)' = (\underline{x}_{(1)}, \dots, \underline{x}_{(p)})$$

donde

- $\underline{x}_i = (x_{i1}, \dots, x_{ip})'$ es un vector $p \times 1$ con el valor de cada variable para el individuo i-ésimo.
- $\underline{x}_{(j)} = (x_{1j}, \dots, x_{nj})'$ es un vector $n \times 1$ que contiene la muestra de tamaño n de la variable X_j.

Sea $\widehat{\mathbf{\Sigma}}$ la matriz de varianzas y covarianzas muestrales de $\underline{X}$

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})'$$

donde $\overline{\underline{x}}$ es el vector de medias muestrales.

Para $j, s \in \{1, \dots, p\}$ se denota la varianza muestral de $\underline{x}_{(j)}$ como $\hat{\sigma}^2_{\underline{x}_{(j)}}$

$$\hat{\sigma}^2_{\underline{x}_{(j)}} = \hat{\sigma}_{jj} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \overline{x_j})^2$$

La covarianza muestral entre $\underline{x}_{(j)}$ y $\underline{x}_{(s)}$ se denota por $\hat{\sigma}_{\underline{x}_{(j)}, \underline{x}_{(s)}}$

$$\hat{\sigma}_{\underline{x}_{(j)}, \underline{x}_{(s)}} = \hat{\sigma}_{js} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \overline{x_j})(x_{is} - \overline{x_s})$$

Por último, denotamos por $\lambda_1, \lambda_2, ..., \lambda_p$ los autovalores de la matriz $\mathbf{\Sigma}$ y por $\underline{e}_1, \underline{e}_2, \dots, \underline{e}_p$ sus autovectores.
Del mismo modo, denotamos sus análogos muestrales como $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_p$ y $\underline{\hat{e}}_1, \underline{\hat{e}}_2, \dots, \underline{\hat{e}}_p$; respectivamente.

1.2. Componentes Principales

La primera componente principal se define como la combinación lineal de las variables originales con varianza máxima de entre todas las combinaciones posibles.

La segunda componente principal se define como la combinación que maximiza la varianza acumulada y que; además, es incorrelada con la primera componente principal. Se sigue este mismo razonamiento para definir el resto de componentes.

A continuación introducimos notación adecuada para tratar con transformaciones lineales de las variables. Así mismo, se recogen propiedades de dichas transformaciones lineales que se utilizarán en las secciones siguientes.

1.2.1. Transformaciones lineales de las variables originales

Sea $\underline{t}_j = (t_{j1}, \ldots, t_{jp})' \in \mathbb{R}^p$, con $j \in \{1, \ldots, p\}$.

Consideremos la siguiente transformación:

$$\mathbf{X}\underline{t}_j = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} t_{j1} \\ \vdots \\ t_{jp} \end{pmatrix} = \begin{pmatrix} t_{j1}x_{11} + \ldots + t_{jp}x_{1p} \\ \vdots \\ t_{j1}x_{n1} + \ldots + t_{jp}x_{np} \end{pmatrix} = \begin{pmatrix} y_{1j} \\ \vdots \\ y_{nj} \end{pmatrix}$$

El resultado obtenido por esta transformación se denota como $\underline{y}_{(j)}$ y es la muestra de tamaño n de la nueva variable $Y_j = \underline{t}_j'\underline{X}$.

Estas transformaciones de las variables originales presentan las propiedades que se detallan en la siguiente proposición.

Proposición 1.2.1 *Para cada una de las transformaciones $Y_j \quad (j = 1, \ldots, p)$, definimos las siguientes características poblacionales:*

- *Media poblacional:*

$$E(Y_j) = E(\underline{t}_j'\underline{X}) = \underline{t}_j'E(\underline{X}) = \underline{t}_j'\underline{\mu} = t_{j1}\mu_1 + \ldots + t_{jp}\mu_p$$

- *Varianza poblacional:*

$$Var(Y_j) = Var(\underline{t}_j'\underline{X}) = \underline{t}_j'\mathbf{\Sigma}\underline{t}_j$$

- *Covarianza entre variables:*

$$Cov(Y_i, Y_j) = Cov(\underline{t}_i'\underline{X}, \underline{t}_j'\underline{X}) = \underline{t}_i'Cov(\underline{X}, \underline{X})\underline{t}_j = \underline{t}_i'Var(\underline{X})\underline{t}_j = \underline{t}_i'\mathbf{\Sigma}\underline{t}_j$$

Además, definimos los siguientes estadísticos muestrales análogos:

- *Media muestral:* $\overline{y}_{(j)} = \underline{t}_j' \overline{x}$
- *Varianza muestral:* $\hat{\sigma}^2_{\underline{y}_{(j)}} = \underline{t}_j' \widehat{\mathbf{\Sigma}} \underline{t}_j$
- *Dados* $\underline{t}_j, \underline{d}_j \in \mathbb{R}^p$*, la covarianza muestral de las transformaciones lineales* $\underline{t}_j' \underline{X}$ *y* $\underline{d}_j' \underline{X}$ *es:* $\hat{\sigma}_{\underline{t}_j' \underline{X}, \underline{d}_j' \underline{X}} = \underline{t}_j' \widehat{\mathbf{\Sigma}} \underline{d}_j$

Proposición 1.2.2 *La matriz de varianzas y covarianzas* $\mathbf{\Sigma}$ *presenta las siguientes propiedades:*

- $\mathbf{\Sigma}$ *es simétrica y semidefinida positiva. Por tanto, los autovalores son reales y positivos cumpliendo:*
$$\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p \geq 0$$
- *Sea* $\underline{e}_j$ *el autovector unitario asociado al autovalor* λ_j*, con* $\| \underline{e}_j \|^2 = 1$*. Los autovectores de esta matriz son ortogonales, es decir*
$$\underline{e}_i' \underline{e}_j = 0, \qquad si \quad i \neq j$$
Disponemos estos autovectores por columnas en una matriz a la que denotamos por $\mathbf{E}$*.*
$$\mathbf{E}_{p \times p} = (\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_p)$$
- *La matriz* $\mathbf{E}$ *es ortogonal:*
$$\mathbf{E}'\mathbf{E} = \mathbf{E}\mathbf{E}' = \mathbf{I}_p$$

Todas estas propiedades también se cumplen para su equivalente muestral $\widehat{\mathbf{\Sigma}}$

El siguiente teorema nos será muy útil para definir las combinaciones lineales de las variables originales que permiten obtener las componentes principales. Dichas combinaciones lineales se obtendrán a partir de la descomposición espectral de la matriz de varianzas y covarianzas $\mathbf{\Sigma}$, como veremos más adelante.

Teorema 1.2.1 *Teorema de la descomposición espectral*

En las condiciones anteriores, se tiene que:

1. $\mathbf{E}'\mathbf{\Sigma}\mathbf{E} = \mathbf{\Lambda} = diag\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$
2. $\mathbf{\Sigma} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}' = \sum_{j=1}^{p} \lambda_j \underline{e}_j \underline{e}_j'$
3. $Traza(\mathbf{\Sigma}) = \sum_{j=1}^{p} \sigma_{jj} = Traza(\mathbf{\Lambda}) = \sum_{j=1}^{p} \lambda_j$

Demostración:

1. *Sea $\mathbf{\Sigma}$ la matriz de varianzas y covarianzas de $\underline{X}$, $\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$ sus autovalores y $\mathbf{E} = (\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_p)$ la matriz que contiene sus autovectores.*

   Sea $\underline{e}_j$ el j-ésimo autovector de $\mathbf{\Sigma}$. Por la definición de autovector, sabemos que se satisface la siguiente igualdad:

   $$\mathbf{\Sigma}\underline{e}_j = \lambda_j \underline{e}_j$$

   Sea $\mathbf{\Lambda} = diag\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$, podemos expresar la ecuación anterior en forma matricial:

   $$\mathbf{\Sigma E} = \mathbf{E\Lambda}$$

   Por la Proposición 1.2.2 sabemos que $\mathbf{E}$ es ortogonal, luego multiplicando por $\mathbf{E}'$ a izquierda en ambos lados de la igualdad, se obtiene:

   $$\mathbf{E'\Sigma E} = \mathbf{\Lambda}$$

2. *Partiendo de $\mathbf{\Sigma E} = \mathbf{E\Lambda}$ y multiplicando a derecha por $\mathbf{E}'$ en ambos lados de la igualdad, se obtiene $\mathbf{\Sigma} = \mathbf{E\Lambda E}'$.*

   $$\mathbf{\Sigma} = \mathbf{E\Lambda E}' = (\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_p) \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} \begin{pmatrix} \underline{e}_1' \\ \vdots \\ \underline{e}_p' \end{pmatrix} = \sum_{j=1}^{p} \lambda_j \underline{e}_j \underline{e}_j'$$

3. *Por el primer apartado sabemos que $\mathbf{\Lambda} = \mathbf{E'\Sigma E}$ y por lo tanto $Tr(\mathbf{\Lambda}) = Tr(\mathbf{E'\Sigma E})$*

   Para probar esta parte del teorema hay que tener en cuenta que la matriz $\mathbf{E}$ es ortogonal y que $Tr(\mathbf{AB}) = Tr(\mathbf{BA})$ para dos matrices $\mathbf{A}$ y $\mathbf{B}$ para las que se puedan definir ambos productos.
   Por consiguiente, se satisface:

   $$Tr(\mathbf{E'\Sigma E}) = Tr(\mathbf{E'E\Sigma}) = Tr(\mathbf{\Sigma})$$

   Queda así demostrado que $Tr(\mathbf{\Lambda}) = Tr(\mathbf{\Sigma})$

□

Para definir las transformaciones de las variables que se llevan a cabo en la determinación de cada componente principal es necesario calcular el valor de los coeficientes $\underline{t}_j = (t_{j1}, \ldots, t_{jp})' \in \mathbb{R}^p$, con $j \in \{1, \ldots, p\}$.

Comenzaremos calculando los coeficientes para la primera componente principal.

Si la expresión de esta componente es $Y_1 = t_{11}X_1 + t_{12}X_2 + ... + t_{1p}X_p$, y su varianza poblacional es $\underline{t}_1' \mathbf{\Sigma} \underline{t}_1$; esta varianza podría aumentar sin límite elevando el valor de los coeficientes $\underline{t}_1' = (t_{11}, t_{12}, ..., t_{1p})$.

Por consiguiente, para encontrar la varianza máxima es necesario introducir la siguiente restricción en nuestro problema: $\underline{t}_1' \underline{t}_1 = 1$

Si queremos maximizar una función dependiente de un gran número de variables y sujeta a una, o más de una restricción, usamos el método de *Los multiplicadores de Lagrange*.

La técnica de los multiplicadores de Lagrange prueba que, si λ_1 es el mayor autovalor de $\mathbf{\Sigma}$; entonces requiriendo que $\underline{t}_1' \underline{t}_1 = 1$, puede verse que $\underline{t}_1$ es el autovector de la matriz $\mathbf{\Sigma}$ correspondiente a λ_1 y que este autovalor es la varianza de la primera CP.

Del mismo modo, con los estadísticos muestrales correspondientes se obtiene la primera componente principal muestral, que definiremos en la próxima sección.

Veamos más detenidamente el razonamiento que se sigue para obtener las transformaciones de las variables.

Como hemos planteado anteriormente, para la primera CP tenemos que resolver el problema:

$$max(\underline{t}_1' \mathbf{\Sigma} \underline{t}_1)$$

$$s.a. : \underline{t}_1' \underline{t}_1 = 1$$

Consideramos la siguiente función y buscamos el máximo, derivando su expresión e igualándola a cero:

$$L(\underline{t}_1) = \underline{t}_1' \mathbf{\Sigma} \underline{t}_1 - \lambda(\underline{t}_1' \underline{t}_1 - 1) \tag{1.2}$$

$$\frac{\partial(L(\underline{t}_1))}{\partial(\underline{t}_1)} = 2\mathbf{\Sigma}\underline{t}_1 - 2\lambda\mathbf{I}\underline{t}_1 = 0 \tag{1.3}$$

$$(\mathbf{\Sigma} - \lambda\mathbf{I})\underline{t}_1 = 0 \tag{1.4}$$

Por el Teorema de Rouché-Frobenius, la matriz $(\mathbf{\Sigma} - \lambda\mathbf{I})$ ha de ser singular para que el sistema tenga solución distinta de cero.

Luego: $|\mathbf{\Sigma} - \lambda \mathbf{I}| = 0 \Leftrightarrow \lambda$ es autovalor de $\mathbf{\Sigma}$.

A partir de la igualdad dada en 1.4 se obtiene:

$$\mathbf{\Sigma}\underline{t}_1 = \lambda \mathbf{I}\underline{t}_1$$

Sustituimos esta igualdad en la expresión de la varianza:

$$Var(Y_1) = \underline{t}_1'\mathbf{\Sigma}\underline{t}_1 = \underline{t}_1'\lambda\mathbf{I}\underline{t}_1 = \lambda\underline{t}_1'\underline{t}_1 = \lambda$$

Luego, para maximizar la varianza de la combinación lineal que define la primera componente principal hay que tomar el mayor autovalor (λ_1) y el correspondiente autovector ($\underline{e}_1$) de $\mathbf{\Sigma}$.

Para la segunda componente; además, se requiere que sea incorrelada con la primera, es decir:

$$Cov(Y_2, Y_1) = Cov(\underline{t}_2'\underline{X}, \underline{t}_1'\underline{X}) = \underline{t}_2'\mathbf{\Sigma}\underline{t}_1 = 0$$

Por la condición anterior se verifica $\mathbf{\Sigma}\underline{t}_1 = \lambda\mathbf{I}\underline{t}_1$, de forma que sustituyendo esta igualdad en la expresión de la covarianza obtenemos lo siguiente:

$$\underline{t}_2'\mathbf{\Sigma}\underline{t}_1 = \underline{t}_2'\lambda\mathbf{I}\underline{t}_1 = 0 \Leftrightarrow \underline{t}_2'\underline{t}_1 = 0$$

Como sabemos, los autovectores de $\mathbf{\Sigma}$ son ortonormales, así que cumplen ambas condiciones.

Se sigue el mismo razonamiento con todas las componentes, resolviéndose así el problema de maximizar la varianza de la combinación lineal correspondiente que además es incorrelada con las combinaciones anteriores.

Obtenemos, por tanto, la transformación de las variables para cada componente.

Definición 1.2.1 *Se define la j-ésima componente principal como:*

$$Y_j = \underline{e}_j'\underline{X}, \qquad j = 1, \ldots, p$$

con $\underline{e}_j$ autovectores unitarios de $\mathbf{\Sigma}$.

Teorema 1.2.2 *La variabilidad explicada por la j-ésima CP es*

$$\frac{\lambda_j}{\sum_{s=1}^{p} \lambda_s}$$

A su vez, la proporción de variabilidad total no explicada por las k primeras CP es

$$\frac{\sum_{j=k+1}^{p} \lambda_j}{\sum_{s=1}^{p} \lambda_s}$$

1.3. Cálculo y propiedades de las Componentes Principales Muestrales

Una vez resuelto el problema que supone calcular los coeficientes para las transformaciones de las variables originales $\underline{X}$, podemos definir las CP muestrales.

Definición 1.3.1 *Se define la j-ésima componente principal muestral como:*

$$Y_j = \underline{\hat{e}}_j' \underline{X}, \qquad j = 1, \ldots, p.$$

con $\underline{\hat{e}}_j$ autovectores unitarios de $\widehat{\mathbf{\Sigma}}$.

A continuación introducimos notación adecuada para trabajar con las CP muestrales.

- Puntuaciones correspondientes a la j-ésima componente principal

  Estos son los valores muestrales de la j-ésima componente principal. Se obtienen multiplicando el autovector $\underline{\hat{e}}_j'$ por la muestra del individuo correspondiente.

$$\underline{y}_{(j)} = \begin{pmatrix} y_{1j} \\ \vdots \\ y_{nj} \end{pmatrix} = \begin{pmatrix} \underline{\hat{e}}_j' \underline{x}_1 \\ \vdots \\ \underline{\hat{e}}_j' \underline{x}_n \end{pmatrix} = \mathbf{X} \underline{\hat{e}}_j, \qquad j = 1, \ldots, p$$

- Puntuaciones de las p componentes principales se recogen en la siguiente matriz

$$\mathbf{Y} = \begin{pmatrix} \underline{x}_1' \underline{\hat{e}}_1 & \cdots & \underline{x}_1' \underline{\hat{e}}_p \\ \vdots & \ddots & \vdots \\ \underline{x}_n' \underline{\hat{e}}_1 & \cdots & \underline{x}_n' \underline{\hat{e}}_p \end{pmatrix} = \mathbf{X}\widehat{\mathbf{E}} = (\underline{\mathbf{y}}_{(\mathbf{1})}, \ldots, \underline{\mathbf{y}}_{(\mathbf{p})})$$

  Estos son los valores muestrales de las p componentes principales, recogidos en una matriz de dimensión $n \times p$.

  Cada columna es una muestra de tamaño n de la correspondiente CP. Cada fila muestra los valores de todas las CP para cada uno de los individuos de la muestra.

- Transformación de los datos muestrales de las variables originales en los datos muestrales de las CP

  Podemos escribir la matriz anterior por filas, siendo cada fila $\underline{y}_i'$ los valores muestrales de las p componentes principales para el individuo i-ésimo.

$$\mathbf{Y} = \begin{pmatrix} \underline{x}_1' \underline{\hat{e}}_1 & \cdots & \underline{x}_1' \underline{\hat{e}}_p \\ \vdots & \ddots & \vdots \\ \underline{x}_n' \underline{\hat{e}}_1 & \cdots & \underline{x}_n' \underline{\hat{e}}_p \end{pmatrix} = \mathbf{X}\widehat{\mathbf{E}} = \begin{pmatrix} \underline{y}_1' \\ \vdots \\ \underline{y}_n' \end{pmatrix}$$

Los valores muestrales de las CP para el individuo i-ésimo son

$$\underline{y}_i = \widehat{\mathbf{E}}'\underline{x}_i, \qquad i = 1, \ldots, n$$

A continuación introducimos varios resultados que caracterizan a las componentes principales muestrales.

Proposición 1.3.1 *Se mantienen las distancias entre los datos originales y los datos transformados:*

$$d^2(\underline{y}_h, \underline{y}_i) = d^2(\underline{x}_h, \underline{x}_i)$$

Demostración:
La prueba se basa en la ortogonalidad de la matriz $\widehat{\mathbf{E}}$.

$$d^2(\underline{y}_h, \underline{y}_i) = (\underline{y}_i - \underline{y}_h)'(\underline{y}_i - \underline{y}_h) = (\underline{x}_i - \underline{x}_h)'\widehat{\mathbf{E}}\widehat{\mathbf{E}}'(\underline{x}_i - \underline{x}_h) = (\underline{x}_i - \underline{x}_h)'(\underline{x}_i - \underline{x}_h) = d^2(\underline{x}_h, \underline{x}_i)$$

□

Proposición 1.3.2 *Las componentes principales muestrales tienen las siguientes propiedades:*

- *Matriz de varianzas-covarianzas:*

  $$\widehat{\mathbf{\Sigma}}_y = \widehat{\mathbf{E}}'\widehat{\mathbf{\Sigma}}\widehat{\mathbf{E}} = \widehat{\mathbf{\Lambda}} = diag\{\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p\}$$

  Demostración:

  $$\widehat{\mathbf{\Sigma}}_y = Cov(\widehat{\mathbf{E}}'\underline{X}) = \widehat{\mathbf{E}}'Cov(\underline{X})\widehat{\mathbf{E}} = \widehat{\mathbf{E}}'\widehat{\mathbf{\Sigma}}\widehat{\mathbf{E}} = \widehat{\mathbf{\Lambda}}$$

  Esta última igualdad se obtiene por el Teorema de la Descomposición Espectral (Teorema 1.2.1) .

- *La varianza muestral de la j-ésima CP es igual al autovalor correspondiente,* $\hat{\lambda}_j$

  $$\hat{\sigma}^2_{\underline{y}_{(j)}} = \hat{\lambda}_j, \qquad \forall j = 1, ..., p$$

- *Covarianza muestral entre dos CPs:* $\hat{\sigma}_{\underline{y}_{(j)}\underline{y}_{(s)}} = 0, \qquad \forall j \neq s$

- $\sum_{j=1}^{p} \hat{\sigma}^2_{\underline{x}_{(j)}} = tr(\widehat{\mathbf{\Sigma}}) = \sum_{j=1}^{p} \hat{\lambda}_j = \sum_{j=1}^{p} \hat{\sigma}^2_{\underline{y}_{(j)}}$

1.4. Componentes principales para datos estandarizados

En ocasiones, las variables que caracterizan los datos con los que trabajamos están medidas en escalas muy distintas.

La técnica de Análisis de Componentes Principales es sensible a cambios escalares, es decir, si intentamos igualar escalas multiplicando una variable por un escalar obtendremos diferentes autovalores y autovectores. Esto se debe a que la descomposición de autovalores se hace en la matriz de covarianzas y no en la de correlación; por tanto, los resultados que obtendremos usando ambas matrices serán distintos.

Es por ello por lo que a veces es preferible trabajar con los datos estandarizados, porque de esta forma las magnitudes de las variables son similares. Se resuelve así el problema que supone que las variables con mayor varianza influyan más en la determinación de la primera componente principal.

Definición 1.4.1 *La matriz de datos estandarizados es la siguiente:*

$$\mathbf{X}_{(s)} = \begin{pmatrix} \frac{x_{11}-\overline{x}_1}{\hat{\sigma}_1} & \cdots & \frac{x_{1p}-\overline{x}_p}{\hat{\sigma}_p} \\ \vdots & & \vdots \\ \frac{x_{n1}-\overline{x}_1}{\hat{\sigma}_1} & \cdots & \frac{x_{np}-\overline{x}_p}{\hat{\sigma}_p} \end{pmatrix}$$

donde $\overline{x}_j$ es la media muestral y $\hat{\sigma}_j$ la desviación típica de $\underline{x}_{(j)}$, siendo $\underline{x}_{(j)}$ un vector de dimensión $n \times 1$ conteniendo la muestra de tamaño n para la variable X_j $(j = 1, \ldots, p)$.

Si calculamos las componentes principales a partir de la matriz de datos estandarizados, esto equivale a calcular las componentes a partir de la matriz de correlaciones en lugar de la matriz de covarianzas.

En la práctica, sólo debemos utilizar la matriz $\widehat{\boldsymbol{\Sigma}}$ para obtener las componentes principales cuando las variables originales tengan la misma escala de medida.

Si no se cumple esta condición emplearemos la matriz de datos estandarizados, de forma que la suma total de los autovalores será p (pues se asumen varianzas iguales a 1 para todas las variables) y la proporción de variabilidad recogida por la j-ésima componente principal muestral es $\frac{\hat{\lambda}_j}{p}$.

1.4.1. Círculo de correlación

El círculo de correlación es una herramienta que permite conocer la contribución de cada variable en la determinación de una CP.
Esta contribución, si es que existe, puede ser tanto positiva como negativa. Depende de la proximidad de la variable en su representación a la periferia del círculo y al eje definido por la CP.

A su vez, permite conocer relaciones entre las variables.

Gracias a este gráfico podremos visualizar las variables más correladas con las componentes principales que recojan más información.

Para comenzar, definimos la covarianza entre dos conjuntos de variables como:

$$Cov(\underline{X},\underline{Y}) = E(\underline{XY}') - E(\underline{X})E(\underline{Y}') = E(\underline{XX}'\mathbf{E}) - \underline{\mu\mu}'\mathbf{E} = Var(\underline{X})\mathbf{E} =$$

$$= \mathbf{\Sigma E} = \mathbf{E\Lambda E}'\mathbf{E} = \mathbf{E\Lambda}$$

Por tanto, la correlación lineal entre la variable original X_j y la CP Y_k se expresa como:

$$\rho_{X_j,Y_k} = \frac{Cov(X_j,Y_k)}{\sqrt{Var(X_j)Var(Y_k)}} = \frac{\lambda_k e_{kj}}{\sqrt{Var(X_j)Var(Y_k)}} = \frac{\lambda_k e_{kj}}{\sigma_j\sqrt{\lambda_k}} = \frac{\sqrt{\lambda_k}e_{kj}}{\sigma_j}$$

cumpliéndose:

$$\sum_{k=1}^{p}\rho^2_{X_j,Y_k} = \frac{\sum_{k=1}^{p}\lambda_k {e_{kj}}^2}{\sigma_j^2} = \frac{\sigma_j^2}{\sigma_j^2} = 1$$

Es obvio que si estudiamos la correlación de la variable X_j con las dos primeras componentes principales, $\rho^2_{X_j,Y_1} + \rho^2_{X_j,Y_2} <= 1$, por tanto los puntos estarán siempre dentro de un círculo de radio 1.

A continuación, mostraremos con un ejemplo la interpretación del círculo de correlación y su utilidad.

Aplicación

El fichero de datos contiene 6 medidas distintas de billetes de 1000 francos suizos, de manera que 100 de ellos son verdaderos (los primeros) y los restantes 100 falsos.
Los datos se encuentran en el paquete *mclust*.

Llevaremos a cabo un Análisis de Componentes Principales ayudándonos del paquete *FactoMineR*.

```
library(mclust)
library(FactoMineR)
data(banknote)
str(banknote)
```

```
'data.frame': 200 obs. of  7 variables:
 $ Status  : Factor w/ 2 levels "counterfeit",..: 2 2 2 2 ...
```

```
$ Length  : num  215 215 215 215 215 ...
$ Left    : num  131 130 130 130 130 ...
$ Right   : num  131 130 130 130 130 ...
$ Bottom  : num  9 8.1 8.7 7.5 10.4 9 7.9 7.2 8.2 9.2 ...
$ Top     : num  9.7 9.5 9.6 10.4 7.7 10.1 9.6 10.7 ...
$ Diagonal: num  141 142 142 142 142 ...
```

Con la función `str` vemos la estructura de los datos con los que trabajamos. Tenemos un data.frame con 200 observaciones para las que se miden las siguientes variables:

1. $Status$: factor que representa el estado del billete. "genuine" indica que es verdadero, mientras que "counterfeit" indica que no.
2. $Length$: longitud del billete en mm. (X_1)
3. $Left$: anchura del borde izquierdo en mm. (X_2)
4. $Right$: anchura del borde derecho en mm. (X_3)
5. $Bottom$: anchura del margen inferior en mm. (X_4)
6. Top: anchura del margen superior en mm. (X_5)
7. $Diagonal$: longitud de la diagonal en mm. (X_6)

Por tanto, sólo existe una variable categórica ($Status$).

```
summary(banknote)
```

```
        Status        Length          Left          Right
counterfeit:100   Min.   :213.8   Min.   :129.0   Min.   :129.0
genuine    :100   1st Qu.:214.6   1st Qu.:129.9   1st Qu.:129.7
                  Median :214.9   Median :130.2   Median :130.0
                  Mean   :214.9   Mean   :130.1   Mean   :130.0
                  3rd Qu.:215.1   3rd Qu.:130.4   3rd Qu.:130.2
                  Max.   :216.3   Max.   :131.0   Max.   :131.1
    Bottom           Top          Diagonal
Min.   : 7.200   Min.   : 7.70   Min.   :137.8
1st Qu.: 8.200   1st Qu.:10.10   1st Qu.:139.5
Median : 9.100   Median :10.60   Median :140.4
Mean   : 9.418   Mean   :10.65   Mean   :140.5
3rd Qu.:10.600   3rd Qu.:11.20   3rd Qu.:141.5
Max.   :12.700   Max.   :12.30   Max.   :142.4
```

Una vez visto un resumen de cada variable con `summary`, llevamos a cabo un Análisis de Componentes Principales con la función `PCA` de *FactoMineR*.

Además de especificar el data.frame de los datos con los que trabajamos, con el argumento "graph = FALSE" indicamos que no queremos ninguna representación gráfica, y con "quali.sup=1" indicamos que la primera variable es categórica y debe tratarse de forma diferente al resto.

```
res.pca=PCA(banknote, graph = FALSE,quali.sup=1 )
```

Con el comando `summary` podemos ver un resumen del ACP llevado a cabo. Muestra los autovalores y autovectores de la matriz de varianzas y covarianzas muestrales $\widehat{\Sigma}$, correlaciones con componentes principales, así como otra información de utilidad; diferenciando individuos, variables y variables categóricas.

```
summary(res.pca)
```

Veamos este resumen con detalle.
En primer lugar aparecen los autovalores, porcentaje de varianza explicada y porcentaje de varianza acumulado para las 6 CP.

```
 Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4
Variance               2.946   1.278   0.869   0.450
% of var.             49.093  21.301  14.484   7.496
Cumulative % of var.  49.093  70.394  84.878  92.374
                       Dim.5   Dim.6
Variance               0.269   0.189
% of var.              4.478   3.148
Cumulative % of var.  96.852 100.000
```

Podemos observar que las 2 primeras componentes explican un 70.394 % de la variabilidad total.

En la primera columna del apartado "Individuals" se indica la distancia en la representación de cada individuo a la categoría a la que pertenece. Tan sólo se indica información de los 5 primeros individuos.

```
Individuals
             Dist    Dim.1    ctr   cos2    Dim.2    ctr
1       |  3.970 |  1.747  0.518  0.194 |  1.651  1.066
2       |  2.533 | -2.274  0.878  0.806 | -0.539  0.114
3       |  2.457 | -2.277  0.880  0.859 | -0.108  0.005
4       |  2.414 | -2.284  0.885  0.895 | -0.088  0.003
5       |  4.234 | -2.632  1.176  0.386 |  0.039  0.001
```

```
             cos2    Dim.3    ctr   cos2
1           0.173 |  1.424  1.166  0.129 |
2           0.045 |  0.533  0.163  0.044 |
3           0.002 |  0.717  0.296  0.085 |
4           0.001 | -0.606  0.211  0.063 |
5           0.000 |  3.196  5.878  0.570 |
```

En las columnas "Dim.1", "Dim.2" y "Dim.3" se muestran los valores $\underline{y}_{(j)}$ para $j = 1, 2, 3$; en otras palabras, los valores de las 3 primeras CP para los primeros individuos de la muestra.

En la columna "ctr" se representa la contribución de cada individuo en la determinación de cada CP.
La calidad de representación de un elemento (ya sea individuo o variable) en los ejes de un gráfico se mide por el coseno al cuadrado del ángulo existente entre el vector del elemento y su proyección en el eje correspondiente. Estos valores, que a su vez se corresponden con la correlación al cuadrado entre estos elementos y cada CP, aparecen en la columna "cos2".

Además de la contribución y la calidad de representación, en el apartado "Variables" aparecen los valores de $(\hat{\underline{e}}_1, \hat{\underline{e}}_2, \hat{\underline{e}}_3)$ necesarios para la determinación de cada CP. Se muestran en las columnas "Dim.1", "Dim.2" y "Dim.3".

```
 Variables
               Dim.1    ctr   cos2    Dim.2    ctr   cos2
Length      | -0.012  0.005  0.000 |  0.922 66.503  0.850 |
Left        |  0.803 21.880  0.644 |  0.387 11.694  0.149 |
Right       |  0.835 23.686  0.698 |  0.285  6.374  0.081 |
Bottom      |  0.698 16.545  0.487 | -0.301  7.088  0.091 |
Top         |  0.631 13.534  0.399 | -0.103  0.837  0.011 |
Diagonal    | -0.847 24.350  0.717 |  0.310  7.504  0.096 |
             Dim.3    ctr   cos2
Length      -0.016  0.031  0.000 |
Left         0.096  1.069  0.009 |
Right        0.115  1.525  0.013 |
Bottom       0.544 34.052  0.296 |
Top         -0.734 62.027  0.539 |
Diagonal     0.106  1.297  0.011 |
```

Por último, aparece información sobre las categorías del factor $Status$.

```
Supplementary categories
                Dist     Dim.1    cos2  v.test     Dim.2
counterfeit |   1.548 |   1.498   0.937  12.314 |  -0.348
genuine     |   1.548 |  -1.498   0.937 -12.314 |   0.348
```

```
               cos2  v.test     Dim.3    cos2  v.test
counterfeit   0.051  -4.340 |   0.001   0.000   0.012 |
genuine       0.051   4.340 |  -0.001   0.000  -0.012 |
```

Con la columna "Dist" se indica la distancia de cada categoría al centro de los ejes del gráfico.

Se indican las coordenadas y calidad de la representación ("Dim." y "cos2", respectivamente); además de un test "v.test" para los individuos de cada categoría. Veremos este test con más detenimiento en otras aplicaciones prácticas del ACP.

Una vez vistos los resultados obtenidos con el comando summary, podríamos; por lo tanto, definir las dos primeras CP como:

$$Y_1 = -0{,}012X_1 + 0{,}803X_2 + 0{,}835X_3 + 0{,}698X_4 + 0{,}631X_5 - 0{,}847X_6$$

$$Y_2 = 0{,}922X_1 + 0{,}387X_2 + 0{,}285X_3 - 0{,}301X_4 - 0{,}103X_5 + 0{,}310X_6$$

Se observa que la primera CP es esencialmente la suma de las variables $Left$ (X_2) y $Right$ (X_3) y la resta de la variable $Diagonal$ (X_6).

La segunda CP, sin embargo, depende en su mayor parte de la variable $Length(X_1)$.

Representaremos las correlaciones muestrales entre las variables y estas dos CP, que serán los ejes de abcisas y ordenadas; respectivamente. La suma de estas correlaciones al cuadrado para cada variable dará como resultado la longitud del vector en la representación.

Como hemos especificado con anterioridad, las correlaciones al cuadrado aparecen en la columna "cos2" del resumen.

Hacemos uso de la función `plot` y con el comando "choix" especificamos que sólo queremos una representación de las variables.

Puede verse en el gráfico que, efectivamente, la primera CP está muy bien explicada por las variables X_2 ($Left$), X_3 ($Right$) y X_6 ($Diagonal$).

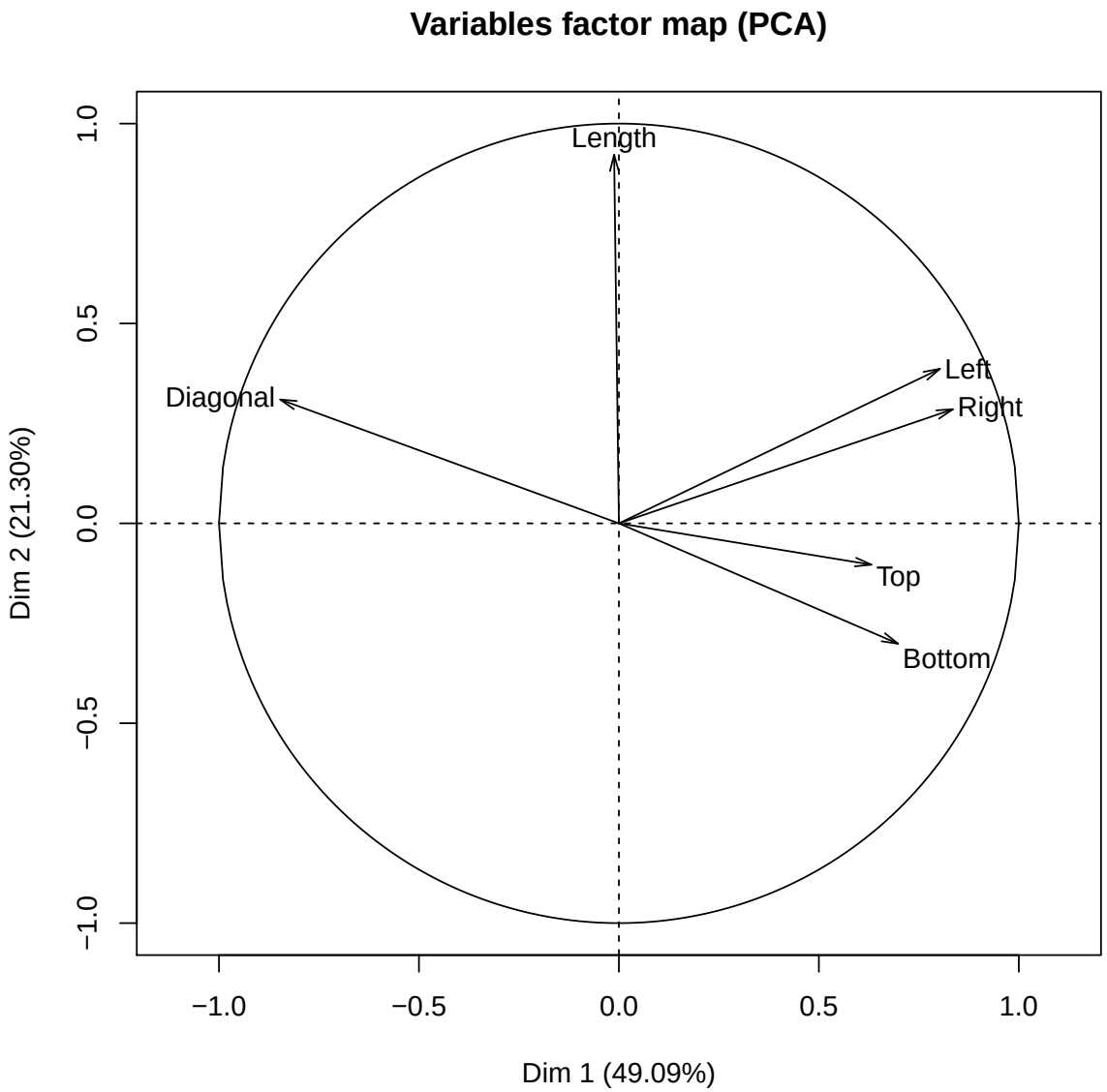
Consideremos la variable $Left$ (X_2), para la que tenemos que:

$$r_{X_2,Y_1} = (cos(X_2, Y_1)) = 0{,}80249611$$

$$r_{X_2,Y_2} = (cos(X_2, Y_2)) = 0{,}386$$

Es sencillo comprobar que se corresponde con la respresentación gráfica. La suma de estas correlaciones muestrales al cuadrado es igual al módulo del vector representado, que en este caso es $0{,}793$.

```
plot(res.pca,choix = "var")
```



Algo muy parecido sucede con la variable $Right$ (X_3), que está un poco más correlacionada con la primera CP y menos con la segunda CP; aunque a efectos de la representación la diferencia es muy pequeña. El módulo del vector es $0{,}779$ en este caso.

Por último, la variable $Diagonal$ (X_6) tiene correlaciones $r_{X_6,Y_1} = (cos(X_6, Y_1)) = -0{,}84675$ y $r_{X_6,Y_2} = (cos(X_6, Y_2)) = 0{,}309838$. Los signos de estas correlaciones los hemos obtenido a partir de la representación gráfica, ya que con la información que obtenemos de la función "res.pca" aparecen las correlaciones al cuadrado.

El valor del módulo del vector es $0{,}813$; lo que significa que es bastante cercano a la periferia del círculo, como se puede observar en el gráfico.

Por otra parte, la segunda CP, Y_2, está explicada en su mayor parte por la variable $Length(X_1)$.

X_1 tiene muy poca correlación con la primera CP (pues $\sqrt{(cos^2(X_1, Y_1))}$ apenas es cero), mientras que $r_{X_1,Y_2} = (cos(X_1, Y_2)) = 0{,}92195$. Además, si sumamos ambas correlaciones a cuadrado, obtenemos que el módulo del vector en el círculo es $0{,}85$.

Variables cercanas a la periferia indican que la suma de sus correlaciones con las CP son muy cercanas a 1, lo que indica que las CP explican una alta proporción de las variables.

También podemos deducir del gráfico que variables como $Left$ y $Right$ o Top y $Bottom$ están muy correladas entre sí, debido a que están muy próximas entre sí en la representación. Por otra parte, las variables $Diagonal$ y $Bottom$ están negativamente correladas de forma significativa; ya que están representadas en lados opuestos. Si existen variables ortogonales, eso indica que no están correladas.

1.5. Componentes Principales de una matriz de datos bivariante

Consideremos una matriz de datos bivariante con dos variables X_1 y X_2 con coeficiente de correlación lineal ρ. La matriz de correlación sería:

$$\mathbf{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

Necesitamos encontrar los autovalores y autovectores de $\mathbf{R}$.
En primer lugar, calculamos los autovalores:

$$|\mathbf{R} - \lambda\mathbf{I}| = 0 \Leftrightarrow (1-\lambda)^2 - \rho^2 = 0$$

El resultado de esta última igualdad es $\lambda_1 = 1 + \rho$, $\lambda_2 = 1 - \rho$.

Para el cálculo de los autovectores, se resuelve en primer lugar la ecuación $\mathbf{R}\underline{e}_1 = \lambda_1\underline{e}_1$. Obtenemos el siguiente sistema:

$$\begin{cases} e_{11} + \rho e_{12} = (1+\rho)e_{11}, \\ \rho e_{11} + e_{12} = (1+\rho)e_{12} \end{cases}$$

Como las dos ecuaciones son idénticas, tomamos $e_{11} = e_{12}$ e introducimos la restricción de normalización $(\underline{e}_1'\underline{e}_1 = 1)$.

Obtenemos que las coordenadas para el primer autovector son $e_{11} = e_{12} = \frac{1}{\sqrt{2}}$

Siguiendo el mismo razonamiento para resolver la ecuación $\mathbf{R}\underline{e}_2 = \lambda_2\underline{e}_2$, obtenemos: $e_{21} = \frac{1}{\sqrt{2}} = -e_{22}$.

Por lo tanto, las dos componentes principales se expresan como sigue:

$$Y_1 = \frac{1}{\sqrt{2}}(X_1 + X_2)$$

$$Y_2 = \frac{1}{\sqrt{2}}(X_1 - X_2)$$

El cálculo de las varianzas de las componentes es sencillo, pues como hemos definido con anterioridad, se corresponde con el autovalor λ_j y tiene como resultado:

$$Var(Y_1) = 1 + \rho, \quad Var(Y_2) = 1 - \rho$$

Si ρ fuera negativo, el orden de los autovalores cambiaría y por tanto las componentes principales se expresarían al contrario. Si ρ fuera nulo, los autovalores serían igual a uno.

Un par de aclaraciones necesarias:

- Hay una forma arbitraria de escoger el signo de los elementos de $\underline{e}_j$. Se suele tomar e_{j1} positivo.

- Los coeficientes que definen las dos componentes no dependen de ρ a pesar de que la proporción de varianza explicada por cada uno si lo haga.

1.6. Análisis de Componentes Principales y Regresión Lineal

Cuando queremos llevar a cabo una regresión lineal, es común que se nos planteen dificultades tales como un número demasiado elevado de predictores para la variable objetivo o bien colinealidad entre estos predictores; en otras palabras, variables con alta correlación entre sí.

Gracias a la técnica de Análisis de Componentes Principales pueden abordarse estos problemas, permitiendo simultáneamente reducir dimensión y explicar un gran porcentaje de variabilidad de la variable respuesta.

Veamos un ejemplo práctico que lo ilustre.

1.6.1. Aplicación

En este ejemplo, trabajamos con un conjunto de datos que recoge la contaminación del aire en varias ciudades de EEUU.

Los datos que manejamos se encuentran en el paquete *HSAUR3*, que se detalla en el último capítulo de esta memoria junto con el resto de paquetes utilizados en las aplicaciones prácticas de las técnicas.

Con el paquete *MVA* podemos representar boxplot bivariantes, por lo que nos será de gran utilidad.

```
oldopt <- options(digits=3)
options(width=60)
on.exit( {options(oldopt)} )
require(HSAUR3)
require(MVA)
```

```
data(USairpollution)
names(USairpollution)

## [1] "SO2"     "temp"    "manu"    "popul"   "wind"
## [6] "precip"  "predays"
```

En nuestros datos se miden 7 variables. Estas variables son:

$SO2$: contenido de SO_2 en el aire (microgramos por metro cúbico)
$temp$: temperatura media anual en grados Fahrenheit
$manu$: número de empresas de producción con 20 trabajadores o más
$popul$: población según el censo de 1970 (miles de personas)
$wind$: velocidad media del viento por año (millas por hora)
$precip$: precipitación media anual (pulgadas)
$predays$: número medio de días con precipitaciones por año

```
str(USairpollution)

## 'data.frame': 41 obs. of  7 variables:
##  $ SO2    : int  46 11 24 47 11 31 110 23 65 26 ...
##  $ temp   : num  47.6 56.8 61.5 55 47.1 55.2 50.6 54 49.7 51.5 ...
##  $ manu   : int  44 46 368 625 391 35 3344 462 1007 266 ...
##  $ popul  : int  116 244 497 905 463 71 3369 453 751 540 ...
##  $ wind   : num  8.8 8.9 9.1 9.6 12.4 6.5 10.4 7.1 10.9 8.6 ...
##  $ precip : num  33.36 7.77 48.34 41.31 36.11 ...
##  $ predays: int  135 58 115 111 166 148 122 132 155 134 ...
```

Observamos que trabajamos con un data.frame de dimensión 41×7, y por lo tanto, medimos las 7 variables anteriores en 41 individuos.
Veamos un resumen de los datos.

```
summary(USairpollution)

##       SO2             temp            manu       
##  Min.   :  8.00   Min.   :43.50   Min.   :  35.0  
##  1st Qu.: 13.00   1st Qu.:50.60   1st Qu.: 181.0  
##  Median : 26.00   Median :54.60   Median : 347.0  
##  Mean   : 30.05   Mean   :55.76   Mean   : 463.1  
##  3rd Qu.: 35.00   3rd Qu.:59.30   3rd Qu.: 462.0  
##  Max.   :110.00   Max.   :75.50   Max.   :3344.0  
##      popul            wind           precip      
##  Min.   :  71.0   Min.   : 6.000   Min.   : 7.05  
##  1st Qu.: 299.0   1st Qu.: 8.700   1st Qu.:30.96  
##  Median : 515.0   Median : 9.300   Median :38.74  
##  Mean   : 608.6   Mean   : 9.444   Mean   :36.77  
##  3rd Qu.: 717.0   3rd Qu.:10.600   3rd Qu.:43.11  
##  Max.   :3369.0   Max.   :12.700   Max.   :59.80  
##     predays      
##  Min.   : 36.0  
##  1st Qu.:103.0  
##  Median :115.0  
##  Mean   :113.9  
##  3rd Qu.:128.0  
##  Max.   :166.0
```

No tendremos en cuenta la variable $SO2$, y nos centraremos en dos variables basadas en la ecología humana ($popul$, $manu$) y cuatro basadas en el clima ($temp$, $wind$, $precip$, $predays$).

Dado que las variables están medidas en escalas muy distintas, obtenemos las componentes a partir de la matriz de correlaciones en lugar de usar la matriz de varianzas y covarianzas.

Como especificamos en la sección anterior, esto es equivalente a trabajar con la matriz de datos estandarizados.

Con la siguiente instrucción, transformamos la variable $temp$ por su equivalente negativa, ganando así en interpretación. La llamaremos $negtemp$.

```
USairpollution$negtemp<-USairpollution$temp*(-1)
USairpollution$temp<-NULL
```

Calculamos la matriz de correlaciones.

```
cor(USairpollution[,-1])

##                manu       popul        wind      precip
## manu     1.00000000  0.95526935  0.23794683 -0.03241688
## popul    0.95526935  1.00000000  0.21264375 -0.02611873
## wind     0.23794683  0.21264375  1.00000000 -0.01299438
## precip  -0.03241688 -0.02611873 -0.01299438  1.00000000
## predays  0.13182930  0.04208319  0.16410559  0.49609671
## negtemp  0.19004216  0.06267813  0.34973963 -0.38625342
##             predays     negtemp
## manu     0.13182930  0.19004216
## popul    0.04208319  0.06267813
## wind     0.16410559  0.34973963
## precip   0.49609671 -0.38625342
## predays  1.00000000  0.43024212
## negtemp  0.43024212  1.00000000
```

Cabe destacar que las variables $popul$ y $manu$ están muy relacionadas, pues su coeficiente de correlación es $0{,}95527$.

Construimos una matriz scatterplot de las 6 variables, incluyendo además el histograma para cada variable en la diagonal. Esta matriz esta formada por una serie de gráficos de puntos dispersos, donde cada eje representa una variable y cada punto un individuo de la muestra. Puede sernos útil para visualizar outliers.

```
data("USairpollution",package="HSAUR3")
panel.hist<-function(x,...){
 usr<-par("usr");on.exit(par(usr))
 par(usr=c(usr[1:2],0,1.5))
 h<-hist(x,plot=FALSE)
 breaks<-h$breaks;nB<-length(breaks)
 y<-h$counts; y<-y/max(y)
 rect(breaks[-nB],0,breaks[-1],y,col="grey",...)
}
USairpollution$negtemp<-USairpollution$temp*(-1)
USairpollution$temp<-NULL
pairs(USairpollution[,-1],diag.panel = panel.hist,pch=".",cex=1.5)
```

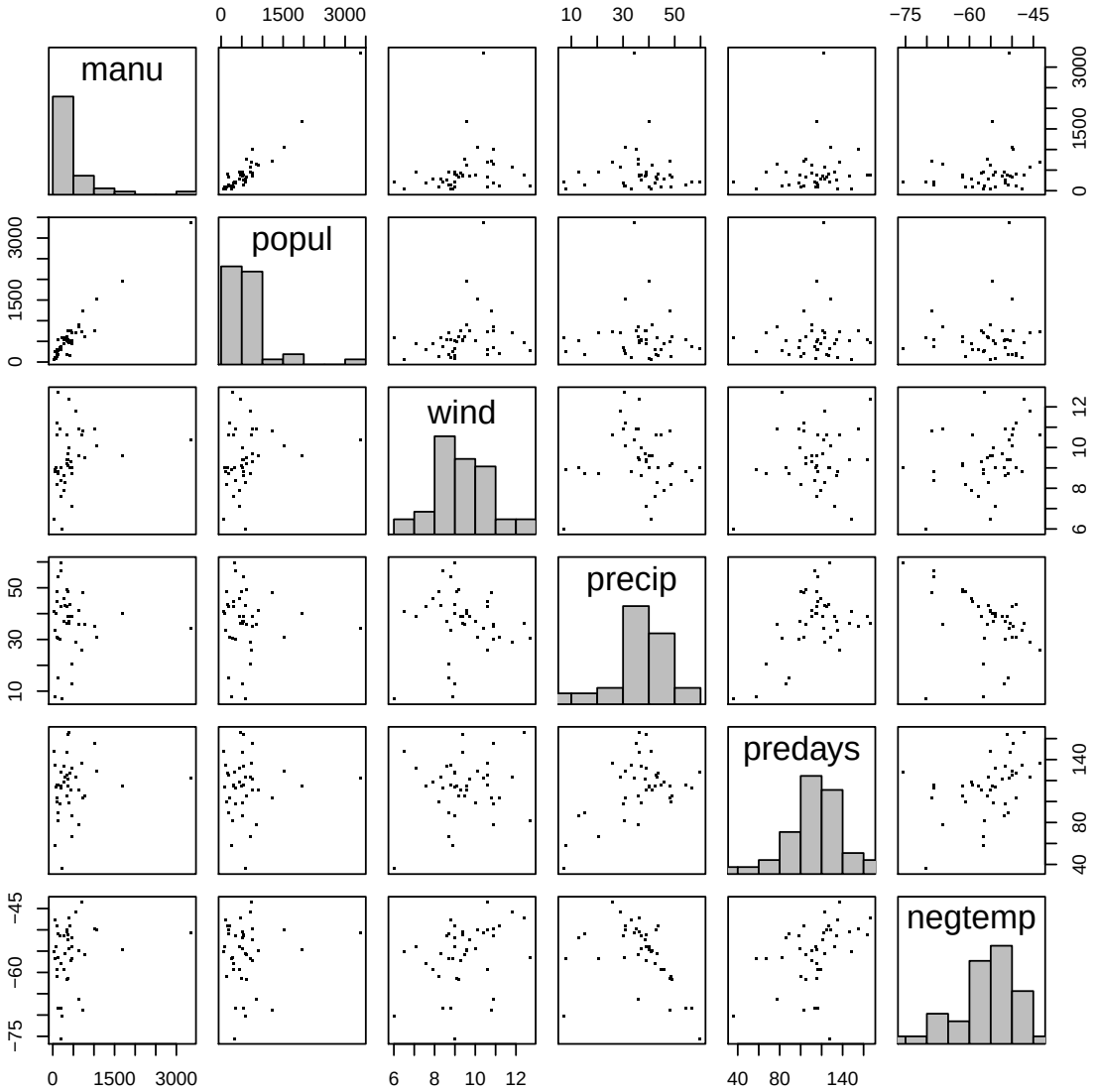


Figura 1.1: Matriz scatterplot de las 6 variables consideradas en los datos.

Ahora, llevamos a cabo el Análisis de Componentes Principales.

```
usair_pca<-princomp(USairpollution[,-1],cor=TRUE)
summary(usair_pca,loadings=TRUE)

## Importance of components:
##                           Comp.1    Comp.2    Comp.3
## Standard deviation     1.4819456 1.2247218 1.1809526
## Proportion of Variance 0.3660271 0.2499906 0.2324415
## Cumulative Proportion  0.3660271 0.6160177 0.8484592
##                           Comp.4     Comp.5      Comp.6
## Standard deviation     0.8719099 0.33848287 0.185599752
## Proportion of Variance 0.1267045 0.01909511 0.005741211
## Cumulative Proportion  0.9751637 0.99425879 1.000000000
##
## Loadings:
##         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## manu    -0.612  0.168 -0.273 -0.137  0.102  0.703
## popul   -0.578  0.222 -0.350                -0.695
## wind    -0.354 -0.131  0.297  0.869 -0.113
## precip         -0.623 -0.505  0.171  0.568
## predays -0.238 -0.708        -0.311 -0.580
## negtemp -0.330 -0.128  0.672 -0.306  0.558 -0.136
```

Gracias al comando `summary` se observa que las tres primeras componentes explican el 0.8484592 % de la varianza.

Si queremos interpretar los datos usando etiquetas para las componentes, tendremos que examinar los coeficientes que definen cada una de ellas y que están recogidos en las columnas "Comp." del resumen.
Por ejemplo, la primera componente podría etiquetarse como "calidad de vida", la segunda como "tiempo húmedo" (dado que los coeficientes más altos acompañan a las variables $precip$ y $predays$), y la tercera como "tipo de clima" (las variables con más peso son $precip$ y $negtemp$).

Podemos representar los individuos mediante un boxplot bivariante con ejes las 3 primeras componentes principales, ya que es un método más objetivo que las matrices scatterplot para etiquetar outliers.
Está basado en la construcción de un par de elipses concéntricas, de forma que en la de menor radio se incluye el 50 % de los datos y en la otra se incluyen los datos que puedan ser outliers.

Podemos deducir que "Chicago" es un outlier; así como "Phoenix" y "Philadelphia".
"Phoenix" ofrece la mejor calidad de vida y "Buffalo" sería la ciudad con el ambiente más húmedo.

```
pairs(usair_pca$scores[,1:3],ylim=c(-6,4),xlim=c(-6,4),
    panel=function(x,y,...){
      text(x,y,abbreviate(row.names(USairpollution)),
           cex=0.6)
      bvbox(cbind(x,y),add=TRUE)
    })
```

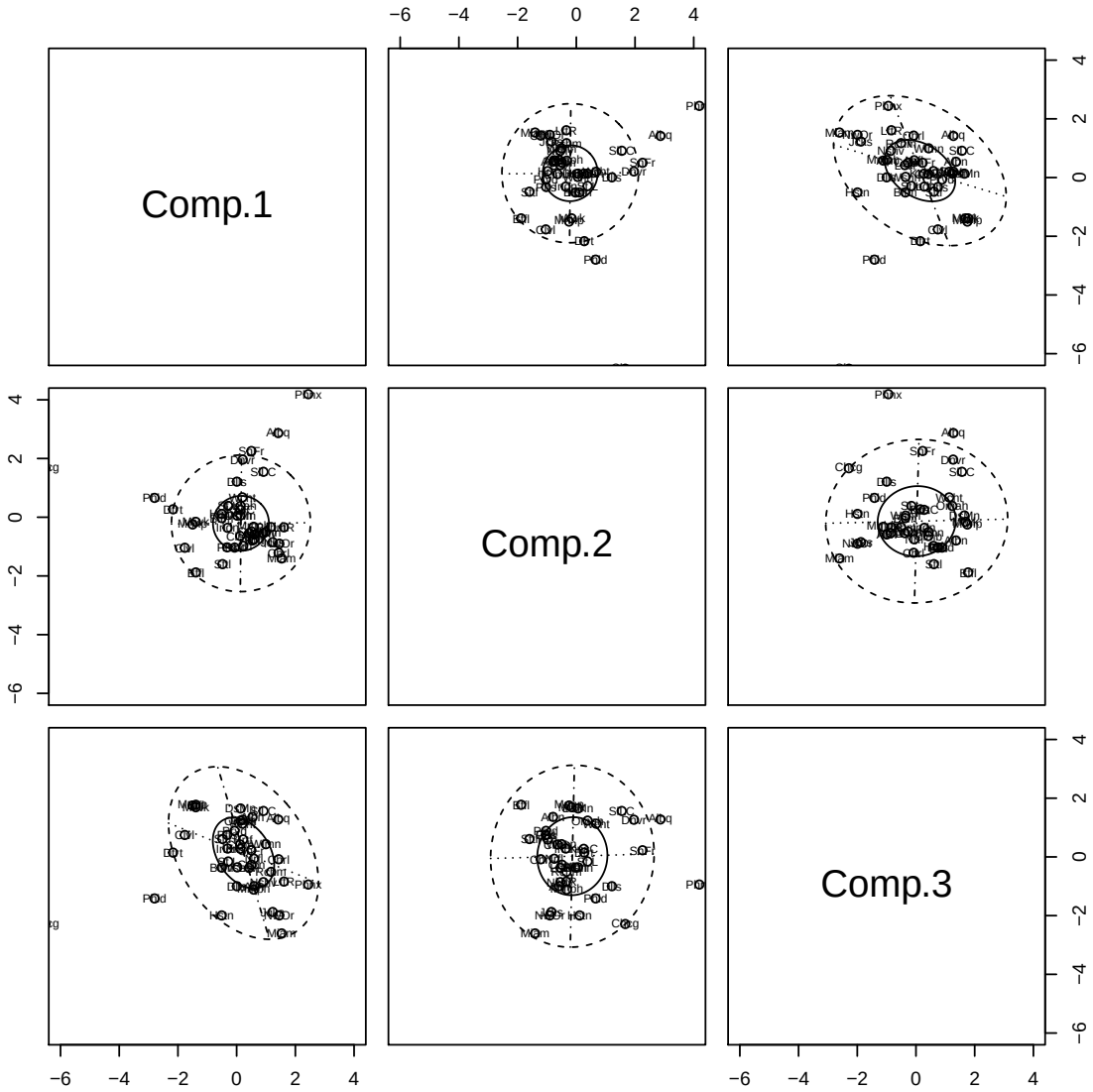


Figura 1.2: Boxplots bivariantes de las tres primeras componentes principales.

A continuación, determinaremos qué variables son las mejores predictoras del grado de contaminación en el aire de una ciudad (variable $SO2$).

Esta pregunta suele responderse con regresión lineal múltiple, pero no puede aplicarse en este caso debido a un problema de colinealidad en los datos (como apuntamos anteriormente, las variables $popul$ y $manu$ están altamente correlacionadas).

Podríamos prescindir de estas dos variables, aunque es mejor aplicar regresión lineal múltiple a las componentes principales de las variables originales. Se sigue este paso porque las componentes principales son incorreladas entre sí.

Pero, para resolver el problema de esta forma necesitamos preguntarnos: ¿Cuántas componentes principales necesitaría usar como variables explicativas en la regresión?

Llevamos a cabo la regresión lineal en R con las 3 primeras CP que, como indicamos antes, explican casi el 85 % de la variabilidad.

```
usair_reg<-lm(SO2~usair_pca$scores[,1:3],
              data=USairpollution)
summary(usair_reg)
```

```
Call:
lm(formula = SO2 ~ usair_pca$scores[, 1:3], data = USairpollution)

Residuals:
   Min     1Q Median     3Q    Max
-36.42 -10.98  -3.18  12.09  61.27

Coefficients:
                                Estimate Std. Error t value
(Intercept)                       30.049      2.907   10.34
usair_pca$scores[, 1:3]Comp.1     -9.942      1.962   -5.07
usair_pca$scores[, 1:3]Comp.2     -2.240      2.374   -0.94
usair_pca$scores[, 1:3]Comp.3     -0.375      2.462   -0.15
                                Pr(>|t|)
(Intercept)                      1.8e-12 ***
usair_pca$scores[, 1:3]Comp.1    1.1e-05 ***
usair_pca$scores[, 1:3]Comp.2        0.35
usair_pca$scores[, 1:3]Comp.3        0.88
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 18.6 on 37 degrees of freedom
Multiple R-squared:  0.418,Adjusted R-squared:  0.371
F-statistic: 8.87 on 3 and 37 DF,  p-value: 0.000147
```

Claramente, la puntuación de la primera componente principal es la más predictiva, pues es la que tiene menor error.
Etiquetada como "calidad de vida", esta combinación de las variables originales es la que mejor predice el grado de contaminación en el aire (SO_2).

1.7. Escogiendo el número de componentes

Hay distintos criterios para decidir cuantas componentes principales necesitamos. Estos son:

- Criterio del porcentaje.

  Tomamos las componentes suficientes para explicar un porcentaje alto de la variabilidad de las variables originales. La variabilidad explicada por cada componente (autovalores) suele estabilizarse en su representación a partir de un cierto número. Éste índica la última componente que tendremos en cuenta.

  Normalmente, tomamos componentes hasta que la varianza explicada esté en torno a un 70 % o 90 %; aunque el punto óptimo de corte suele ser $\lambda^* = 0'7 * v$, siendo $v = \frac{tr(\Sigma)}{p}$

- Criterio de Kaiser.

  Cuando trabajamos con la matriz $\mathbf{R}$, dado que se asumen varianzas iguales a uno no tendremos en cuenta los autovalores menores que esa cantidad. Estudios de Montecarlo afirman que el valor óptimo de λ^* para el punto de corte en la gráfica es 0'7.

- Excluir las componentes cuyos autovalores estén por debajo de la media de los autovalores.

Todos estos métodos dan lugar a diferentes conclusiones.

1.8. Aplicación práctica del ACP

Cuando el tipo de estudio que llevamos a cabo es de descubrimiento de clases, necesitamos definir distancias entre genes o muestras para reconocer patrones o identificar efecto lote.

Es por ello por lo que surge la necesidad de reducir la dimensión de la matriz de datos, para facilitar la interpretación mediante la observación de tendencias en la visualización gráfica.
La forma de hacerlo es mediante técnicas de estadística descriptiva como ACP o clustering.

En estos ejemplos veremos el Análisis de Componentes Principales. Para hacer una reducción efectiva nos basamos en el porcentaje de variabilidad explicada, ya que puede servirnos para

alcanzar los objetivos anteriores.

Este método suele usarse como una reducción previa; es decir, como una técnica exploratoria necesaria para interpretar luego los datos ayudándonos de otras técnicas.

Datos

Esta sección introduce los datos que usaremos en los siguientes ejercicios prácticos.

1. El conjunto de datos `chicken` contiene datos de expresión génica de 7406 genes (variables) medidos en 43 pollos (muestras). También hay una variable categórica que contiene 6 dietas en función de distintos niveles.

2. `breastCancer` es un conjunto de datos de expresión génica (microarrays, tipo 'hgu95-A') obtenidos en un estudio donde se ha analizado el efecto de diferentes tratamientos y tiempos de exposición en niveles de expresión génica de un grupo de mujeres afectadas por cáncer de pecho. El estudio está disponible en la base de datos GEO (Gene Expression Omnibus). Para este estudio nos hemos basado tan sólo en 18 muestras, guardadas en el fichero `Breast-Cancer.txt`.

Paquetes

1. FactoMineR
   Es un paquete de R que permite realizar Análisis Exploratorio de Datos Multivariante y Minería de Datos. En concreto, tiene implementados métodos para realizar Análisis de Componentes Principales, Análisis de Correspondencias, y Análisis de Correspondencias Múltiples, así como otros métodos estadísticos multivariantes avanzados. Destacamos que con él se pueden obtener gráficos de gran calidad. Este paquete ha sido desarrollado por Husson; Josse, Lê, y Mazet (2007). La última versión es FactomineR 1.31.5 publicada el 07/01/2016, y que se encuentra disponible en https://cran .rproject.org/web/packages/FactoMineR/index.html.

1.8.1. Exploración de datos

Cargamos el paquete *FactoMineR* y construimos el data.frame "poulet" a partir del fichero "chicken.txt".

También calculamos la dimensión de la matriz y obtenemos como resultado 7406 filas y 43 columnas.

Dado que en la estructura de un `data.frame` las columnas se corresponden con las variables y las filas con los individuos, trasponemos la matriz de partida y la renombramos; obteniendo 43 observaciones de 7406 variables.

```
require(FactoMineR)
poulet <- read.table(file.path("data", "chicken.txt"),
            header=T,sep="\t", dec=".", row.names=1)
dim(poulet)

## [1] 7406   43

poulet=as.data.frame(t(poulet))
dim(poulet)

## [1]   43 7406
```

Creamos una nueva variable categórica definiendo 6 dietas y, convirtiéndola previamente en un factor, la añadimos como columna con la instrucción `cbind.data.frame` a nuestro data frame "poulet".
Con la función `colnames` especificamos el nombre de la variable.

```
poulet =cbind.data.frame(as.factor(c(rep("N",6),
rep("J16",5), rep("J16R5",8), rep("J16R16",9),
rep("J48",6), rep("J48R24",9))), poulet)
colnames(poulet)[1] = "Diet"
```

Ahora, procedemos a hacer el ACP, indicando que representamos los individuos con diferente color basándonos en nuestra variable categórica "Dieta"; que es la primera, y que no introducimos gráficos.
Hemos convertido la matriz de datos al formato data.frame para poder ahora aplicar la función PCA.

```
res.pca = PCA(poulet, quali.sup=1, graph=F)
```

El comando `summary` nos dará información sobre el ACP.

```
summary(res.pca)
```

Mostraremos los resultados más significativos.

En primer lugar aparecen los autovalores (varianza, porcentaje de varianza explicada y porcentaje de varianza acumulado) para las 42 CP que considera.

```
Eigenvalues
                        Dim.1    Dim.2    Dim.3    Dim.4
Variance             1453.572  692.788  536.208  434.453
% of var.              19.627    9.354    7.240    5.866
Cumulative % of var.   19.627   28.981   36.222   42.088

                       Dim.41   Dim.42
Variance               38.259   33.786
% of var.               0.517    0.456
Cumulative % of var.   99.544  100.000
```

Para los primeros individuos se calcula la distancia a la categoría a la que pertenecen, además de los valores de $\underline{y}_{(j)}, j = 1, 2, 3$, que aparecen en las columnas "Dim.1", "Dim.2" y "Dim.3"; respectivamente.

```
Individuals
               Dist     Dim.1     ctr    cos2     Dim.2
N_1        |  72.888 |   6.168   0.061   0.007 | -17.356
N_2        |  69.634 |  15.472   0.383   0.049 | -14.354
N_3        |  70.814 |  16.951   0.460   0.057 | -14.533
j16_3      |  82.817 |   8.092   0.105   0.010 |   9.506
j16_4      |  74.501 |  28.090   1.262   0.142 |   4.704

               ctr    cos2     Dim.3     ctr    cos2
N_1          1.011   0.057 | -47.024   9.590   0.416 |
N_2          0.692   0.042 | -37.041   5.951   0.283 |
N_3          0.709   0.042 | -41.914   7.619   0.350 |
j16_3        0.303   0.013 |  43.142   8.072   0.271 |
j16_4        0.074   0.004 |  41.583   7.499   0.312 |
```

También se indica la contribución de cada individuo en la determinación de cada CP ("ctr"), así como la calidad de la representación de estos ("cos2").

Para las variables se obtienen los primeros autovectores (las columnas "Dim.1" y "Dim.2" se corresponden con $\underline{\hat{e}}_1$ y $\underline{\hat{e}}_2$). También aparecen la contribución de cada variable en una CP determinada y la calidad de representación o correlación al cuadrado entre ellas; como hemos comentado en ejemplos anteriores.

```
Variables
                Dim.1    ctr   cos2    Dim.2    ctr   cos2
A4GALT       | -0.526  0.019  0.276 | -0.060  0.001  0.004
A4GNT        | -0.124  0.001  0.015 |  0.217  0.007  0.047
AACS         |  0.332  0.008  0.110 |  0.036  0.000  0.001
AADACL1      |  0.253  0.004  0.064 |  0.284  0.012  0.081
AADACL2      |  0.510  0.018  0.260 | -0.486  0.034  0.236
```

Por último, para cada categoría de la variable "Dieta" aparece la distancia al centro de los ejes de la representación gráfica. Cabe señalar que cada categoría se proyecta en el baricentro de los individuos que pertenecen a la misma.

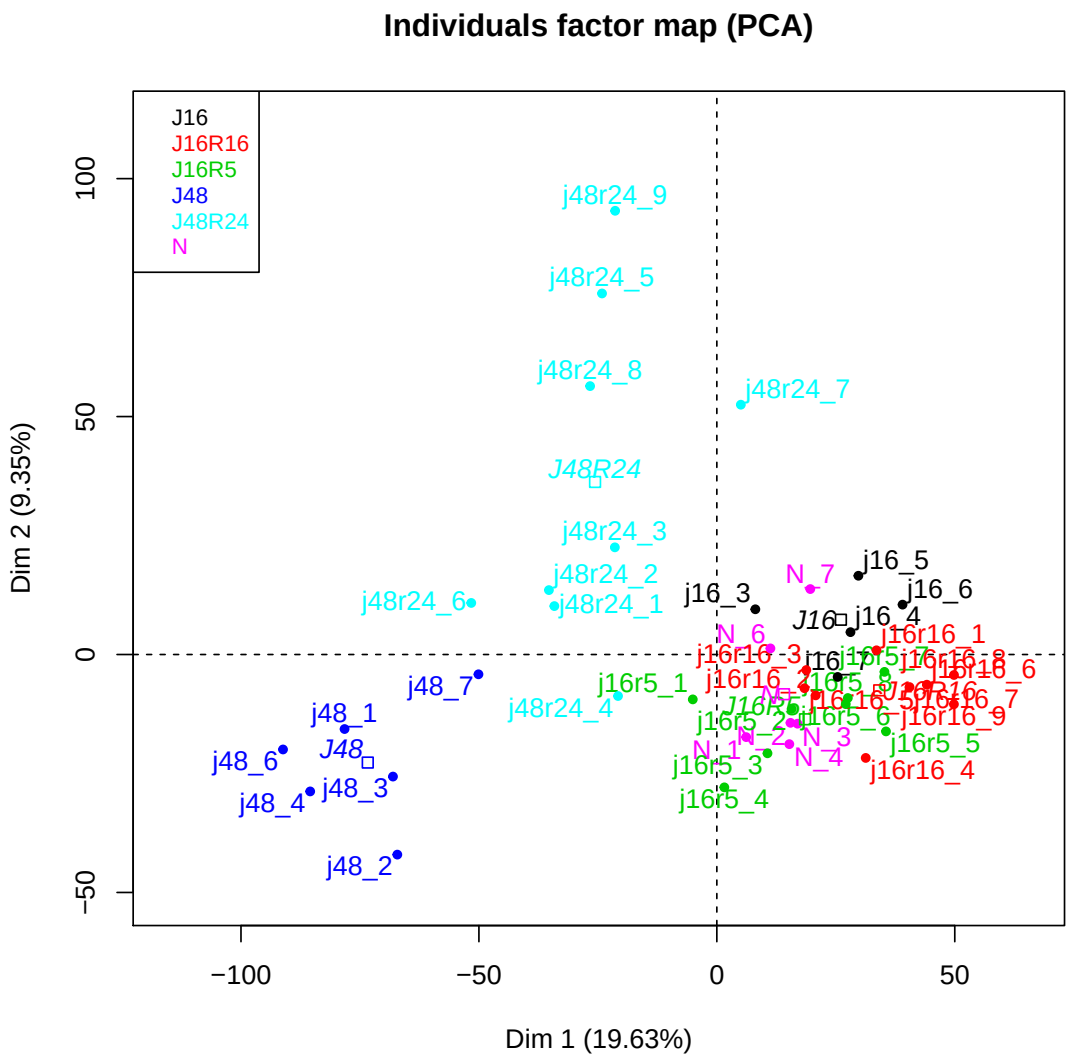
Aparecen también las coordenadas de las categorías para las primeras CP, la correlación entre ellas y un test de normalidad "v.test"; considerando como variable de agrupación la variable categórica. En este test se basan las elipses de confianza que se mostrarán posteriormente.

```
Supplementary categories
               Dist     Dim.1    cos2  v.test
J16         |  55.795 |  26.069   0.218   1.607
J16R16      |  42.365 |  34.124   0.649   2.984
J16R5       |  41.933 |  18.534   0.195   1.506
J48         |  79.310 | -73.356   0.855  -5.021
J48R24      |  48.622 | -25.591   0.277  -2.238
N           |  46.001 |  14.122   0.094   0.967

             Dim.2  cos2  v.test
J16       |   7.307 0.017   0.653
J16R16    |  -7.524 0.032  -0.953
J16R5     | -13.613 0.105  -1.602
J48       | -22.706 0.082  -2.251
J48R24    |  36.262 0.556   4.594
N         |  -8.340 0.033  -0.827
```

En el siguiente gráfico aparecen las proyecciones de los individuos que resultan del ACP así como las distintas categorías del factor "Dieta". Cada individuo se representa con un color distinto en función de la dieta que siga.

Se representan las proyecciones porque, como sabemos por teoría, las distancias al cuadrado entre elementos originales y transformados se conservan ($d^2(\underline{y}_h, \underline{y}_i) = d^2(\underline{x}_h, \underline{x}_i)$).

```
plot(res.pca, choix="ind", habillage=1)
```

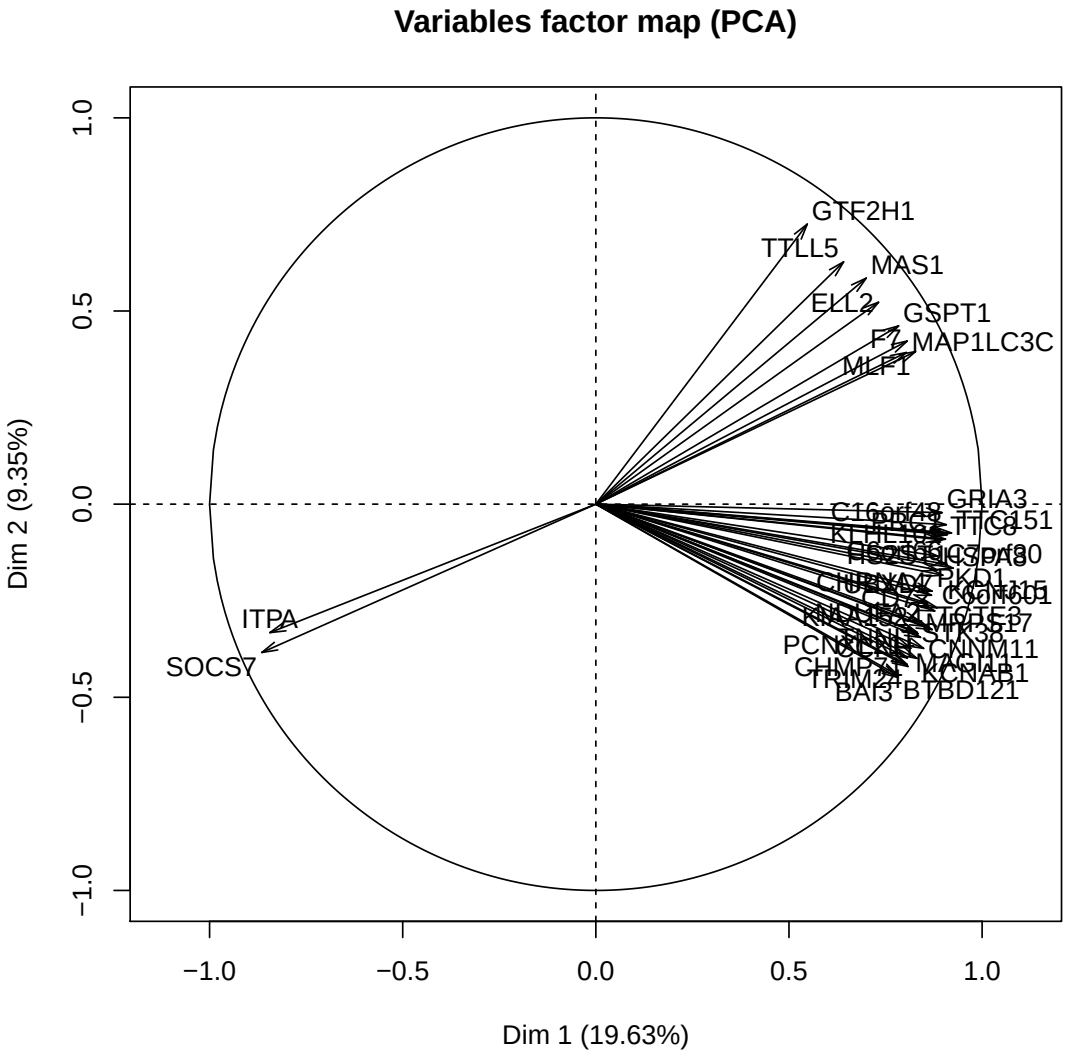


Podemos ver que los ejes “Dim 1” y “Dim 2” son las CP1 y CP2 respectivamente, y aparece indicado el porcentaje de variabilidad explicado por cada una de ellas.
La primera CP distingue los individuos con las dietas “j48” y “j48R24” del resto; mientras que la segunda CP distingue los individuos que siguen las dietas “j16” y “j48R24” de los restantes individuos de la muestra.

A continuación, representamos en un círculo de correlación la relación entre las variables iniciales y las dos primeras CP, teniendo en cuenta sólo aquellas con correlación igual o

superior a $0{,}8$.

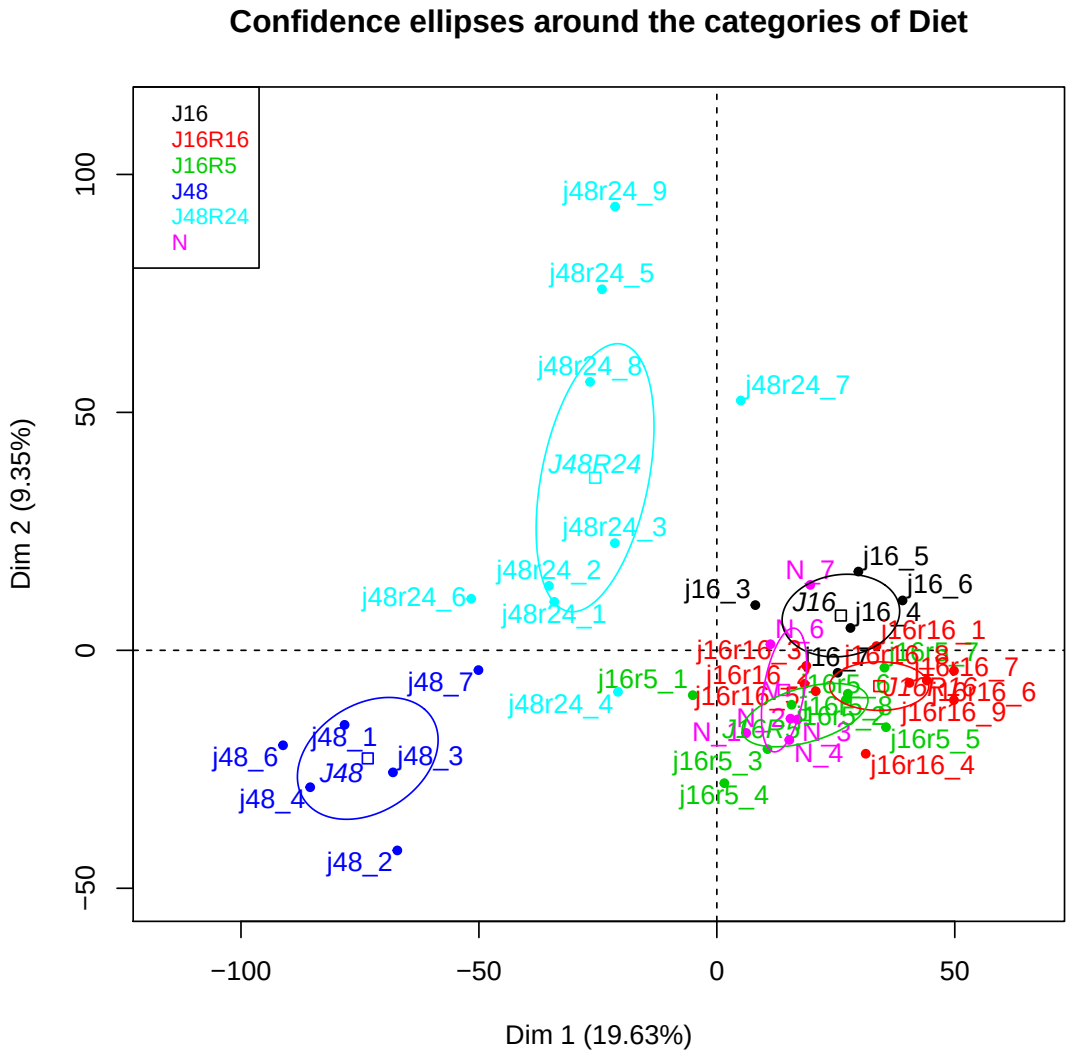
```
plot(res.pca, choix="var", lim.cos2.var=0.8)
```



Como ya vimos en este capítulo, variables próximas en el círculo de correlación están muy correladas entre sí, si son ortogonales no están correladas y si están en lados opuestos están negativamente correladas de forma significativa. Además, cuando las variables están cercanas al centro, la información está en otros ejes (componentes principales).

Por último, visualizamos las elipses de confianza de las clases de la variale categórica tras el ACP con el comando `plotellipses()`. Se calculan por defecto a un nivel de confianza $0{,}95$ suponiendo la normalidad de los datos originales.

```
plotellipses(res.pca)
```



Estas elipses son útiles para comparar categorías, decidir si una observación procede o no de una determinada población con distribución normal bivariante y además, detectar puntos anómalos.
En este caso, las elipses de las categorías “J48R24” y “J48” no se solapan, lo que significa que estas son significativamente diferentes. Sin embargo, las elipses correspondientes a las 4 categorías restantes muestran que no existen diferencias significativas entre ellas.

1.8.2. Detección de efecto lote en experimentos ómicos

Los datos producidos en estudios de tipo microarrays pueden confundirse por el conocido como *efecto lote*. Este efecto se debe al procesamiento de los datos en distintos lotes (días, platos, operarios,...) y puede crear confusión si oculta el efecto de los tratamientos objeto de estudio.

Cuando hay diferentes condiciones experimentales, si individuos de cada grupo están bien equilibrados entre los lotes el problema no es grave (o bien se compensa o bien se puede eliminar con técnicas ANOVA).
Sin embargo, si lotes y grupos están mezclados no se sabe si las diferencias que detectemos se deben a condiciones experimentales diferentes o a haber sido procesados en diferentes lotes. Esto conlleva que los datos puedan perder su valor y el estudio estadístico no sea válido.

A continuación, se recoge un ejemplo en el que utilizamos ACP para detectar el efecto lote. Trabajaremos con el conjunto "Breast_Cancer.txt" que, tal y como especificamos en el apartado datos de esta sección 1.8, contiene datos de expresión génica medidos en 18 mujeres con cáncer de pecho.

```
bcData <- read.table(file.path("data",
"Breast_Cancer.txt"), head=T, sep="\t", row.names=1,
as.is=TRUE)
class(bcData)

## [1] "data.frame"

dim(bcData)

## [1]    18 12630
```

Podemos observar que "bcData" es un data.frame con 18 filas y 12630 columnas.

Se ha utilizado la función `substr`, que permite seleccionar determinadas posiciones de una cadena de caracteres; y hemos recortado los nombres de las filas de nuestra matriz para facilitar la lectura de las salidas.

```
rownames(bcData)<- substr(rownames(bcData),4,8)
```

No vamos a considerar dos de las primeras filas de la matriz, pues son datos de control que carecen de interés en nuestro estudio. Por tanto, nuestro data.frame será de dimensión

16×12630.

Además, vamos a considerar las 3 primeras variables como categóricas (Tratamiento, Tiempo y Lote).

```
bcData<- bcData[-(1:2),]
head(names(bcData))

## [1] "Treatment"                "Time"
## [3] "Batch"                    "Treatment.Combination"
## [5] "X100_g_at"                "X101_at"

for(i in 1:3) bcData[,i]<- as.factor(bcData[,i])
```

Como hemos convertido las variables Tratamiento, Tiempo y Lote en factores, la función `summary` nos muestra una tabla de frecuencia con cada una de sus categorías.
La variable "Batch" que indica el lote en que se procesaron los datos, es la que nos interesará para detectar si los datos son válidos o no.

```
summary(bcData$Treatment)

##      E2 E2+ICI E2+Ral E2+TOT
##       4       4       4       4

summary(bcData$Time)

##  8 48
##  8  8

summary(bcData$Batch)

## A B
## 8 8
```

Una vez visto esto, llevamos a cabo el ACP sin incluir la variable "Treatment.Combination", pues aporta información redundante.

```
res.pca = PCA(bcData[,-4], quali.sup=1:3, graph=F)
summary(res.pca)
```

Con el comando summary, hemos obtenido un resumen del ACP.

En particular, información sobre los autovalores, los individuos y las variables. Además; información sobre las 3 variables categóricas y cada clase (E2, E2+ICI, E2+Ral, E2+TOT, Time8, Time48, A y B).

Mostraremos tan sólo los resultados más relevantes sobre algunos de los individuos y variables. La información obtenida es la misma que en la aplicación anterior de ACP (datos "poulet").

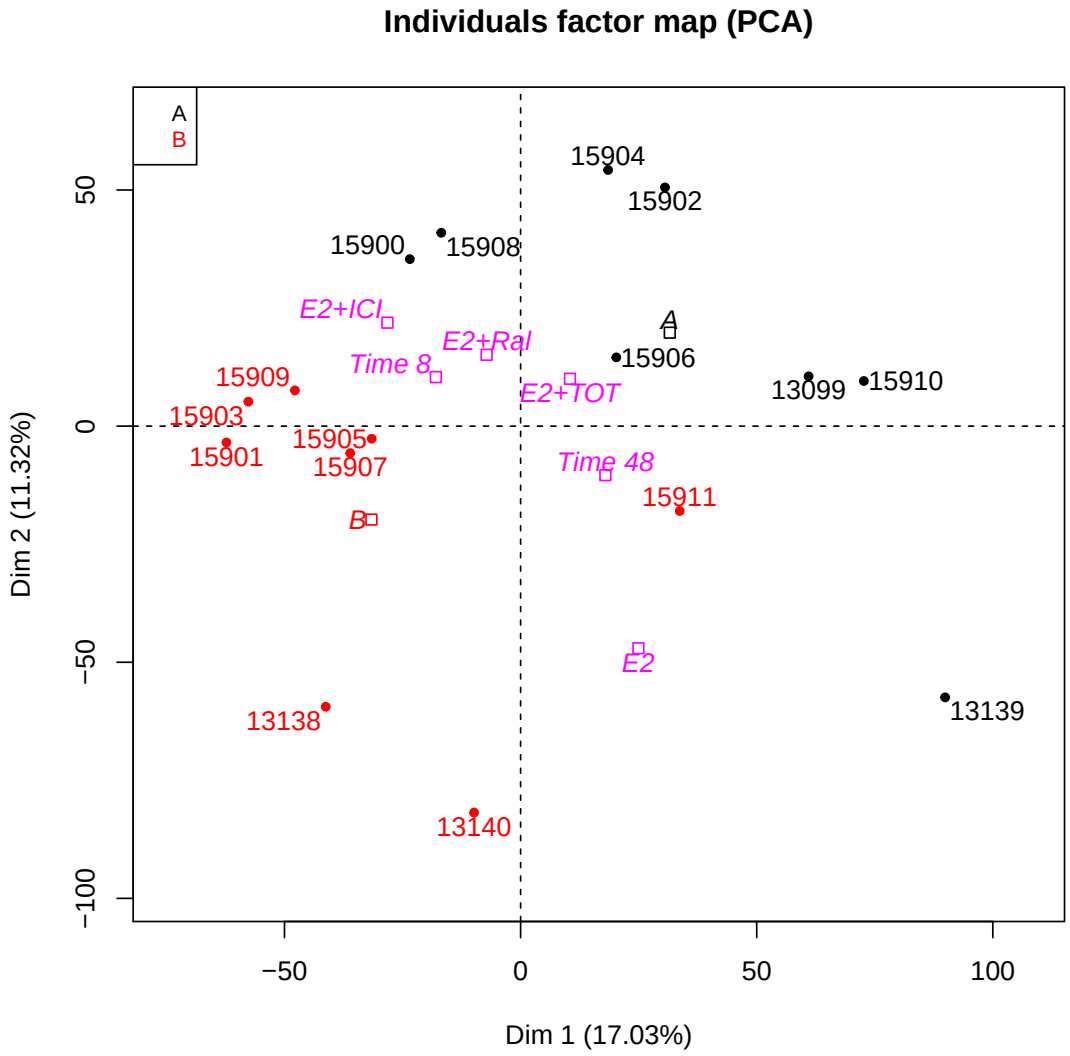
```
Eigenvalues
                       Dim.1   Dim.2   Dim.3 |  Dim.15
Variance             2145.25 1426.41 1249.83 |  452.18
% of var.              17.03   11.32    9.92 |    3.59
Cumulative % of var.   17.03   28.35   38.27 |  100.00

Individuals
           Dist     Dim.1     ctr    cos2     Dim.2
99     | 124.565 |  60.994  10.839   0.240 |  10.537
38     | 116.159 | -41.266   4.961   0.126 | -59.426
39     | 133.028 |  89.899  23.545   0.457 | -57.436

             ctr    cos2     Dim.3     ctr    cos2
99         0.486   0.007 | -20.336   2.068   0.027 |
38        15.473   0.262 | -32.974   5.437   0.081 |
39        14.454   0.186 | -21.334   2.276   0.026 |

Variables
              Dim.1    ctr   cos2    Dim.2    ctr   cos2
X100_g_at |  0.364  0.006  0.132 | -0.018  0.000  0.000 |
X101_at   | -0.199  0.002  0.040 | -0.459  0.015  0.210 |
X102_at   |  0.768  0.027  0.589 |  0.004  0.000  0.000 |

              Dim.3    ctr   cos2
X100_g_at    -0.343  0.009  0.118 |
X101_at       0.286  0.007  0.082 |
X102_at      -0.182  0.003  0.033 |

Supplementary categories
                Dist     Dim.1    cos2  v.test     Dim.2
E2          |  63.713 |  24.946   0.153   1.204 | -47.038
Time 8      |  33.218 | -17.948   0.292  -1.501 |  10.387
A           |  41.098 |  31.590   0.591   2.642 |  19.788
B           |  41.098 | -31.590   0.591  -2.642 | -19.788
```

	cos2	v.test	Dim.3	cos2	v.test	
E2	0.545	-2.785	-25.005	0.154	-1.582	
Time 8	0.098	1.065	-12.985	0.153	-1.423	
A	0.232	2.029	-16.493	0.161	-1.807	
B	0.232	-2.029	16.493	0.161	1.807	

Es sencillo observar que las 2 primeras CP explican un 28.347 % de la variabilidad total.

Representamos los resultados del ACP obtenidos para los individuos, dando diferentes colores para aquellos que fueron procesados en el lote A de los que fueron procesados en el B.

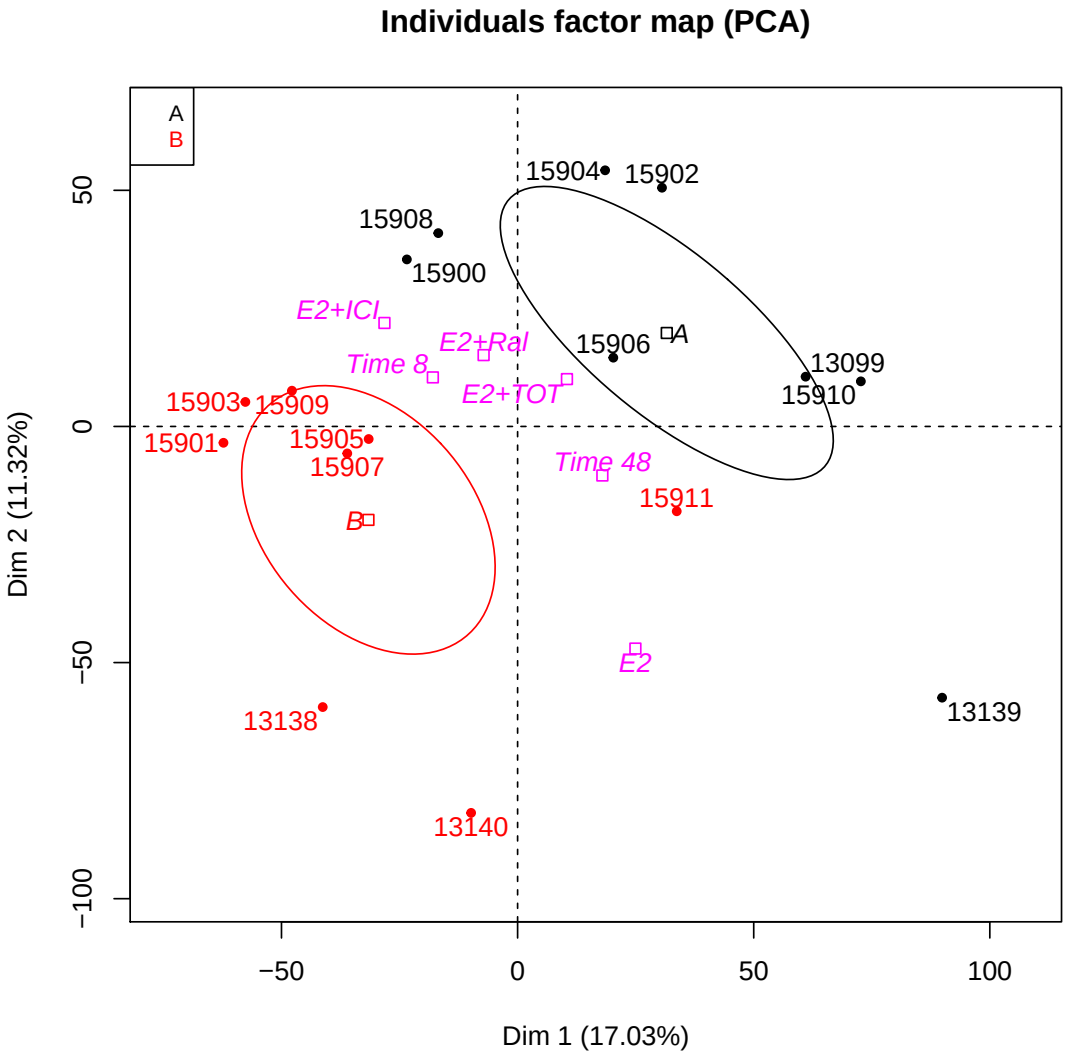
```
plot(res.pca, choix="ind", habillage=3)
```



Los individuos que se corresponden con el lote A están representados en color negro, y los que se corresponden con B en color rojo.
Además, las clases de cada una de las 3 variables categóricas que hemos considerado están representadas en color rosa.

Añadimos ahora las elipses de confianza:

```
aa=cbind.data.frame(bcData[,3],res.pca$ind$coord)
bb=coord.ellipse(aa, bary=TRUE)
plot.PCA(res.pca, habillage=3, ellipse=bb)
```



Podemos ver que las elipses no se solapan, lo que indica que existen diferencias significativas entre los datos del lote A y el lote B. Esto significa, por tanto; que hay efecto lote.

Capítulo 2

Descomposición en Valores Singulares

La Descomposición en Valores Singulares (DVS) es un método algebraico que permite descomponer una matriz rectangular como producto de otras tres matrices. Es una de las aproximaciones que se utiliza en Análisis Multivariante cuando la dimensión del conjunto de variables con el que se trabaja supone un problema.
Por consiguiente, puede decirse que es un paso intermedio que se utiliza en multitud de técnicas estadísticas.

Así lo hemos utilizado en Análisis de Correspondencias, para descomponer el estadístico χ^2 y poder obtener las coordenadas principales y standard de los elementos de la tabla de contingencia.
Del mismo modo, este método se ha empleado en Análisis de Correlación Canónica para obtener las variables de correlación canónica a partir de la matriz $\mathbf{\Sigma_{XX}}^{-1/2}\mathbf{\Sigma_{XY}}\mathbf{\Sigma_{YY}}^{-1/2}$.

2.1. Descomposición

Comenzaremos definiendo la descomposición en valores singulares de una matriz $\mathbf{X}$.

Definición 2.1.1 *Sea* $\mathbf{X}$ *una matriz de orden* $n \times p$ *y rango* $p \leq n$*, entonces la descomposición en valores singulares de* $\mathbf{X}$ *es:*

$$\mathbf{X} = \mathbf{UDV}'$$

donde $\mathbf{U}$ *y* $\mathbf{V}$ *son matrices de dimensión* $n \times p$ *y* $p \times p$*; respectivamente, y* $\mathbf{D}$ *es una matriz de dimensión* $p \times p$ *conteniendo los valores singulares de* $\mathbf{X}$*.*

En $\mathbf{U}$ se recogen los denominados vectores singulares izquierdos de $\mathbf{X}$ y en $\mathbf{V}$ los vectores singulares derechos. Ambas matrices son ortogonales, es decir:

$$\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}$$

$$\mathbf{V}'\mathbf{V} = \mathbf{V}\mathbf{V}' = \mathbf{I}$$

Esta ortogonalidad que caracteriza a las matrices $\mathbf{U}$ y $\mathbf{V}$ nos será de gran utilidad para explicar otras propiedades de la descomposición en valores singulares para una matriz rectangular.

La matriz $\mathbf{D}$ se define como:

$$\mathbf{D} = \begin{pmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_p \end{pmatrix}, \qquad \text{siendo } \alpha_1 \geq \alpha_2 \cdots \geq \alpha_p \text{ los valores singulares de } \mathbf{X}.$$

2.1.1. Propiedades de la descomposición

En el siguiente Lema se recogen las principales propiedades de las matrices que resultan de la descomposición. Nos serán de gran utilidad para las secciones siguientes.

Lema 2.1.1 *En las condiciones anteriores, se tiene que:*

a) $\mathbf{V}_{p\times p}$ *es la matriz de autovectores de* $\mathbf{X}'\mathbf{X}$

b) $\mathbf{U}_{n\times p}$ *es la matriz de autovectores de* $\mathbf{X}\mathbf{X}'$

c) *Los cuadrados de los valores singulares de* $\mathbf{X}$, $\alpha_1^2, \alpha_2^2, \cdots, {\alpha_p}^2$, *son los autovalores de* $\mathbf{X}'\mathbf{X}$ *y* $\mathbf{X}\mathbf{X}'$.

d) *Si* $\mathbf{X}$ *es simétrica y n=p, las matrices* $\mathbf{U}$ *y* $\mathbf{V}$ *son iguales.*

Demostración:

a) *Sea* $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$ *la descomposición en valores singulares de* $\mathbf{X}$. *Por la ortogonalidad de la matriz* $\mathbf{U}$, *podemos expresar el producto* $\mathbf{X}'\mathbf{X}$ *como:*

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{U}\mathbf{D}\mathbf{V}' = \mathbf{V}\mathbf{D}^2\mathbf{V}'$$

Multiplicando a derecha en ambos lados de la igualdad por $\mathbf{V}$ *y por la ortogonalidad de esta matriz, obtenemos:*

$$\mathbf{X}'\mathbf{X}\mathbf{V} = \mathbf{V}\mathbf{D}^2$$

Esta expresión satisface la definición de autovector para $\mathbf{X}'\mathbf{X}$*, con* $\mathbf{V}$ *el conjunto de autovectores de esta matriz y* $\mathbf{D}^2$ *sus autovalores.*

Luego queda probado que $\mathbf{V}$ *es el conjunto de autovectores de* $\mathbf{X}'\mathbf{X}$*.*

b) *Análoga al caso anterior.*

c) *Siguiendo el razonamiento anterior, las matrices* $\mathbf{X}'\mathbf{X}$ *y* $\mathbf{X}\mathbf{X}'$ *pueden expresarse como:*

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{U}\mathbf{D}\mathbf{V}' = \mathbf{V}\mathbf{D}^2\mathbf{V}'$$

$$\mathbf{X}\mathbf{X}' = \mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{V}\mathbf{D}\mathbf{U}' = \mathbf{U}\mathbf{D}^2\mathbf{U}'$$

Teniendo de nuevo en cuenta la definición de autovector, es sencillo ver que $\mathbf{D}^2$ *es la matriz de autovalores para ambas matrices.*

Por tanto, si denotamos por $\lambda_1, \ldots, \lambda_p$ *los autovalores de* $\mathbf{X}\mathbf{X}'$ *y* $\mathbf{X}'\mathbf{X}$*, puede verse que los valores singulares de* $\mathbf{X}$ *recogidos en la matriz* $\mathbf{D}$ *son equivalentes la raíz de estos autovalores. Es decir,* $\alpha_j = +\sqrt{\lambda_j}$*, con* $j = 1, \ldots, p$*.*

d) $\mathbf{X}$ *es simétrica si y sólo si* $\mathbf{X}' = \mathbf{X}$

Si es así, también se cumplirá lo siguiente:

$$\mathbf{X}'\mathbf{X} = \mathbf{X}\mathbf{X}' \Leftrightarrow \mathbf{V}\mathbf{D}^2\mathbf{V}' = \mathbf{U}\mathbf{D}^2\mathbf{U}'$$

Queda probado por tanto que en este caso los vectores singulares izquierdos y derechos para $\mathbf{X}$ *son los mismos.*

□

En el siguiente Corolario se especifica la relación entre la Descomposición en Valores Singulares y el Análisis de Componentes Principales.

Corolario 2.1.1 a) *Las Componentes Principales de* $\mathbf{X}'\mathbf{X}$ *están relacionadas con los autovectores de* $\mathbf{X}\mathbf{X}'$

b) *Las Componentes Principales de* $\mathbf{X}\mathbf{X}'$ *están relacionadas con los autovectores de* $\mathbf{X}'\mathbf{X}$.

Demostración:

a) *Las componentes principales de* $\mathbf{X}'\mathbf{X}$ *están dadas por* $\mathbf{Z} = \mathbf{X}\mathbf{V}$*, pues como hemos visto en el primer capítulo de la memoria, las combinaciones lineales de máxima varianza de las variables originales se obtienen con los autovectores y* $\mathbf{V}$ *es el conjunto de autovectores de* $\mathbf{X}'\mathbf{X}$.

Volviendo a hacer uso de la ortogonalidad de $\mathbf{V}$*, podemos expresar* $\mathbf{Z}$ *como:*

$$\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{V} = \mathbf{U}\mathbf{D}$$

Estas CP son, por tanto, una versión reescalada de $\mathbf{U}$*; vectores singulares izquierdos de* $\mathbf{X}$ *y autovectores de* $\mathbf{X}\mathbf{X}$*'.*

b) *Análoga al caso anterior.*

□

2.1.2. Aproximación matricial

En este apartado explicaremos cómo aproximar la matriz inicial por una de menor dimensión sin perder información relevante.

Si queremos trabajar con una matriz $\widehat{\mathbf{X}}$ aproximación de menor rango de la matriz $\mathbf{X}$, podemos calcularla mediante la aproximación de mínimos cuadrados.

El problema a resolver sería:

$$min(tr[(\mathbf{X} - \widehat{\mathbf{X}})(\mathbf{X} - \widehat{\mathbf{X}})'])$$

La matriz que buscamos es $\widehat{\mathbf{X}} = \mathbf{U}\mathbf{D}_r\mathbf{V}' = \sum_{j=1}^{r} \alpha_j \underline{u}_j \underline{v}_j'$, siendo $\mathbf{D}_r$ la matriz diagonal con los $r \leq p$ primeros valores singulares de $\mathbf{X}$.

Por las propiedades de la descomposición especificadas con anterioridad, la solución al problema puede expresarse también como $\widehat{\mathbf{X}} = \mathbf{Z}_r\mathbf{V}_r'$, con $\mathbf{Z}_r$ las r primeras componentes principales de $\mathbf{X}'\mathbf{X}$ y $\mathbf{V}_r'$ sus r primeros autovectores.

2.2. Biplots y matriz de aproximación

En este apartado, mostramos cómo la aproximación de la matriz $\mathbf{X}$ puede representarse en un biplot.
Recordemos que un biplot es una representación gráfica que combina filas y columnas de una matriz de datos, pudiendo así interpretar la distancia entre los mismos y obtener conclusiones.

Sea $\hat{\mathbf{X}} = \mathbf{U}_r\mathbf{D}_r\mathbf{V}_r'$, las columnas de $\mathbf{V}_r$ dan información sobre las columnas o variables, mientras que las columnas de $\mathbf{U}_r$ dan información sobre las filas o muestras de $\mathbf{X}$.

Esto se debe a la relación entre componentes principales y autovectores de una matriz que vimos en la sección anterior.
Si $\mathbf{Z} = \mathbf{UD}$ son las componentes principales de la matriz $\mathbf{X}'\mathbf{X}$, entonces la matriz $\mathbf{X}$ se expresará como:

$$\mathbf{X} = \mathbf{UDV}' = \mathbf{ZV}'$$

y por tanto podremos obtener información sobre las filas de la matriz $\mathbf{X}$.
Del mismo modo, si $\mathbf{W} = \mathbf{VD}$ son las componentes principales de la matriz $\mathbf{XX}'$, entonces a partir de $\mathbf{X}' = \mathbf{WU}'$ podremos obtener información sobre las columnas de $\mathbf{X}$.

Cada componente principal definirá un eje y los inviduos serán puntos dispersos.

Para poder visualizar el problema, normalmente sólo se consideran dos CP.

Capítulo 3

Análisis de Correspondencias

El Análisis de Correspondencias (AC) es una técnica iniciada por J.P. Benzécri en 1930 que analiza las asociaciones entre filas y columnas de una tabla de contingencia. Ésta se define como una tabla con dos entradas donde se muestran las frecuencias de cada una de las clases de dos variables cualitativas. En general, consideraremos tablas de dimensión $(n \times p)$.

El Análisis de Correspondencias tiene como objetivo principal desarrollar índices simples que muestren las relaciones entre las categorías de las filas y las columnas, indicando qué categorías columnas tienen más peso en una categoría fila y viceversa.
A su vez, es útil para reducir la dimensión de la tabla de contingencia. Para ello, se calcularán los índices en orden decreciente de importancia, de manera que podamos resumir esta tabla sin perder información relevante. Si sólo usamos dos índices, podremos mostrar los resultados en gráficos de dos dimensiones.

3.1. El método

Este método fue desarrollado para variables cualitativas exclusivamente.
Cada entrada de la tabla es el número de observaciones de la muestra que se corresponde simultáneamente con la i-ésima categoría fila y la j-ésima categoría columna (para $i = 1, ..., n$ y $j = 1, ..., p$).

Una vez que tenemos establecido el tipo de variables que medimos, podemos plantearnos dos tipos de análisis para los datos: homogeneidad de poblaciones o independencia de caracteres.

•Homogeneidad de poblaciones

Se considera una variable categórica B con p modalidades.
El problema se estudia para n poblaciones, siendo la hipótesis de homogeneidad:

$$H_0 : P_1(B_j) = \cdots = P_n(B_j), \qquad j = 1, \cdots, p$$

Seleccionamos una muestra N_i de cada población y construimos la tabla de contingencia.

Definición 3.1.1 *Sea N_{ij} el número de elementos de la muestra tomada de la población P_i que presentan la categoría B_j. Se tiene que $N_{ij} \sim Bi(N_i, P_i(B_j))$.*

Se construye la tabla de contingencia como:

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

donde $N_i = N_{i\cdot} = \sum_{j=1}^{p} N_{ij}$ y $N_{\cdot j} = \sum_{i=1}^{n} N_{ij}$

•Independencia de caracteres

En este problema se consideran dos variables categóricas A y B con n modalidades y p modalidades respectivamente, siendo la hipótesis de independencia:

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

Para definir la tabla de contingencia, seleccionamos una muestra de tamaño N y observamos (A, B) en cada elemento.

Definición 3.1.2 *Sea N_{ij} el número de elementos de la muestra que presentan $A_i \cap B_j$. Se tiene que $N_{ij} \sim Bi(N, P(A_i \cap B_j))$ y se define la tabla de contingencia como:*

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

donde $N_{i\cdot} = \sum_{j=1}^{p} N_{ij}$ *y* $N_{\cdot j} = \sum_{i=1}^{n} N_{ij}$

A continuación, haremos uso del concepto de tabla de contingencia para obtener la que se conoce como tabla de frecuencias relativas, matriz con la que trabajamos y que es equivalente para los dos tipos de estudio.
Se mostrarán las frecuencias relativas conjuntas y marginales.

Definición 3.1.3 *Sea $f_{ij} = \frac{N_{ij}}{N}$, se puede definir la siguiente tabla de frecuencias relativas:*

$$\mathbf{F} = \begin{pmatrix} f_{11} & \cdots & f_{1j} & \cdots & f_{1p} \\ \vdots & & \vdots & & \vdots \\ f_{i1} & \cdots & f_{ij} & \cdots & f_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ f_{n1} & \cdots & f_{nj} & \cdots & f_{np} \end{pmatrix} \quad \begin{matrix} f_{1.} \\ \vdots \\ f_{i.} = \sum_{j=1}^{p} f_{ij} \\ \vdots \\ f_{n.} \end{matrix}$$

$$\begin{matrix} f_{.1} & \cdots & f_{.j} & \cdots & f_{.p} \end{matrix}$$

Estas frecuencias marginales filas $\begin{pmatrix} f_{1.} \dots f_{i.} \dots f_{n.} \end{pmatrix}$' y columnas $\begin{pmatrix} f_{.1} \cdots & f_{.j} \cdots & f_{.p} \end{pmatrix}$ también se conocen como pesos fila y columna, respectivamente.

Como comentamos en la introducción al capítulo, el objetivo de esta técnica es definir la relación entre filas y columnas de la tabla mediante unos índices. ¿Cómo se miden estas relaciones entre categorías filas y columnas?.

Podemos estudiarlas gráficamente, representando las categorías e interpretando las posiciones relativas de los puntos en función de los pesos correspondientes a cada fila o columna. Más adelante en este capítulo veremos cómo interpretar las distancias entre elementos del

gráfico.

Obtendremos las coordenadas para la representación de cada clase gracias a un sistema de índices similares al ACP, pero descomponiendo el valor del estadístico "chi-cuadrado" en lugar de la varianza total.

En la siguiente sección se mostrará cómo obtener este estadístico y su descomposición, además de algunas propiedades.

3.2. Descomposición Chi-Cuadrado

Para calcular el valor del estadístico χ^2, en primer lugar se estima el valor esperado de cada elemento de la tabla de contingencia E_{ij}, para luego comparar estos valores con los valores observados correspondientes N_{ij}.

El valor esperado de cada elemento N_{ij} dependerá del tipo de análisis de los datos.

• Homogeneidad de poblaciones:

$\hat{E}_{H_0}(N_{ij}) = N_{i\cdot}P(B_j) = \frac{N_{i\cdot}N_{\cdot j}}{N}$

• Independencia de caracteres:

$\hat{E}_{H_0}(N_{ij}) = NP(A_i \cap B_j) = N(\frac{N_{i\cdot}}{N})(\frac{N_{\cdot j}}{N}) = \frac{N_{i\cdot}N_{\cdot j}}{N}$

Hemos comprobado que el valor de E_{ij} es el mismo para ambos casos. Así, el estadístico propuesto es:

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(N_{ij} - E_{ij})^2}{E_{ij}} \tag{3.1}$$

siendo $E_{ij} = \hat{E}_{H_0}(N_{ij}) = \frac{N_{i\cdot}N_{\cdot j}}{N}$.

La razón por la que se divide por E_{ij} en la expresión del estadístico es porque ha de considerarse la distinta precisión de cada coordenada fila o columna. Debemos tener en cuenta que cada fila o columna tiene un peso distinto.

3.2.1. Descomposición y propiedades

Para calcular los índices que permiten interpretar las relaciones entre clases, comenzaremos definiendo una matriz a partir del estadístico χ^2.

Sea $\mathbf{C}$ una matriz de elementos $c_{ij} = \frac{(N_{ij}-E_{ij})}{E_{ij}^{1/2}}$, calcularemos su descomposición en valores singulares.

Esta descomposición se define como

$$\mathbf{C} = \mathbf{\Gamma D \Delta}',$$

siendo $\mathbf{\Gamma}_{(n\times p)}$ la matriz de autovectores de $\mathbf{CC}'$, $\mathbf{\Delta}_{(p\times p)}$ los de $\mathbf{C}'\mathbf{C}$ y $\mathbf{D}_{(p\times p)}$ la matriz diagonal de valores singulares $\alpha_j, \quad j \in 1, \ldots, p$.

Estos p valores singulares cumplen $\alpha_j = +\sqrt{\lambda_j}$, con λ_j los autovalores de $\mathbf{CC}'$ y $\mathbf{C}'\mathbf{C}$, como vimos en el Lema 2.1.1 del segundo capítulo de esta memoria.

Tras la descomposición de la matriz $\mathbf{C}$, podemos deducir:

$$c_{ij} = \sum_{k=1}^{R} \lambda_k^{1/2} \gamma_{ik} \delta_{jk} \tag{3.2}$$

con γ_{ik} elemento de la matriz $\mathbf{\Gamma}$, δ_{jk} elemento de $\mathbf{\Delta}$ y $R = rango(\mathbf{C}) \leq \text{mín}\{(n-1),(p-1)\}$.

Por la normalidad de los autovectores contenidos en $\mathbf{\Gamma}$ y $\mathbf{\Delta}$ y la ecuación (3.2), se satisface la siguiente igualdad:

$$tr(\mathbf{CC}') = \sum_{k=1}^{R} \lambda_k = \sum_{i=1}^{n}\sum_{j=1}^{p} c_{ij}^2 = \chi^2$$

Se demuestra así que el estadístico χ^2 definido en (3.1) puede expresarse como la suma de los R autovalores no nulos de la matriz $\mathbf{CC}'$.

Denotemos por $\mathbf{A} = diag(N_{1\cdot}, \ldots, N_{n\cdot})$ y $\mathbf{B} = diag(N_{\cdot 1}, \ldots, N_{\cdot p})$. Es sencillo comprobar que se cumple

$$\begin{cases} \mathbf{C}\sqrt{(N_{\cdot 1}, ..., N_{\cdot p})'} = \mathbf{C}\sqrt{\mathbf{B}1_p} = \underline{0} \\ \mathbf{C}'\sqrt{(N_{1\cdot}, ..., N_{n\cdot})'} = \mathbf{C}\sqrt{\mathbf{A}1_n} = \underline{0} \end{cases}$$

Una vez descompuesto el estadístico χ^2, los autovectores $\underline{\delta}_k$ y $\underline{\gamma}_k$ son objeto de interés para calcular los índices que permiten analizar la correspondencia entre filas y columnas. Por las propiedades de la descomposición en valores singulares, sabemos que los autovectores $\underline{\gamma}_k$ proporcionan información sobre las filas de la tabla de contingencia y $\underline{\delta}_k$ sobre las columnas.
A partir de esto, obtenemos el siguiente resultado.

Proposición 3.2.1 *Las relaciones de dualidad entre el espacio fila y columna (para $k = 1, ..., R$) están dadas por:*

$$\begin{cases} \underline{\delta}_k = \frac{1}{\sqrt{\lambda_k}} \mathbf{C}' \underline{\gamma}_k, \\ \underline{\gamma}_k = \frac{1}{\sqrt{\lambda_k}} \mathbf{C} \underline{\delta}_k \end{cases}$$

mientras que las proyecciones de las filas y las columnas de $\mathbf{C}$ están dadas por:

$$\begin{cases} \mathbf{C} \underline{\delta}_k = \sqrt{\lambda_k} \underline{\gamma}_k, \\ \mathbf{C}' \underline{\gamma}_k = \sqrt{\lambda_k} \underline{\delta}_k \end{cases}$$

Las aproximaciones de χ^2 que se propongan dependerán de la información que podamos recoger con los mayores autovalores.

Si suponemos que el primer autovalor es dominante, eso significa que podemos hacer la aproximación $c_{ij} \approx \lambda_1^{1/2} \gamma_{i1} \delta_{j1}$; de manera que si δ_{j1} y γ_{i1} toman valores grandes y del mismo signo; esto indica una asociación positiva entre las categorías fila i y columna j. Si por el contrario toman valores de signo contrario entonces la asociación será negativa.

3.3. Coordenadas Principales y Coordenadas Standard

Hay dos posibles representaciones de las entradas de la tabla de contingencia, en función de si asumimos o no que existe relación entre los datos.

Cuando partimos de que filas y columnas en la tabla están relacionadas, representaremos estas filas y columnas utilizando las llamadas coordenadas principales. La solución que obtenemos se conoce como una solución simétrica del problema.

En el caso contrario, usaremos las denominadas coordenadas standard para las columnas, obteniéndose así lo que se conoce como una solución asimétrica.

Comenzaremos definiendo las Coordenadas Principales.

Definición 3.3.1 *Sean* $\underline{r}_{k_{(n\times 1)}}$ *las proyecciones de* $\mathbf{A}^{-1/2}\mathbf{C}$ *en* $\underline{\delta}_k$ *y* $\underline{s}_{k_{(p\times 1)}}$ *las proyecciones de* $\mathbf{B}^{-1/2}\mathbf{C}'$ *en* $\underline{\gamma}_k$.
Los vectores $\underline{r}_k$ *y* $\underline{s}_k$ *se conocen como coordenadas principales de filas y columnas, respectivamente.*

$$\begin{cases} \underline{r}_k = \mathbf{A}^{-1/2}\mathbf{C}\underline{\delta}_k = \sqrt{\lambda_k}\mathbf{A}^{-1/2}\underline{\gamma}_k \\ \underline{s}_k = \mathbf{B}^{-1/2}\mathbf{C}'\underline{\gamma}_k = \sqrt{\lambda_k}\mathbf{B}^{-1/2}\underline{\delta}_k \end{cases}$$

con $k = 1, \ldots, R$.

Proposición 3.3.1 *Estos vectores satisfacen las siguientes igualdades:*

$$\begin{cases} \underline{r}_k'(N_{1.}, ..., N_{n.})' = 0 \\ \underline{s}_k'(N_{.1}, ..., N_{.p})' = 0 \end{cases}$$

A partir de las ecuaciones que definen las coordenadas principales $\underline{r}_k$ y $\underline{s}_k$, sabemos que:

$$\underline{r}_k'\mathbf{A}\underline{r}_k = \lambda_k, \quad \underline{s}_k'\mathbf{B}\underline{s}_k = \lambda_k$$

De la relación de dualidad entre los autovectores de la Descomposición singular obtenemos:

$$\underline{r}_k = \frac{1}{\sqrt{\lambda_k}}\mathbf{A^{-1/2}CB^{1/2}}\underline{s}_k = \sqrt{\frac{N}{\lambda_k}}\mathbf{A^{-1}X}\underline{s}_k$$

$$\underline{s}_k = \frac{1}{\sqrt{\lambda_k}}\mathbf{B^{-1/2}C'A^{1/2}}r_k = \sqrt{\frac{N}{\lambda_k}}\mathbf{B^{-1}X'}\underline{r}_k$$

Proposición 3.3.2 *Se satisfacen las relaciones:*

$$\underline{r}_i = \frac{1}{\lambda_i}\sum_{j=1}^{p}\underline{s}_j\frac{N_{ij}}{N_{i.}}$$

$$\underline{s}_j = \frac{1}{\lambda_j}\sum_{i=1}^{n}\underline{r}_i\frac{N_{ij}}{N_{.j}}$$

En la siguiente proposición se detallan la media y varianza de las Coordenadas Principales.

Proposición 3.3.3 *La media de los vectores* $\underline{r}_k$ *y* $\underline{s}_k$ *es:*

$$\overline{\underline{r}}_k = \frac{1}{N}\underline{r}_k'(N_{1.}, ..., N_{n.})' = 0$$

$$\overline{\underline{s}}_k = \frac{1}{N}\underline{s}_k'(N_{.1}, ..., N_{.p})' = 0$$

La varianza de estos es:

$$Var(\underline{r}_k) = \frac{1}{N}\sum_{i=1}^{n} N_{i.} r_{ki}^2 = \frac{r_k' \mathbf{A} \underline{r}_k}{N} = \frac{\lambda_k}{N}, \tag{3.5}$$

$$Var(\underline{s}_k) = \frac{1}{N}\sum_{j=1}^{p} N_{.j} s_{kj}^2 = \frac{s_k' \mathbf{B} \underline{s}_k}{N} = \frac{\lambda_k}{N} \tag{3.6}$$

Para la solución asimétrica del problema necesitamos definir las Coordenadas Standard.

Definición 3.3.2 *Se definen las Coordenadas Standard como:*

$$\begin{cases} \underline{r}_{ks} = \mathbf{A}^{-1/2} \underline{\gamma}_k \\ \underline{s}_{ks} = \mathbf{B}^{-1/2} \underline{\delta}_k \end{cases}$$

Por último, haremos uso de las contribuciones para evaluar el peso de cada fila (o columna) en las varianzas de los factores dadas en (3.5) y (3.6).

Definición 3.3.3 *La contribución de la fila i a la varianza del factor $\underline{r}_k$ se define como:*

$$C_a(i, \underline{r}_k) = \frac{N_{i.} r_{ki}^2}{\lambda_k}$$

mientras que la contribución de la columna j a la varianza del factor $\underline{s}_k$ es:

$$C_a(j, \underline{s}_k) = \frac{N_{.j} s_{kj}^2}{\lambda_k}$$

3.4. Análisis de Correspondencias en la Práctica

Como hemos mencionado anteriormente, la representación gráfica en los k ejes de las n filas y las p columnas de $\mathbf{X}$ viene dada por los elementos $\underline{r}_k$ y $\underline{s}_k$; respectivamente.

Las coordenadas filas y columnas pueden ser representados en un gráfico bidimensional. Si están representados próximos el uno al otro (lejanos al origen), esto indicaría que la i-ésima categoría fila tiene una alta frecuencia condicional $N_{ij}/N_{.j}$ en la expresión de la j-ésima categoría columna, y la j-ésima categoría columna tiene una alta frecuencia condicional $N_{ij}/N_{i.}$ en la expresión de la i-ésima categoría fila; lo que indicaría una asociación positiva entre ambas.

Razonando del mismo modo, si ambas categorías estuvieran representadas muy lejos la una de la otra, la asociación entre ellas sería negativa y por tanto la contribución condicional

de frecuencia sería muy pequeña.

Normalmente, un gráfico bidimensional explica satisfactoriamente los datos.

Dependiendo del gráfico, la interpretación de los datos se puede resumir como sigue:

• Gráfico para filas de la tabla de contingencia: Se representan los elementos en los ejes $\underline{r}_1$ y $\underline{r}_2$. La proximidad entre clases de una misma variable categórica indica frecuencias similares.

• Gráfico para columnas de la tabla de contingencia: Se representan los elementos en los ejes $\underline{s}_1$ y $\underline{s}_2$. La interpretación de este gráfico es similar al anterior. Proximidad entre clases de una misma variable categórica indica frecuencias similares.

• Gráfico conjunto: La proximidad entre filas y columnas indica el peso que tiene un elemento de la matriz sobre el otro. Una fila que está bastante distante de una columna particular indica que no hay casi observaciones en la columna para esta fila (y viceversa). Filas y columnas con fuerte asociación se proyectan en la misma dirección desde el origen. El origen es la media de $\underline{r}_k$ y $\underline{s}_k$.

Todas las interpretaciones anteriores dependerán de la calidad de la representación gráfica que se evalua, como en ACP, usando el porcentaje de varianza acumulada.

3.5. Aplicación práctica del AC

Esta es una aplicación del Análisis de Correspondencias simétrico, por lo que asumiremos que existe relación entre filas y columnas y calcularemos las coordenadas principales para ambas.

Nos basamos en el porcentaje de variabilidad explicada para elegir el número k de ejes que recogen la mayor cantidad de información posible.

Datos

Se introducen los datos que usaremos como ilustración de la técnica expuesta.

La base de datos `khan` está disponible en el paquete `made4`. El conjunto de datos `Khan` contiene 2308 perfiles de expresión génica (individuos) y valores de expresión para 6567 clones, de los cuales 3789 eran genes conocidos y 2778 eran EST (“expressed sequence

tag"); que se introducen para estudiar la expresión de genes en los 4 tipos de tumores de células pequeñas azules de la infancia:

a) Neuroblastoma (NB)

b) Rabdomiosarcoma (RMS)

c) Un conjunto de linfomas No Hodgkin, llamado Linfoma Burkitt (BL)

d) Familia de tumores Ewing (EWS)

En este caso, el subconjunto de datos que analizamos está formando por 306 genes para muestras de 64 pacientes.

Paquetes

El paquete *made4* facilita el análisis multivariante de datos de microarrays de expresión génica. Proporciona un conjunto de funciones que utilizan y extienden funciones gráficas y estadísticas multivariantes disponibles en *ade4*. Además, acepta datos de expresión génica en una amplia variedad de formatos de entrada, incluyendo los de Bioconductor, AffyBatch, ExpressionSet, marrayRaw y data.frame o matrix.
made4 requiere instalar *ade4* y además *scatterplot3d*. El paquete se encuentra en bioconductor: "http://bioconductor.org/biocLite.R"

En primer lugar realizamos un análisis exploratorio de los datos y posteriormente un Análisis de Correspondencias Simétrico.

3.5.1. Exploración de los datos

Comenzamos cargando los paquetes necesarios para ejecutar el Análisis de Correspondencias.

```
library(ade4)
library(made4)
```

Este subconjunto del conjunto de datos original es una lista con 7 elementos. Con la función `names`, vemos los nombres de cada uno de ellos.

```
data(khan)
class(khan)

## [1] "list"

names(khan)
```

```
## [1] "train"                 "test"
## [3] "train.classes"         "test.classes"
## [5] "annotation"            "gene.labels.imagesID"
## [7] "cellType"
```

Veamos con detenimiendo cada uno de estos elementos.

a) "khan$train" es un data.frame de dimensión 306×64, lo que significa que hay 64 muestras para el conjunto de entrenamiento.

b) "khan$test" es un data.frame de dimensión 306×25, lo que indica que hay 25 muestras para el conjunto test.

c) "khan$train.classes" es un factor de longitud 64 con diferentes niveles, mostrando los distintos tipos de cáncer que presentan los individuos del conjunto de entrenamiento.

d) "khan$test.classes" es un factor de longitud 25 con diferentes niveles que indican los tipos de cáncer de cada uno de los individuos del conjunto test.

e) "khan$annotation" es un data.frame (306×8) con 8 anotaciones distintas para los 306 genes que estudiaremos.

f) "khan$gene.labels.imagesID" es un objeto de tipo carácter de longitud 306 con información sobre imágenes etiquetadas de los genes que estudiaremos.

g) "khan$cellType" es un objeto de tipo carácter de longitud 64 con información sobre el tipo de célula de cada elemento del conjunto de entrenamiento. Veremos de qué se trata esto más adelante.

```
summary(khan)
```

```
##                      Length Class      Mode
## train                 64    data.frame list
## test                  25    data.frame list
## train.classes         64    factor     numeric
## test.classes          25    factor     numeric
## annotation             8    data.frame list
## gene.labels.imagesID 306    -none-     character
## cellType              64    -none-     character
```

Como hemos comentado anteriormente, trabajamos con conjunto entrenamiento y test en los que medimos 306 genes, de manera que los individuos de cada conjunto son distintos y en ningún momento se solapan.

Ahora, veamos "khan$cellType" con más profundidad.

```
khan$cellType
```

```
[1] "T" "T" "T" "T" "T" "T" "T" "T" "T" "T" "T" "T" "T" "C"
[15]"C" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C"
[29]"C" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C"
[43]"C" "C" "C" "C" "C" "C" "C" "C" "C" "C" "C" "T" "T" "T"
[57]"T" "T" "T" "T" "T" "T" "T" "T"
```

Al ejecutar esta orden, aparecen los tipos de células de las 64 muestras del conjunto de entrenamiento. “T” indica que la muestra ha sido obtenida por una biopsia del tumor, mientras que “C” indica que se ha obtenido a partir de una línea celular.
Es fácil comprobar que en el estudio de los tumores de tipo “RMS” y “EWS” se han usado ambos métodos, mientas que en los otros dos la muestra se ha obtenido a partir de la línea celular.

En las siguientes instrucciones hemos especificado que sólo trabajaremos con el conjunto de entrenamiento y que “k.class” son los tipos de cáncer que presentan las 64 muestras.

```
k.data<-khan$train
dim(k.data)

## [1] 306  64

k.class<-khan$train.classes
```

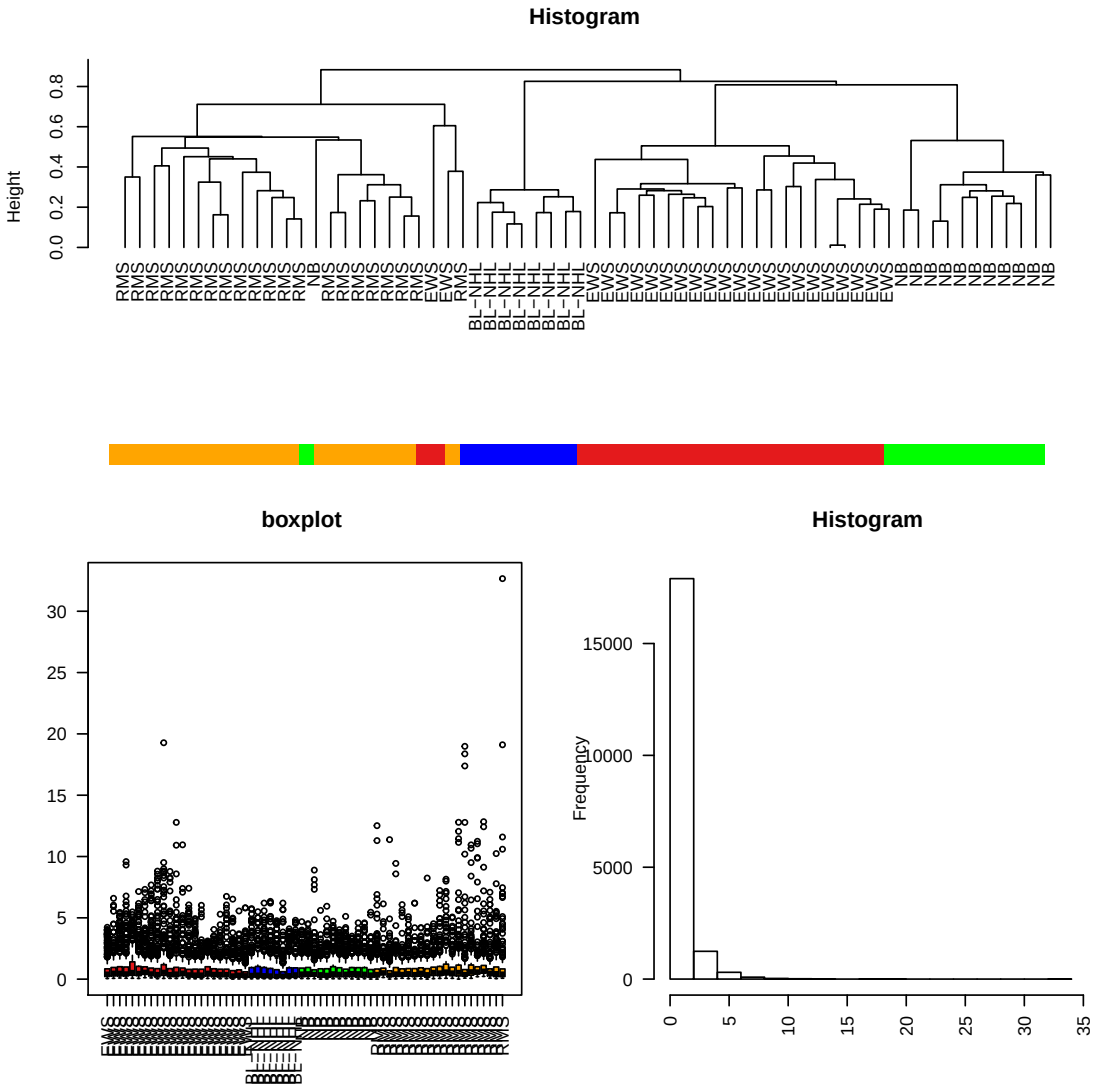
La función `overview()` da una visión general de los datos. Dibuja un diagrama de cajas, un histograma y un dendograma de análisis jerárquico.

El dendograma de análisis jerárquico es una herramienta de análisis cluster, método que permite construir grupos a partir de objetos multivariantes. Para ello, se construye una matriz que contenga las medidas de similitud; aunque cabe destacar que normalmente las variables categóricas conducen a valores próximos mientras que las continuas dan lugar a valores distantes.

Este análisis se puede plantear o bien para clasificar filas o bien para clasificar columnas de la tabla de contingencia. En este caso se clasifican columnas, que contienen las 64 muestras. Muestras que a priori son indistinguibles pueden tener alguna subclase que gracias a este análisis podemos identificar.

Se lleva a cabo un clustering jerárquico aglomerativo, donde se empieza por la partición más fina posible y se va agrupando. Se usa la correlación de Pearson como medida de similitud.

```
overview(k.data,classvec=k.class, labels=k.class)
```



El dendograma representa las clases de cáncer de cada individuo, agrupadas por similitud. Se establecen dos grupos, uno de ellos conteniendo en su mayor parte individuos con el tipo de tumor “RMS” y el otro conteniendo los 3 tipos restantes.

El diagrama de cajas y bigotes representa los valores que toman las variables para cada individuo en función del primer y tercer cuantil.

El histograma representa en el eje de abcisas el valor que pueden tomar las variables y en el eje de ordenadas la frecuencia con que se toma este valor para los 64 individuos.

3.5.2. Análisis de Correspondencias

Procedemos al Análisis de Correspondencias de los datos. Emplearemos la función `ord` del paquete *made4* para especificar que ha de llevarse a cabo este análisis y el argumento "type=coa" para indicar que es un análisis de tipo simétrico.

```
k.coa<-ord(k.data, type="coa")
```

Analicemos los resultados de esta instrucción más detenidamente.

```
k.coa$ord

## Duality diagramm
## class: coa dudi
## $call: dudi.coa(df = data.tr, scannf = FALSE, nf = ord.nf)
##
## $nf: 63 axis-components saved
## $rank: 63
## eigen values: 0.1713 0.1383 0.1032 0.05995 0.04965 ...
##   vector length mode    content
## 1 $cw    64     numeric column weights
## 2 $lw    306    numeric row weights
## 3 $eig   63     numeric eigen values
##
##   data.frame nrow ncol content
## 1 $tab       306  64   modified array
## 2 $li        306  63   row coordinates
## 3 $l1        306  63   row normed scores
## 4 $co        64   63   column coordinates
## 5 $c1        64   63   column normed scores
## other elements: N
```

Los resultados de ordenación contienen una lista de longitud 14. Esta lista incluye 63 autovalores (eig), el vector de pesos columnas de longitud 64 (cw) y el vector de pesos filas de longitud 306 (lw). Además, las Coordenadas Principales filas $\underline{r}_j$ (li) y columnas $\underline{s}_j$ (co), así como sus equivalentes normalizadas $(l1)$ y $(c1)$.

En la siguiente instrucción, hemos ejecutado los valores de las 3 primeras Coordenadas Principales filas y columnas para los dos primeros genes e individuos; respectivamente.

```
k.coa$ord$li[1:2,1:3]

##             Axis1       Axis2      Axis3
## 25725   0.7001400 -0.08693461 -0.1278178
## 193913 -0.2486831  0.64857220 -0.2381108

k.coa$ord$co[1:2,1:3]

##             Comp1      Comp2       Comp3
## EWS.T1 -0.3923287 -0.3255404 -0.02589332
## EWS.T2 -0.3461686 -0.2956390  0.05687684
```

La suma de los autovalores será equivalente al estadístico *chi-cuadrado* de la tabla de contingencia, tal y como vimos en el desarrollo teórico de la técnica.

Veamos el porcentaje de varianza explicada por cada eje j de los 63 que hay en total.

```
k.coa$ord$eig*100/sum(k.coa$ord$eig)

##  [1] 16.947461298 13.683297888 10.207338717  5.931017086
##  [5]  4.911896717  3.810337299  3.027912076  3.004793647
##  [9]  2.419837082  2.319143755  2.042227963  1.886343953
## [13]  1.841744206  1.695622185  1.585943486  1.437215463
## [17]  1.342418983  1.251040188  1.167146174  1.114077872
## [21]  1.076647536  0.989724516  0.919956542  0.865911339
## [25]  0.812460108  0.789851377  0.755875491  0.736710674
## [29]  0.694142462  0.652011530  0.614248981  0.582822698
## [33]  0.548383426  0.513845412  0.493681883  0.446536643
## [37]  0.440410135  0.425398824  0.416579264  0.398563871
## [41]  0.376038015  0.367570772  0.356010253  0.332082395
## [45]  0.327998743  0.318501534  0.295054657  0.283405148
## [49]  0.257741008  0.241029558  0.227885174  0.225137815
## [53]  0.214519150  0.203081716  0.179324545  0.172844643
## [57]  0.168069405  0.151730905  0.142298857  0.129021644
## [61]  0.115421971  0.106023909  0.008629433
```

Con `cumsum` calculamos la varianza acumulada por los primeros ejes:

```
head(cumsum(k.coa$ord$eig*100/sum(k.coa$ord$eig)))

## [1] 16.94746 30.63076 40.83810 46.76911 51.68101 55.49135
```

Por tanto, el 30.63076 % de la varianza está explicada por los dos primeros ejes.

Análisis de Correspondencias- Visualización de resultados

Hay muchas funciones en el paquete *made4* que permiten visualizar resultados de un análisis de ordenación. La forma más simple de hacerlo es usando la función `plot`. Representaremos los autovalores, variables y casos (microarrays).

• Gráfico de autovalores: en el gráfico superior izquierdo se muestran los autovalores. Sólo tendremos en cuenta aquellos hasta que comienza a estabilizarse la gráfica, en este caso los dos primeros.

• Gráfico para muestras: Se representan los individuos en los ejes $\underline{s}_1$ y $\underline{s}_2$, pues son las columnas de la matriz inicial. Perfiles próximos en la representación indican que existe similitud entre ellos.

El eje $\underline{s}_1$ distingue la familia de tumores de tipo RMS de los demás perfiles de expresión génica, mientras que el eje $\underline{s}_2$ separa las muestras pertenecientes al tipo Neuroblastoma (NB) y Linfoma Burkitt (BL) frente al resto.

• Gráfico para variables: Se representan los genes (filas de la matriz inicial) en los ejes $\underline{r}_1$ y $\underline{r}_2$. Genes próximos en la representación indican que existe similitud entre ellos.
Podemos observar los grupos de genes que separa cada una de los dos ejes.

• Gráfico conjunto: Se proyectan filas y columnas de la tabla de contingencia en el mismo gráfico. Genes y muestras con fuerte asociación se proyectan en la misma dirección desde el origen. Se revela así la relación existente entre los genes y los tipos de cáncer que se asocian.

```
plot(k.coa, classvec = k.class)
```

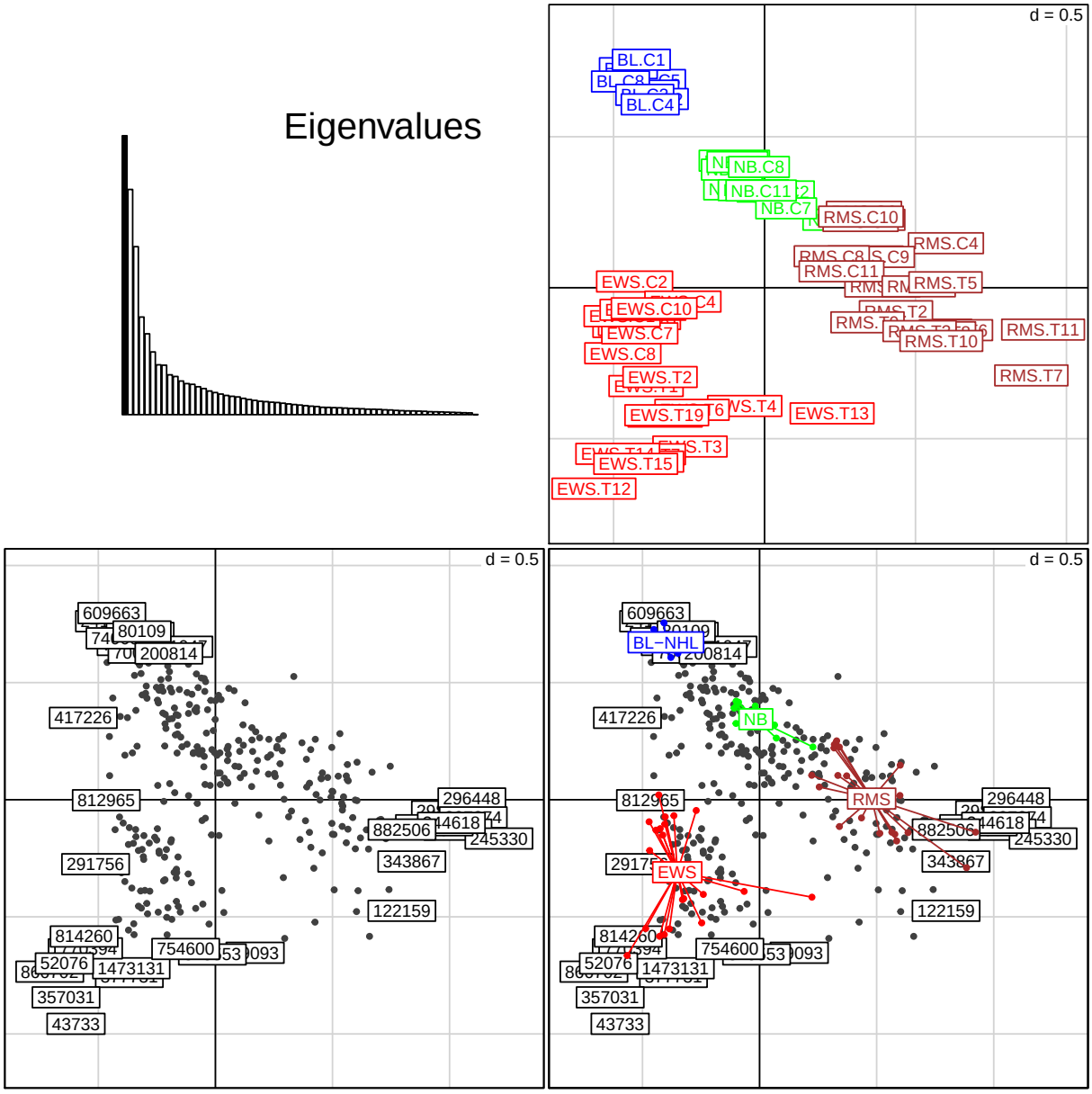


Figura 3.1: Plot del Análisis de Correspondencias.
A. autovalores
B. proyección de muestras microarrays de pacientes con varios tipos de tumores
C. proyección de genes
D.biplot mostrando genes y muestras

Capítulo 4

Análisis de Coinercia

El Análisis de Coinercia (ACoi) es una técnica de análisis multivariante que permite identificar tendencias en conjuntos de datos que contienen las mismas muestras, es decir, que se han recogido sobre los mismos individuos. A su vez, permite reducir la dimensión de los conjuntos más similares.

Esta herramienta no se limita al análisis de conjuntos que contienen el mismo número de variables y puede aplicarse en caso de que haya más variables que muestras.

El Análisis de Correspondencias es un paso previo y necesario para su aplicación.

4.1. El Método

Sean $\mathbf{X}_{n\times p}$ e $\mathbf{Y}_{n\times q}$ dos matrices de datos con los mismos n individuos a los que se miden p y q variables, respectivamente. Nos interesa estudiar la relación entre estos dos conjuntos de variables.

Como ya sabemos, la relación entre dos variables cuantitativas se mide con el coeficiente de correlación lineal ρ, que toma valores entre -1 y 1 dependiendo de la robustez y el sentido de la relación entre ellas.

Para estudiar la relación entre conjuntos de variables mediante el Análisis de Coinercia, necesitamos definir un índice conocido como $coinercia$.

Para definir este índice, necesitamos definir previamente una coestructura entre las tablas de datos iniciales con las que trabajamos. Pasaremos de las matrices de datos originales $\mathbf{X}$

e $\mathbf{Y}$ a considerar la matriz producto $\mathbf{X'Y}$ de dimensiones $p \times q$.
Esquemáticamente, tiene la siguiente forma:

$$\boxed{\mathbf{X}}\boxed{\mathbf{Y}} \rightarrow \boxed{\mathbf{X'Y}}$$

de manera que la matriz con la que trabajaremos tendrá dimensión $p \times q$ y por tanto sus filas recogerán las p variables que caracterizan a $\mathbf{X}$ y sus columnas las q variables que caracterizan a $\mathbf{Y}$.
También es posible trabajar con las q variables de $\mathbf{Y}$ como filas y las p variables de $\mathbf{X}$ como columnas de esta nueva matriz, que en ese caso tendrá dimensión $q \times p$.

El cálculo de la coinercia se basará en los tripletes estadísticos $(\mathbf{X}, \mathbf{Q_X}, \mathbf{D})$ e $(\mathbf{Y}, \mathbf{Q_Y}, \mathbf{D})$, siendo $\mathbf{Q_X}_{p\times p}$ una matriz diagonal representando los pesos de las p variables de $\mathbf{X}$ y $\mathbf{Q_Y}_{q\times q}$ los pesos de las q variables de $\mathbf{Y}$. Estas matrices diagonales representan por tanto los pesos de las columnas de las matrices iniciales. $\mathbf{D}_{n\times n}$ es una matriz diagonal que recoge los pesos de los n individuos (filas).

Nótese que hemos definido con la misma matriz $\mathbf{D}$ los pesos de las filas para ambas tablas de datos pues aunque se midan distintas variables se trabaja con los mismos individuos.

Cada triplete es resultado de un Análisis de Correspondencias, de forma que las matrices $\mathbf{X}$ e $\mathbf{Y}$ serán las tablas χ^2 derivadas de los conjuntos de datos originales.

Observación 4.1.1 *En las matrices $\mathbf{X}$ e $\mathbf{Y}$ se recogen los denominados residuos de Pearson obtenidos para cada casilla como:*

$$\frac{(N_{ij} - E_{ij})}{\sqrt{E_{ij}}}$$

siendo N_{ij} el valor observado en la tabla de contingencia y E_{ij} su valor esperado correspondiente, según se vio en la sección 3.2 del tercer capítulo.

A partir de estos elementos, podemos definir los conceptos de $inercia$ y $coinercia$.

La inercia es una medida de la variabilidad en los datos y se define como la distancia χ^2 entre un elemento y su perfil medio, teniendo en cuenta el peso de cada elemento.

Definición 4.1.1 *La expresión de la inercia en el contexto que nos ocupa es:*

$$\begin{cases} in_{\mathbf{X}} = traza(\mathbf{X}\mathbf{Q_X}\mathbf{X'}\mathbf{D}) \\ in_{\mathbf{Y}} = traza(\mathbf{Y}\mathbf{Q_Y}\mathbf{Y'}\mathbf{D}) \end{cases}$$

Definición 4.1.2 *Se define la coinercia entre dos espacios como:*

$$Coin(\mathbf{XY}) = traza(\mathbf{XQ_Y X'DYQ_Y Y'D})$$

Si los datos están centrados, la inercia será una suma de varianzas mientras que la coinercia será una suma de covarianzas al cuadrado.

A continuación, definiremos los pasos que se llevan a cabo para la identificación de tendencias.

Estas tablas de datos $\mathbf{X}$ e $\mathbf{Y}$ se representan en dos hiperespacios de dimensiones p y q, respectivamente.
Con un análisis previo, se maximiza la inercia (o variabilidad explicada) de ambos hiperespacios. Esto se lleva a cabo mediante un análisis de corespondencias de cada conjunto de datos. Como resultado se obtienen dos conjuntos de ejes, uno para cada conjunto de datos.

El primer par de ejes en el que se representarán las relaciones entre ambos conjuntos de variables se elige maximizando la coinercia, de forma que represente la tendencia más importante entre los conjuntos de datos.
El segundo par se obtiene con el mismo procedimiento, pero requiriendo también que sea ortogonal al primer par. Así se va repitiendo el proceso con todos los ejes.

¿Cómo podemos medir la similitud entre ordenaciones?

Para ello utilizaremos el conocido como coeficiente de correlación vector, denominado coeficiente R-V. Este coeficiente es una extensión multivariante del coeficiente de correlación de Pearson, con la diferencia fundamental de que mide la correlación existente entre tablas de datos en lugar de entre variables.

Definición 4.1.3 *El valor del coeficiente de correlación vector R-V es equivalente al cuadrado del coeficiente de correlación lineal ρ y tiene la siguiente expresión:*

$$\frac{Coin(\mathbf{XY})}{\sqrt{Coin(\mathbf{XX})Coin(\mathbf{YY})}}$$

Toma valores entre 0 y 1, pues se define para matrices semidefinidas positivas. Mientras más cercano a 1 sea el valor de este coeficiente, mayor será la correlación entre los dos conjuntos de variables.

4.2. Aplicación práctica del ACoi

En esta sección vemos como el Análisis de Coinercia puede revelar patrones o tendencias que siguen conjuntos de datos donde el número de variables excede en gran medida del número de muestras, como es el caso de los análisis de microarrays.

Comenzaremos aplicando Análisis de Correspondencias a los conjuntos iniciales para identificar gráficamente la relación entre variables de cada matriz de datos.

Una vez calculados los ejes con coinercia máxima, podremos representar gráficamente la divergencia entre perfiles de expresión génica obtenidos desde diferentes plataformas microarrays. Los genes que definan las tendencias principales en el análisis podrán identificarse fácilmente.

Datos

Examinamos dos conjuntos de datos de expresión génica con las mismas 60 líneas celulares, que serán las muestras en nuestro estudio. Estas líneas derivan de pacientes con 9 fenotipos tumorales distintos (leucemia (LE), melanomas (ME), cáncer de ovarios (OV), de pecho (BR), próstata (PR), pulmón (LU), renal (RE), colon (CO) y del sistema central nervioso (CNS)).

La expresión génica de estas células fue analizada por los grupos Ross2000 y Staunton2001 en diferentes plataformas microarrays.

Compararemos los resultados obtenidos por la compañía Affymetrix (Staunton et al., 2001) con los resultados obtenidos usando spotted arrays (o microarrays de dos colores) por la compañía Stanford (Ross et al., 2000).
Los datos completos de Stanford y Affymetrix pueden encontrarse en `http://discover.nci.nih.gov/datasetsNature2000.jsp` y en `http:\www-genome.wi.mit.edu/mpr/NCI60/`; respectivamente.

1. El conjunto de datos del grupo Ross contiene un subconjunto de los genes originales de tamaño 1375. Aquellos con más del 15 % de valores perdidos se han quitado y el resto se han calculado por el método de los k vecinos más cercanos, considerando 16 vecinos y la distancia euclídea. Además, el conjunto de datos pre-procesado contiene valores logarítmicos.

2. El conjunto de datos Staunton contiene valores para un subconjunto de 1517 del total de genes estudiados. Algunos pasos llevados a cabo en el preprocesamiento de los datos

han sido reemplazar los valores que se diferencien de la media menos de 100 unidades por 100, eliminar genes cuya expresión ha sido invariante a lo largo de las 60 líneas celulares y seleccionar los subconjuntos de genes que conlleven un cambio mínimo de al menos 500 unidades de diferencia con la media. También se han centrado los datos y se ha aplicado una transformación logarítmica (base 2).

Ambos conjuntos de datos están disponibles en el fichero *NCI60* del paquete *made4*.

En esta aplicación práctica se han seleccionado 144 variables de las 1375 que contiene la base de datos original del grupo Ross, y 144 de las 1517 originales del grupo Staunton.

Consideramos por tanto 144 variables distintas para cada laboratorio medidas sobre los mismos individuos, cada uno de los cuales presenta uno de los 9 tipos de cáncer que estudiamos.

El Análisis de Coinercia permitirá visualizar los genes con patrones de expresión similares a través de las dos plataformas consideradas.

Paquetes

Para llevar a cabo este análisis en R, hay que ejecutar la instrucción `cia` en el paquete *ade4* descrito en ejemplos anteriores.

4.2.1. Análisis de Correspondencias

En este caso, no hay asunción previa de la relación entre dos conjuntos de variables. Nos interesa integrar estos dos conjuntos de datos y escoger variables simultáneamente.

Por tanto, llevaremos a cabo un análisis asimétrico de los datos.

Cargamos la librería *made4*, y ejecutamos los datos que necesitamos.

```
library(made4)
data(NCI60)
class(NCI60)

## [1] "list"
```

Vemos que los datos están almacenados en una lista.

```
names(NCI60)

## [1] "Ross"    "Affy"    "classes" "Annot"
```

Esta lista contiene los siguientes elementos:

a) "NCI60$Ross" es un data.frame de dimensión 144×60. Se corresponde con las medidas de los 144 genes que interesan para el estudio llevadas a cabo por el laboratorio Ross.

b) "NCI60$Affy" es un data.frame de dimensión 144×60. Se corresponde con medidas de 144 genes llevadas a cabo por el laboratorio Affymetrix.

c) "NCI60$classes" es una matriz de dimensión 60×2 que contiene información sobre el tipo de cáncer de cada individuo.

d) "NCI60$Annot" es un data.frame de dimensión 144×4 conteniendo 4 anotaciones distintas de los genes medidos en cada una de las muestras.

Con la función `summary` obetenemos información del análisis llevado a cabo por los dos laboratorios sobre los mismos 60 individuos.

```
summary(NCI60)

##         Length Class      Mode
## Ross     60    data.frame list
## Affy     60    data.frame list
## classes 120    -none-     character
## Annot     4    data.frame list
```

Veamos los datos de ambos grupos para las 4 primeras variables y las 4 primeras muestras.

```
NCI60$Ross[1:4,1:4]

##        BREAST_BT549 BREAST_HS578T BREAST_MCF7
## 484963        0.713         0.544      -1.643
## 510395       -2.293        -2.584      -2.471
## 76539         1.130         1.443      -0.062
## 509732        0.837         0.824      -0.057
##        BREAST_MCF7ADRr
## 484963           1.162
## 510395          -2.863
## 76539            0.318
## 509732           0.227
```

```
NCI60$Affy[1:4,1:4]

##                BREAST_BT549 BREAST_HS578T BREAST_MCF7
## V00594_s_at       0.7131644     0.6079392 -4.40948789
## M60854_at        -0.2952515    -0.3567663  0.05465057
## X53331_at         2.1143670     0.3103401  0.29865832
## L19686_rna1_at    0.5050693     0.1812558  0.53823619
##                BREAST_MCF7ADRr
## V00594_s_at         0.261889367
## M60854_at           0.001557987
## X53331_at           0.000000000
## L19686_rna1_at     -0.027256449
```

Con la función `table` se obtiene una tabla de frecuencias para cada clase de cáncer.

```
table(NCI60$classes[,2])

##
##   BREAST      CNS    COLON     LEUK    MELAN    NSCLC
##        8        6        7        6        8        9
##     OVAR PROSTATE    RENAL
##        6        2        8
```

También podemos ver las anotaciones de las primeras variables de cada conjunto:

```
NCI60$Annot[1:4,]

##           Ross.ImageID Symbol.Source   Affy.Affy ID
## Hs.418241       484963          MT2A    V00594_s_at
## Hs.397609       510395         RPS16      M60854_at
## Hs.365706        76539           MGP      X53331_at
## Hs.407995       509732           MIF L19686_rna1_at
##           Symbol.AnnAffy
## Hs.418241           MT2A
## Hs.397609          RPS16
## Hs.365706            MGP
## Hs.407995            MIF
```

Procedamos al Análisis de Correspondencias Asimétrico de los datos obtenidos por el grupo Ross.

La función `ord` simplifica el funcionamiento de métodos de ordenación canónicos como Análisis de Componentes Principales, Correspondencias y Correpondencias Asimétrico. Proporciona una "envoltura" con la cual se puede llamar a cada uno de esos métodos.
Para llevar a cabo el Análisis de Correspondencias Asimétrico hemos especificado "type="nsc'.

```
data.coa1 <-ord(NCI60$Ross, type="nsc")
```

Veamos los resultados de la ordenación:

```
data.coa1$ord

## Duality diagramm
## class: nsc dudi
## $call: dudi.nsc(df = data.tr, scannf = FALSE, nf = ord.nf)
##
## $nf: 59 axis-components saved
## $rank: 59
## eigen values: 0.007109 0.004584 0.003671 0.002675 0.001829 ...
##   vector length mode    content
## 1 $cw    60     numeric column weights
## 2 $lw    144    numeric row weights
## 3 $eig   59     numeric eigen values
##
##   data.frame nrow ncol content
## 1 $tab       144  60   modified array
## 2 $li        144  59   row coordinates
## 3 $l1        144  59   row normed scores
## 4 $co        60   59   column coordinates
## 5 $c1        60   59   column normed scores
## other elements: N
```

Esta lista de longitud 14 incluye los 59 autovalores que resultan de la DVS de la matriz inicial, los pesos de las columnas y de las filas, además de las nuevas coordenadas para filas y columnas. Ya vimos con profundidad la estructura de los resultados en la aplicación para el Análisis de Correspondencias del capítulo anterior.

Para el Análisis de Correspondencias asimétrico se trabaja con las Coordenadas Principales y Coordenadas Standard, como vimos en teoría.

Hemos hallado también los valores de coordenadas filas ($\underline{r}_j$) y columnas ($\underline{s}_j$) para los primeros genes e individuos; respectivamente.

```
data.coa1$ord$li[1:4,1:3]

##              Axis1       Axis2        Axis3
## 484963 -0.08471002 -0.02782602  0.013984051
## 510395  0.18763603 -0.00994279 -0.008648991
## 76539  -0.03424693  0.01577408 -0.027110886
## 509732 -0.01198044  0.06460366  0.005213023

data.coa1$ord$co[1:4,1:3]

##                        Comp1       Comp2       Comp3
## BREAST_BT549    -0.10275027 -0.05317834 -0.03500724
## BREAST_HS578T   -0.18105947 -0.09919005 -0.03431942
## BREAST_MCF7      0.11365401  0.02823559 -0.08839012
## BREAST_MCF7ADRr -0.03574366 -0.03563459 -0.04106029
```

La varianza acumulada por los primeros ejes viene dada por el siguiente resultado:

```
head(cumsum(data.coa1$ord$eig*100/sum(data.coa1$ord$eig)))

## [1] 17.91786 29.47025 38.72162 45.46232 50.07146 54.35127
```

Por tanto, el 29.47025 % de la varianza está explicada por los dos primeros ejes.

Ahora, llevamos a cabo un Análisis de Correspondencias Asimétrico de los datos; considerando sólo los resultados obtenidos por el grupo Affymetrix.

```
data.coa2 <-ord(NCI60$Affy, type="nsc")
```

Al igual que con el grupo Ross, en los resultados de ordenación se obtienen autovalores, pesos filas y columnas, coordenadas filas y columnas, coordenadas normalizadas,etc.
La varianza acumulada viene dada por

```
head(cumsum(data.coa2$ord$eig*100/sum(data.coa2$ord$eig)))

## [1] 17.54391 27.38023 34.10021 39.07819 43.68462 47.91515
```

Por tanto, el 27.38023 % de la varianza está explicada por los dos primeros ejes

Análisis de Correspondencias- Visualización de resultados

Hay muchas funciones en *made4* para visualizar resultados de un análisis de ordenación. La forma más simple de ver gráficamente los resultados de `ord` es usando la función `plot`. Dibujaremos los autovalores, variables y casos (microarrays).
Ya vimos la interpretación de estos gráficos con detalle en el capítulo anterior.

```
plot(data.coa1, classvec = NCI60$classes[,2])
```

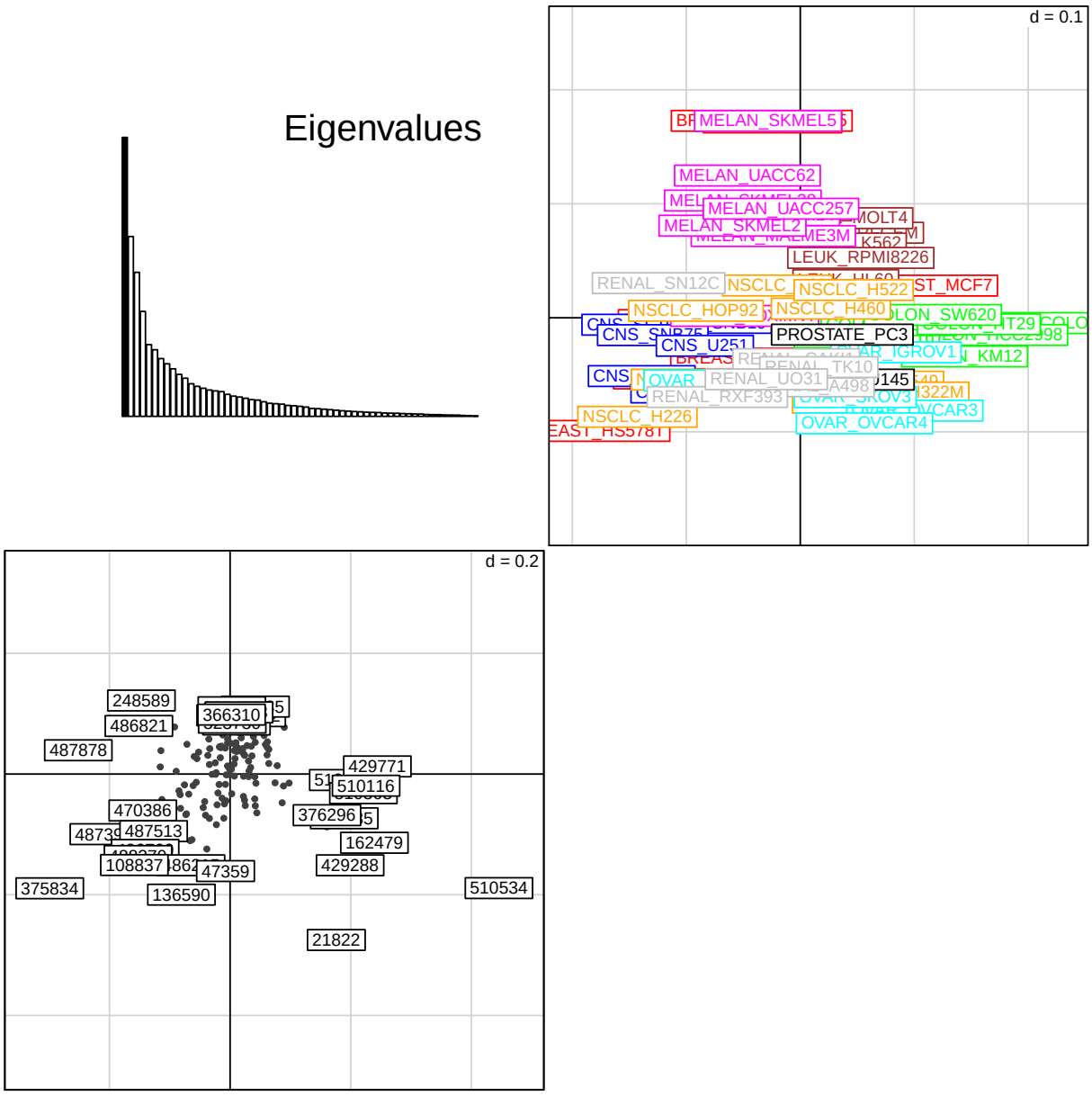
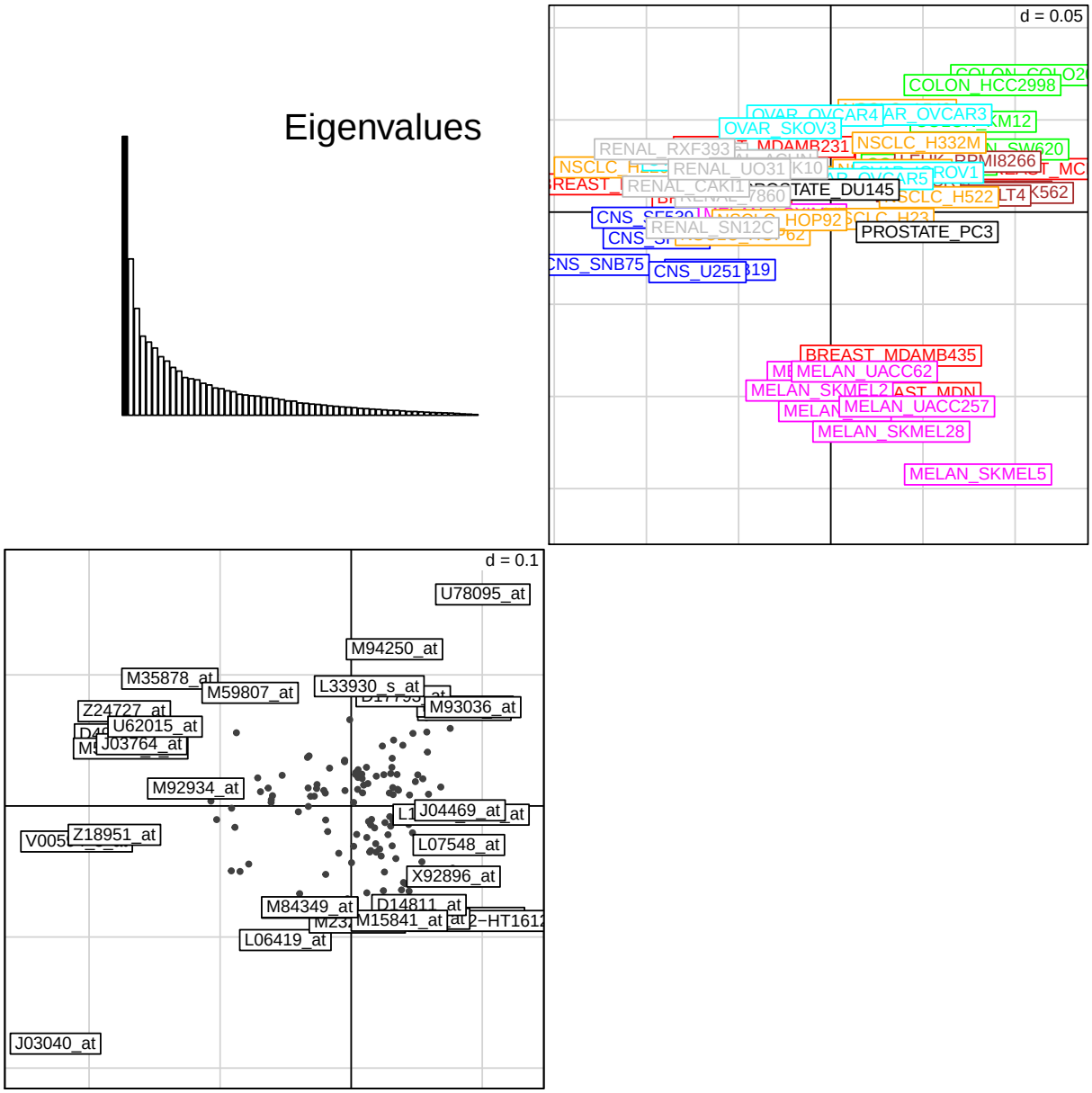


Figura 4.1: Plot del Análisis de Correspondencias(Ross).
A. autovalores
B. proyección de muestras microarrays (proyecciones columnas)
C. proyección de genes (proyecciones filas).

```
plot(data.coa2, classvec = NCI60$classes[,2])
```

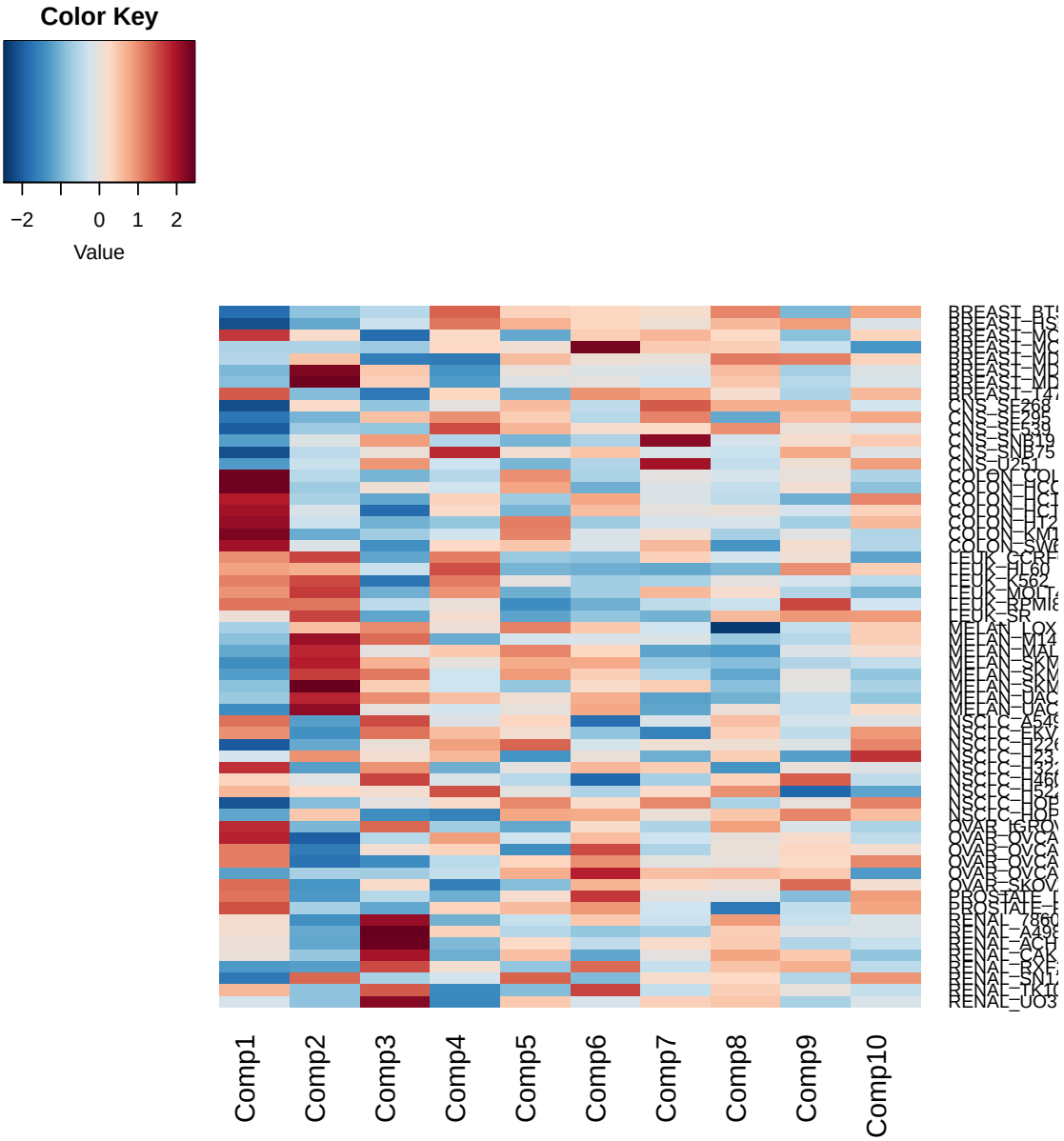


Plot del Análisis de Correspondencias(Affy).
A. autovalores
B. proyección de muestras microarrays (proyecciones columnas)
C. proyección de genes (proyecciones filas).
En ambas representaciones de las muestras analizadas podemos observar que las líneas celulares correspondientes a melanomas (ME) se distinguen del resto por el segundo eje.

A veces es útil tener una visión de separación por grupos de cada una de las Coordenadas. La función `heatplot` representa en el eje de abcisas las coordenadas (Comp)) y en el eje de

ordenadas los pesos columnas (60 muestras).

```
heatplot(data.coa1$ord$co[,1:10],dend="none",
         classvec=NCI60$classes, scale="none")
```



Con este gráfico vemos qué peso tiene cada uno de los individuos en cada uno de los 10 primeros ejes.
Aquellos que presentan el fenotipo tumoral leucemia (LE) son los que más peso tienen en el primer eje, mientras que los que presentan fenotipo tumoral colon (CO) contribuyen más a la determinación del segundo eje. Esto coincide con los resultados obtenidos tras el Análisis de Correspondencia de ambos conjuntos de datos.

4.2.2. Análisis de Coinercia

Llevaremos a cabo un Análisis de Coinercia para comparar ambos conjuntos. Con la implementación del Análisis de Coinercia del paquete *ade4* (usando la función `cia`), se supone que los pesos de las filas de ambos conjuntos de datos son los mismos.

Aunque ambos conjuntos tengan genes diferentes, se espera que muestren patrones y tendencias similares.

```
coin <- cia(NCI60$Ross, NCI60$Affy)
names(coin)

## [1] "call"      "coinertia" "coa1"      "coa2"
```

```
coin$coinertia
```

```
Coinertia analysis
call: coinertia(dudiX = coa1, dudiY = coa2, scannf =
cia.scan, nf = cia.nf)
class: coinertia dudi

$rank (rank)     : 59
$nf (axis saved) : 2
$RV (RV coeff)   : 0.7859656

eigenvalues: 2.266e-05 9.904e-06 4.342e-06 2.335e-06  ...

  vector length mode    content
1 $eig   59     numeric Eigenvalues
2 $lw    144    numeric Row weigths (for coa2 cols)
3 $cw    144    numeric Col weigths (for coa1 cols)

   data.frame nrow ncol content
1  $tab       144  144  Crossed Table (CT): cols(coa2) x cols(coa1)
2  $li        144  2    CT row scores (cols of coa2)
3  $l1        144  2    Principal components (loadings for coa2 cols)
4  $co        144  2    CT col scores (cols of coa1)
5  $c1        144  2    Principal axes (loadings for coa1)
6  $lX        60   2    Row scores (rows of coa1 cols)
7  $mX        60   2    Normed row scores (rows of coa1)
8  $lY        60   2    Row scores (rows of coa2)
9  $mY        60   2    Normed row scores (rows of coa2)
```

```
10 $aX        2    2    Corr coa1 axes / coinertia axes
11 $aY        2    2    Corr coa2 axes / coinertia axes

CT rows = cols of coa2 (144) / CT cols = cols of coa1 (144)
```

Obtenemos un lista con los 59 autovalores, pesos filas (lw) y columnas (cw).
Se obtienen también las coordenadas para las variables de los conjuntos del grupo Affymetrix (li) y Ross (co); así como las coordenadas para las muestras.
Las coordenadas para los individuos para el grupo Ross se recogen en $\$lX$, mientras que las del grupo Affymetrix se recogen en $\$lY$.

El coeficiente RV es una medida de similitud global entre ambos conjuntos de datos. Mientras más cercano a 1, mayor será la correlación entre ambos conjuntos.

```
coin$coinertia$RV

## [1] 0.7859656
```

Hemos obtenido un valor igual a $0{,}7859656$, que indica una correlación bastante fuerte.

Para visualizar las líneas celulares con perfiles de expresión génica similares o distintas usamos `plotarrays`.

Las muestras están representadas en las coordenadas dadas por lX y lY para cada matriz. Cada línea celular está coloreada por su fenotipo (por ejemplo, colon de verde, pecho de rojo, melanoma de rosa, etc).

En el gráfico podemos ver puntos (Ross) y flechas (Affy) unidas por líneas. La distancia entre estos puntos y flechas indica la similitud entre perfiles.

```
plotarrays(coin, classvec=NCI60$classes[,2], lab="", cpoint=3)
```

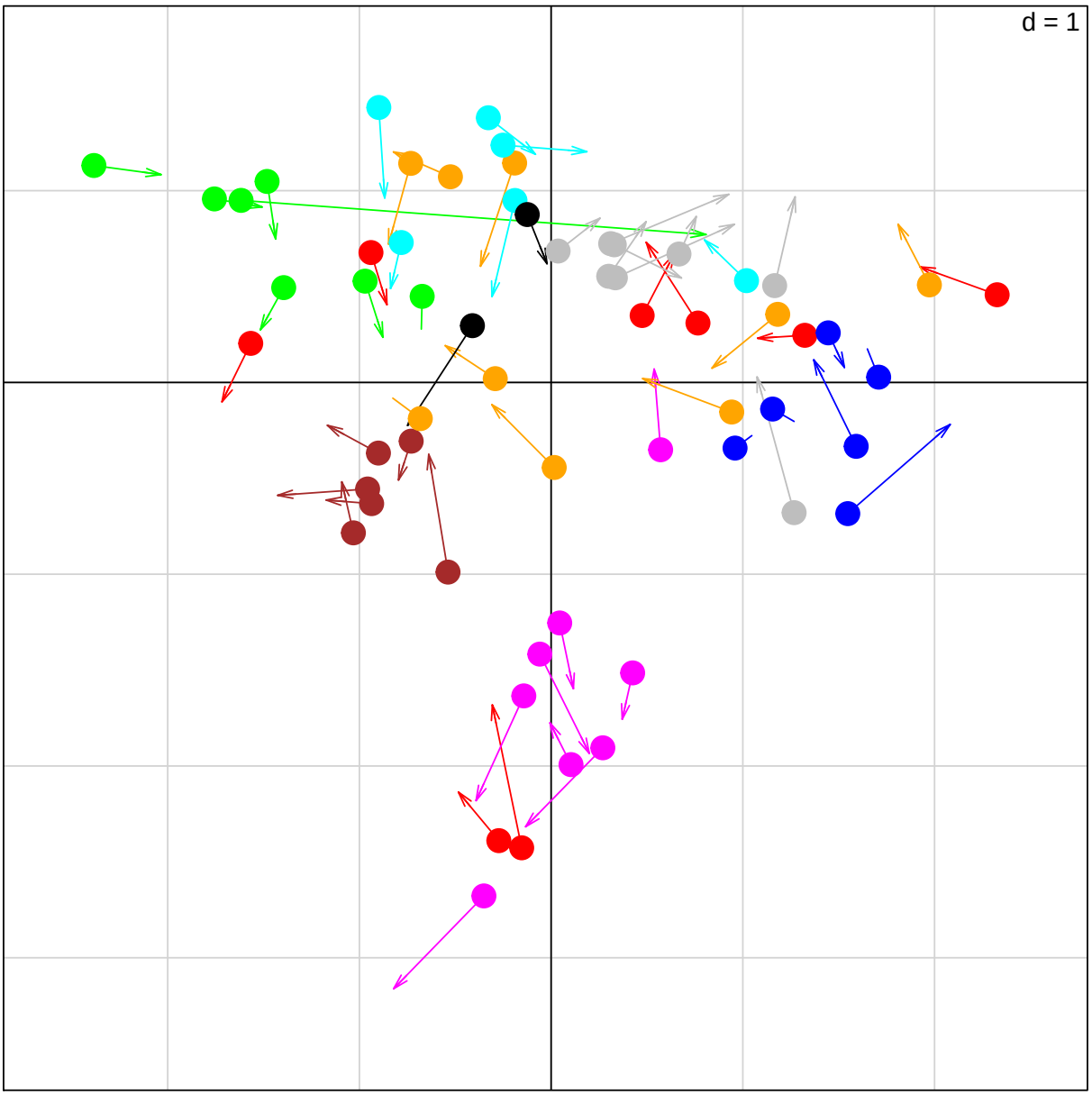


Figura 4.2: Comparación de los perfiles de expresión génica de spotted arrays y Affymetrix de NCI60.

Con la función `plot`, se representan simultáneamente el gráfico anterior y las proyecciones génicas para cada conjunto de datos.
En el primer gráfico puede verse como el primer eje separa las líneas celulares correspondientes a leucemia (LE, color marrón) y colon (CO, color verde) del resto de fenotipos tumorales y que las líneas celulares correspondientes a melanomas (ME, color rosa) se distinguen del resto por el segundo eje.

Respecto a los genes representados en los gráficos B y C, podemos afirmar que aquellos localizados al final de los ejes son los que más intervienen en la definición de los mismos.

La interpretación conjunta de estos tres gráficos revela que genes y líneas celulares proyectadas en la misma dirección desde el origen tienen una fuerte asociación.

```
plot(coin, classvec=NCI60$classes[,2])
```

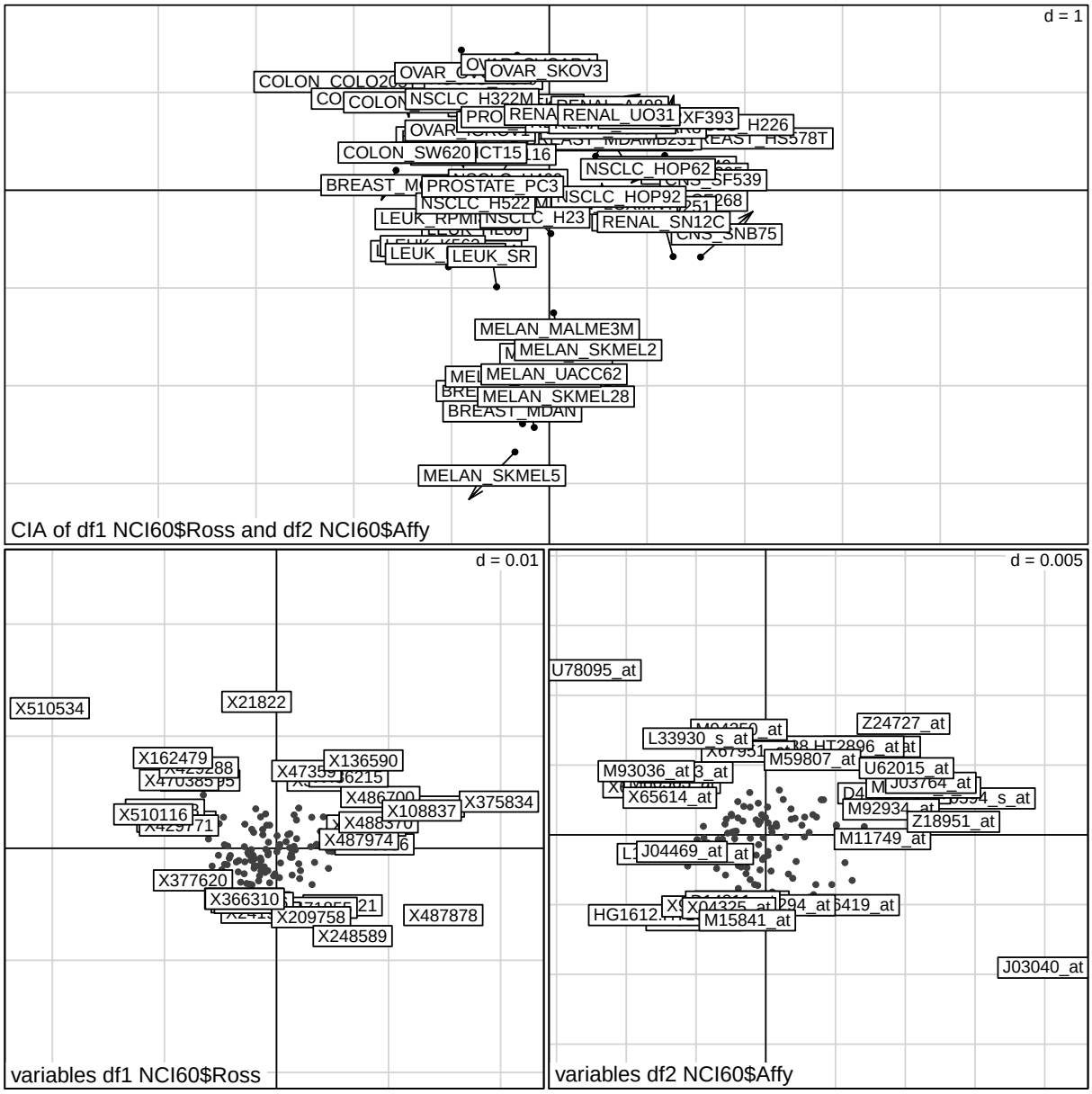


Figura 4.3: Análisis de Coinercia.
A. Representa las 60 muestras microarrays proyectadas en un espacio único. Los 60 círculos representan el conjunto Ross y las 60 flechas representan el conjunto Affy.
B. proyecciones génicas para Ross.
C. proyecciones génicas para Affy.

Capítulo 5

Análisis de Correlación Canónica

El principal propósito del Análisis de Correlación Canónica (ACC) es la exploración de la correlación entre conjuntos de variables.

Existen tres tipos de correlación:

- Correlación simple, si se relacionan dos variables.
- Correlación múltiple, si se relacionan una variable y un vector aleatorio.
- Correlación canónica, si se relacionan dos vectores aleatorios.

Esta herramienta del análisis estadístico multivariante desarrollada en un principio por Hotelling se basa en las proyecciones, ya que estructuras de datos multivariantes muy complejas son más fáciles de comprender si estudiamos proyecciones de menor dimensión.

La proyección de una variable define un índice que tiene correlación máxima con el índice de otra variable para cada muestra separadamente. El objetivo es maximizar la correlación entre las proyecciones de menor dimensión de los dos conjuntos de datos.
Los vectores de correlación canónica se obtendrán a partir del análisis de covarianza conjunto de las dos variables.

5.1. El método

Supongamos que tenemos dos conjuntos de variables aleatorias $\underline{X} \in \mathbb{R}^p$ y $\underline{Y} \in \mathbb{R}^q$,y por tanto dos matrices $\mathbf{X}$ e $\mathbf{Y}$ de orden $n \times p$ y $n \times q$, respectivamente. Las columnas de éstas corresponderían a las variables y las filas a las unidades experimentales.

Asumiremos, sin pérdida de generalidad, que las variables de ambas matrices están estandarizadas y que $p < q$. Supondremos además que ningún conjunto de variables explica el otro, planteando por tanto el problema de forma simétrica.

En este análisis, el problema es encontrar un índice que describa la posible relación lineal entre $\underline{X}$ e $\underline{Y}$. Estos índices se conocen como variables canónicas y exactamente podemos contruir $m = \text{mín}\{p, q\}$ parejas de variables.

Estas variables son las combinaciones de las variables originales $\underline{a}_i'\underline{X}$ y $\underline{b}_i'\underline{Y}$, con $i = 1, \ldots, m$. Se escogen los coeficientes $\underline{a}_i$ y $\underline{b}_i$, conocidos como vectores canónicos, de forma que la correlación lineal entre estos índices sea máxima.

El problema por tanto se plantea cómo la búsqueda de $\underline{a}_i = (a_{i1}, \cdots, a_{ip})'$ y $\underline{b}_i = (b_{i1}, \cdots, b_{iq})'$ tal que:

$$\eta_i = \underline{a}_i'\underline{X} = a_{i1}X_1 + \cdots + a_{ip}X_p, \qquad \varphi_i = \underline{b}_i'\underline{Y} = b_{i1}Y_1 + \cdots + b_{iq}Y_q$$

y la correlación lineal entre η_i y φ_i sea máxima.

Veamos cómo expresar el coeficiente de correlación entre las proyecciones de $\underline{X}$ e $\underline{Y}$.

Si suponemos que :

$$\begin{pmatrix} \underline{X} \\ \underline{Y} \end{pmatrix} \sim \left(\begin{pmatrix} \underline{\mu} \\ \underline{\nu} \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma_{XX}} & \mathbf{\Sigma_{XY}} \\ \mathbf{\Sigma_{YX}} & \mathbf{\Sigma_{YY}} \end{pmatrix} \right)$$

Entonces sabemos que se cumplen las siguientes propiedades:

$$Var(\underline{X}) = \mathbf{\Sigma_{XX}}_{(p\times p)}$$

$$Var(\underline{Y}) = \mathbf{\Sigma_{YY}}_{(q\times q)}$$

$$Cov(\underline{X}, \underline{Y}) = E((\underline{X} - \underline{\mu})(\underline{Y} - \underline{\nu})') = \mathbf{\Sigma_{XY}} = \mathbf{\Sigma'_{YX}}_{(p\times q)}$$

De esta forma, es fácil obtener la expresión de $\rho(\eta_i, \varphi_i)$ que buscamos:

$$\rho(\eta_i, \varphi_i) = \frac{Cov(\eta_i, \varphi_i)}{\sqrt{Var(\eta_i)}\sqrt{Var(\varphi_i)}} = \frac{Cov(\underline{a}_i'\underline{X}, \underline{b}_i'\underline{Y})}{\sqrt{Var(\underline{a}_i'\underline{X})}\sqrt{Var(\underline{b}_i'\underline{Y})}} = \frac{\underline{a}_i'\mathbf{\Sigma_{XY}}\underline{b}_i}{(\underline{a}_i'\mathbf{\Sigma_{XX}}\underline{a}_i)^{1/2}(\underline{b}_i'\mathbf{\Sigma_{YY}}\underline{b}_i)^{1/2}}$$

Para cualquier escalar $c \in \mathbb{R}^+$ se tiene que:

$$\rho(c\eta_i, \varphi_i) = \rho(\eta_i, c\varphi_i) = \rho(\eta_i, \varphi_i)$$

Sabido esto, podemos plantear apropiadamente el problema a resolver.
Comenzaremos calculando el primer par de vectores de correlación canónica $\underline{a}_1$ y $\underline{b}_1$ con detalle.

5.1.1. Primer par de vectores de correlación canónica

El problema que tratamos resolver es el siguiente:

$$\max_{\underline{a}_1,\underline{b}_1}\{\underline{a}_1'\mathbf{\Sigma_{XY}}\underline{b}_1\}$$

bajo las restricciones

$$\begin{cases}\underline{a}_1'\mathbf{\Sigma_{XX}}\underline{a}_1 = 1\\ \underline{b}_1'\mathbf{\Sigma_{YY}}\underline{b}_1 = 1\end{cases}$$

Las restricciones impuestas sirven para expresar que las varianzas de η_1 y φ_1 están normalizadas, pues si aumentaramos el valor de $\underline{a}_1$ y $\underline{b}_1$ las varianzas correspondientes lo harían también.

Resolvemos este problema por el método de los multiplicadores de Lagrange.

Consideramos la siguiente función L y buscamos el máximo.

$$L = (\underline{a}_1'\mathbf{\Sigma_{XY}}\underline{b}_1) - \frac{\lambda}{2}(\underline{a}_1'\mathbf{\Sigma_{XX}}\underline{a}_1 - 1) - \frac{\mu}{2}(\underline{b}_1'\mathbf{\Sigma_{YY}}\underline{b}_1 - 1)$$

Para calcular el máximo igualamos las derivadas parciales de esta expresión a cero.

$$\frac{\partial L}{\partial \underline{a}_1} = \mathbf{\Sigma_{XY}}\underline{b}_1 - \lambda\mathbf{\Sigma_{XX}}\underline{a}_1 = 0$$
$$\frac{\partial L}{\partial \underline{b}_1} = \mathbf{\Sigma_{YX}}\underline{a}_1 - \mu\mathbf{\Sigma_{YY}}\underline{b}_1 = 0$$

Despejándo obtenemos las siguientes relaciones:

$$\begin{cases}\mathbf{\Sigma_{XY}}\underline{b}_1 = \lambda\mathbf{\Sigma_{XX}}\underline{a}_1\\ \mathbf{\Sigma_{YX}}\underline{a}_1 = \mu\mathbf{\Sigma_{YY}}\underline{b}_1\end{cases}$$

Multiplicando por $\underline{a}_1'$ en la primera ecuación y por $\underline{b}_1'$ en la segunda se obtiene:

$$\begin{cases}\underline{a}_1'\mathbf{\Sigma_{XY}}\underline{b}_1 = \lambda\underline{a}_1'\mathbf{\Sigma_{XX}}\underline{a}_1\\ \underline{b}_1'\mathbf{\Sigma_{YX}}\underline{a}_1 = \mu\underline{b}_1'\mathbf{\Sigma_{YY}}\underline{b}_1\end{cases}$$

Puesto que $\mathbf{\Sigma_{XY}}'$ es equivalente a $\mathbf{\Sigma_{YX}}$, podemos afirmar que $\lambda = \mu$, y por tanto el sistema de ecuaciones resultante es:

$$\underline{a}_1'\mathbf{\Sigma_{XY}}\underline{b}_1 = \lambda\underline{a}_1'\mathbf{\Sigma_{XX}}\underline{a}_1 \tag{5.2}$$
$$\underline{b}_1'\mathbf{\Sigma_{YX}}\underline{a}_1 = \lambda\underline{b}_1'\mathbf{\Sigma_{YY}}\underline{b}_1 \tag{5.3}$$

Si despejamos $\underline{b}_1$ de 5.3, se obtiene la relación $\underline{b}_1 = \lambda^{-1}\mathbf{\Sigma_{YY}}^{-1}\mathbf{\Sigma_{YX}}\underline{a}_1$.

Sustituyendo en la ecuación 5.2 se satisface la siguiente igualdad:

$$\mathbf{\Sigma_{XY}}(\lambda^{-1}{\mathbf{\Sigma_{YY}}}^{-1}\mathbf{\Sigma_{YX}}\underline{a}_1) = \lambda\mathbf{\Sigma_{XX}}\underline{a}_1$$

que equivale a $({\mathbf{\Sigma_{XX}}}^{-1}\mathbf{\Sigma_{XY}}{\mathbf{\Sigma_{YY}}}^{-1}\mathbf{\Sigma_{YX}})\underline{a}_1 = \lambda^2\underline{a}_1$

Es sencillo observar que $\underline{a}_1$ es el autovector correspondiente al valor propio λ^2 de la matriz $({\mathbf{\Sigma_{XX}}}^{-1}\mathbf{\Sigma_{XY}}{\mathbf{\Sigma_{YY}}}^{-1}\mathbf{\Sigma_{YX}})$.

Operando de la misma forma con la ecuación restante obtenemos la relación correspondiente a $\underline{b}_1$. A partir de esto podemos afirmar el siguiente resultado.

Teorema 5.1.1 *Los primeros vectores canónicos verifican las siguientes relaciones*

$$({\mathbf{\Sigma_{XX}}}^{-1}\mathbf{\Sigma_{XY}}{\mathbf{\Sigma_{YY}}}^{-1}\mathbf{\Sigma_{YX}})\underline{a}_1 = \lambda^2\underline{a}_1$$

$$({\mathbf{\Sigma_{YY}}}^{-1}\mathbf{\Sigma_{YX}}{\mathbf{\Sigma_{XX}}}^{-1}\mathbf{\Sigma_{XY}})\underline{b}_1 = \lambda^2\underline{b}_1,$$

siendo λ uno de los multiplicadores de Lagrange asociado al problema anteriormente planteado.

Teorema 5.1.2 *Los vectores canónicos, cumpliendo $\underline{a}_1'\mathbf{\Sigma_{XX}}\underline{a}_1 = 1$ y $\underline{b}_1'\mathbf{\Sigma_{YY}}\underline{b}_1 = 1$, se relacionan por*

$$\underline{a}_1 = \lambda^{-1}{\mathbf{\Sigma_{XX}}}^{-1}\mathbf{\Sigma_{XY}}\underline{b}_1,$$
$$\underline{b}_1 = \lambda^{-1}{\mathbf{\Sigma_{YY}}}^{-1}\mathbf{\Sigma_{YX}}\underline{a}_1.$$

Definiéndose la primera correlación canónica como $\rho_1 = {\lambda_1}^{1/2}$, donde λ_1 es el mayor autovalor de ${\mathbf{\Sigma_{XX}}}^{-1}\mathbf{\Sigma_{XY}}{\mathbf{\Sigma_{YY}}}^{-1}\mathbf{\Sigma_{YX}}$.

Demostración: La demostración puede verse en [1]

El procedimiento para obtener el resto de vectores que definen las variables de correlación canónica es similar, suponiendo además que cada variable de correlación canónica es incorrelada con todas las anteriores.
Esto permite obtener cada par de variables ordenadas por importancia.

Aparte de las relaciones entre los vectores de correlación canónica obtenidas mediante el método de los multiplicadores de Lagrange, podemos abordar el problema planteado usando Descomposición en Valores Singulares; tal y como veremos a continuación.

5.1.2. Correlación Canónica y Descomposición Singular

Para resolver el problema planteado al comienzo de esta sección, partiremos de la siguiente matriz:

$$\mathbf{K} = \mathbf{\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}}$$

siendo $\mathbf{\Sigma_{XX}}$ la varianza del vector $\underline{X}$, $\mathbf{\Sigma_{YY}}$ la del vector $\underline{Y}$ y $\mathbf{\Sigma_{XY}} = \mathbf{\Sigma'_{YX}}$ la covarianza entre ambos conjuntos de variables $\underline{X}$ e $\underline{Y}$.

Llevando a cabo la descomposición en valores singulares de $\mathbf{K}$ obtenemos:

$$\mathbf{K} = \mathbf{\Gamma D \Delta'}$$

siendo $\mathbf{\Gamma} = (\underline{\gamma}_1, \underline{\gamma}_2, \ldots, \underline{\gamma}_m)$ autovectores de $\mathbf{KK'}$, $\mathbf{\Delta} = (\underline{\delta}_1, \underline{\delta}_2, \ldots, \underline{\delta}_m)$ autovectores de $\mathbf{K'K}$ y $\mathbf{D}$ matriz diagonal con la raíz de los m autovalores no nulos de $\mathbf{KK'}$.

Teorema 5.1.3 *Los vectores de correlación canónica $\underline{a}_1$ y $\underline{b}_1$ son:*

$$\underline{a}_1 = \mathbf{\Sigma_{XX}^{-1/2}}\underline{\gamma}_1$$

$$\underline{b}_1 = \mathbf{\Sigma_{YY}^{-1/2}}\underline{\delta}_1$$

siendo $\underline{\gamma}_1$ y $\underline{\delta}_1$ los autovectores recogidos en $\mathbf{\Gamma}$ y $\mathbf{\Delta}$ que están asociados al mayor autovalor de $\mathbf{KK'}$.

Demostración:

Sabemos por las propiedades de la descomposición en valores singulares que $\mathbf{\Gamma} = (\underline{\gamma}_1, \underline{\gamma}_2, \ldots, \underline{\gamma}_m)$ es el conjunto de autovectores de $\mathbf{KK'}$ cumpliendo $\mathbf{KK'} = \mathbf{\Gamma D^2 \Gamma'}$.

Desarrollamos el producto $\mathbf{KK'}$ como sigue

$$\mathbf{KK'} = \mathbf{\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}} = \mathbf{\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2}}$$

Si $\underline{\gamma}_1$ es el autovector de $\mathbf{KK'}$ asociado al mayor autovalor se cumple:

$$\mathbf{\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2}}\underline{\gamma}_1 = \lambda_1\underline{\gamma}_1$$

Multiplicamos a izquierda por $\mathbf{\Sigma_{XX}^{-1/2}}$ en ambos lados de la igualdad:

$$\mathbf{\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}}(\mathbf{\Sigma_{XX}^{-1/2}}\underline{\gamma}_1) = \lambda_1\mathbf{\Sigma_{XX}^{-1/2}}\underline{\gamma}_1$$

Comparando esto con la relación para $\underline{a}_1$ descrita en el Teorema 5.1.1 queda probada la primera igualdad.

Siguiendo el mismo razonamiento con $\mathbf{K}'\mathbf{K}$ probaremos la segunda igualdad.

□

Es posible probar que las correlaciones canónicas son invariantes por transformaciones lineales y, por tanto, pueden calcularse partiendo de las matrices de correlaciones.

En la siguiente sección veremos con detalle la expresión explícita de los m pares de vectores y variables de correlación canónica.

5.2. Variables de correlación canónica

Definimos estas variables a partir de los vectores de correlación canónica.

Definición 5.2.1 *Los m vectores de correlación canónica $\underline{a}_i$ y $\underline{b}_i$, se definen como:*

$$\underline{a}_i = \mathbf{\Sigma_{XX}^{-1/2}} \underline{\gamma}_i$$

$$\underline{b}_i = \mathbf{\Sigma_{YY}^{-1/2}} \underline{\delta}_i,$$

con $i = 1, \ldots, m \quad (m = min\{p, q\})$

$\underline{\gamma}_i$ es el i-ésimo autovector recogido en $\mathbf{\Gamma}$ y $\underline{\delta}_i$ el i-ésimo autovector en $\mathbf{\Delta}$ asociado al autovalor λ_i correspondiente.

Una vez que conocemos el valor exacto de nuestros índices, podemos definir las *variables de correlación canónica*

Definición 5.2.2 *Las variables de correlación canónica se definen como*

$$\eta_i = \underline{a}_i' \underline{X}$$

$$\varphi_i = \underline{b}_i' \underline{Y}$$

siendo $\rho_i = \lambda_i^{1/2}$ el correspondiente coeficiente de correlación canónica, con $i = 1, \ldots, m$.

El siguiente resultado recoge propiedades fundamentales de este método.

Teorema 5.2.1 *Si suponemos $\rho_1 > \rho_2 > \cdots > \rho_m$, se cumple:*

1. *Las variables canónicas $\eta_1, \cdots, \eta_m$ y $\varphi_1, \cdots, \varphi_m$ están incorreladas.*

$$Cov(\eta_i, \eta_j) = \underline{a}_i' \mathbf{\Sigma_{XX}} \underline{a}_j = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$$

$$Cov(\varphi_i, \varphi_j) = \underline{b}_i' \mathbf{\Sigma_{YY}} \underline{b}_j = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$$

2. *La primera correlación canónica ρ_1 es la correlación máxima entre una proyección de $\underline{X}$ y una de $\underline{Y}$.*
3. *ρ_2 es la correlación máxima entre proyecciones de $\underline{X}$ incorreladas con η_1 y proyecciones de $\underline{Y}$ incorreladas con φ_1.*
4. *$Cov(\eta_i, \varphi_j) = 0 \quad si \quad i \neq j$*

Demostración:

1. Inmediata por la ortogonalidad que caracteriza a los autovectores.

4. Sea $Cov(\eta_i, \varphi_j) = E(\underline{a}_i' \underline{X} \underline{Y}' \underline{b}_j) = \underline{a}_i' \mathbf{\Sigma_{XY}} \underline{b}_j$.

Por el Teorema 5.1.2 obtuvimos una expresión del vector $\underline{b}_j$. Sustituyendo en la expresión de la covarianza se muestra:

$$Cov(\eta_i, \varphi_j) = \underline{a}_i' (\lambda_j \mathbf{\Sigma_{XX}} \underline{a} j) = 0$$

□

Teorema 5.2.2 *Sean η_i y φ_i las i-ésimas variables de correlación canónica, con $i = 1, \cdots, m$. Definimos $\underline{\eta} = (\eta_1, \cdots, \eta_m)$ y $\underline{\varphi} = (\varphi_1, \cdots, \varphi_m)$ los vectores m-dimensionales. Entonces la matriz de varianzas y covarianzas es:*

$$Var \begin{pmatrix} \underline{\eta} \\ \underline{\varphi} \end{pmatrix} = \begin{pmatrix} I_k & \mathbf{\Lambda} \\ \mathbf{\Lambda} & I_m \end{pmatrix}$$

con $\mathbf{\Lambda} = diag(\lambda_1^{1/2}, \cdots, \lambda_m^{1/2})$

Este teorema prueba que los coeficientes de correlación canónica $\rho_i = \lambda_i^{1/2}$ son las covarianzas entre las variables canónicas η_i y φ_i y los índices $\eta_1 = \underline{a}_1' \underline{X}$ y $\varphi_1 = \underline{b}_1' \underline{Y}$ tienen covarianza máxima $\sqrt{\lambda_1} = \rho_1$.

5.3. Contrastes de significación

Una vez obtenidas las m relaciones canónicas, nos puede interesar decidir cuáles son significativas.
Suponiendo normalidad multivariante, pueden realizarse contrastes como el test secuencial de Barlett-Lawley o contrastes de independencia de los conjuntos de variables originales.

•Test de Bartlett-Lawley:

Suponemos $\underline{X} \sim N_p(\underline{0}, \mathbf{\Sigma_{XX}})$ e $\underline{Y} \sim N_q(\underline{0}, \mathbf{\Sigma_{YY}})$

En este test se plantea que las primeras k correlaciones canónicas obtenidas a partir de la matriz de varianzas y covarianzas de $\underline{X}$ e $\underline{Y}$ son no nulas, a diferencia del resto.

Para $k = 0, 1, \cdots, m$; y $\rho_0 = 1$, se plantea

$$H_0^k : \rho_k > \rho_{k+1} = \cdots = \rho_m = 0$$

Si se acepta la hipótesis nula, se acepta por tanto que las k primeras variables de correlación canónica expresan satisfactoriamente la dependencia entre variables; mientras que si se acepta la hipótesis alternativa quiere decir que no podemos prescindir de ninguna variable de correlación canónica.

Sea L_k es el estadístico del contraste

$$L_k = -\left[n - 1 - k - \frac{1}{2}(p + q + 1) + \sum_{i=1}^{k} \rho_i^{-2}\right] log\left[\prod_{i=k+1}^{m} (1 - \rho_i^2)\right]$$

se aceptará la hipótesis nula cuando este no sea significativo.

Contrastes de independencia

Este contraste se lleva a cabo suponiendo normalidad multivariante de $\underline{X}$ e $\underline{Y}$ y planteando la hipótesis nula de que $\underline{X}$ e $\underline{Y}$ son independientes.

Puesto que se supone normalidad, la hipótesis nula será equivalente a plantear que las variables son incorreladas ($\mathbf{\Sigma_{XY}} = 0$). Por tanto, el contraste es el siguiente:

$$\begin{cases} H_0 : \mathbf{\Sigma_{XY}} = 0 \\ H_1 : \mathbf{\Sigma_{XY}} \neq 0 \end{cases}$$

Para resolverlo tenemos dos opciones:

• Test de razón de verosimilitudes

Sea $\underline{x} = (\underline{x}_1, \cdots, \underline{x}_n)$ una muestra aleatoria de $\underline{X}$ con función de densidad definida como $f(\underline{x}, \underline{\theta})$ y $\underline{\theta} \in \mathbf{\Theta}$ un parámetro desconocido. Si existe una subregión paramétrica $\mathbf{\Theta_0}$ de $\mathbf{\Theta}$, se puede plantear el test:

$$\begin{cases} H_0 : \underline{\theta} \in \mathbf{\Theta_0} \\ H_1 : \underline{\theta} \in \mathbf{\Theta} - \mathbf{\Theta_0} \end{cases}$$

La razón de verosimilitud es el estadístico:

$$\lambda_R = \frac{L(\underline{x}, \hat{\underline{\theta}}_0)}{L(\underline{x}, \hat{\underline{\theta}})}$$

que toma valores entre 0 y 1, siendo $\hat{\underline{\theta}}$ el estimador de máxima verosimilitud de $\underline{\theta} \in \mathbf{\Theta}$, $\hat{\underline{\theta}}_0$ el correspondiente a $\underline{\theta} \in \mathbf{\Theta_0}$ y $L(\underline{x}, \underline{\theta})$ la función de verosimilitud definida como:

$$L(\underline{x}, \underline{\theta}) = \prod_{i=1}^{n} f(x_i, \underline{\theta})$$

Aplicando este test a nuestro caso, el estadístico que obtendríamos sería:

$$\lambda_R = \frac{|\widehat{\mathbf{\Sigma}}|}{|\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}||\widehat{\mathbf{\Sigma}}_{\mathbf{YY}}|} = \frac{|\widehat{\mathbf{R}}|}{|\widehat{\mathbf{R}}_{\mathbf{XX}}||\widehat{\mathbf{R}}_{\mathbf{YY}}|}$$

que sigue una distribución Lambda de Wilks $\Lambda(p, n-1-q, q)$ y cuando toma valores pequeños y significativos supone el rechazo de la hipótesis nula H_0.

Puede establecerse esta igualdad porque el estadístico es invariante por transformaciones lineales.

La distribución Lambda de Wilks se define a partir de dos conjuntos de variables independientes con distribución Wishart: $\mathbf{A} \sim W_p(\mathbf{\Sigma}, m), \mathbf{B} \sim W_p(\mathbf{\Sigma}, n)$

$$\lambda_R = \frac{|\mathbf{A}|}{|\mathbf{A}+\mathbf{B}|} = \frac{1}{|(\mathbf{I}+\mathbf{A}^{-1}\mathbf{B})|} \sim \Lambda(p, m, n)$$

• Principio de Unión-Intersección
Partimos de que disponemos de los vectores $\underline{\eta}$ y $\underline{\varphi}$.

El Principio de Unión-Intersección plantea tanto contrastes multivariantes como test univariantes. Por ejemplo, si nuestra hipótesis nula es del tipo $H_0 : \underline{\mu} = \underline{\mu}_0$, haciendo uso del vector $\underline{a}$ podemos transformar el contraste como sigue:

Sustituimos $\underline{X}$ por $\underline{X}_a = \underline{X}a$, de manera que la hipótesis nula sería:

$$H_0 = \cap_a H_0(a), \qquad \text{con } H_0(a) : \mu(a) = \mu_0(a)$$

Si aplicamos este Principio a nuestro caso, la hipótesis nula sería $H_0 : \rho(\underline{\eta}, \underline{\varphi}) = 0$

Esto nos lleva a estudiar la significación de la primera correlación canónica a través del test:

$$\begin{cases} H_0 : \rho_1 = max\rho(\underline{\eta}, \underline{\varphi}) = 0 \\ H_1 : \rho_1 > 0 \end{cases}$$

Sólo se acepta la hipótesis nula en caso de que la correlación no sea significativa.

Observación 5.3.1 *Como hemos mencionado anteriormente, ésta técnica de la que posteriormente derivarían otras técnicas de Análisis Multivariante fue introducida por Hotelling. El objetivo era encontrar el trasfondo de las relaciones entre cuerpo y mente a través de test mentales y medidas biométricas.*

Algunas de las aplicaciones de la técnica se han llevado a cabo en psicología, ecología o en datos génicos.

Los únicos resultados que se conocen sobre la distribución de las correlaciones canónicas son asintóticos (Muirhead, 1982), debido a la gran complejidad que supone.

5.4. Aplicación práctica del ACC

El Análisis de Correlación Canónica es un análisis exploratorio que permite ilustrar las correlaciones entre dos conjuntos de datos en que se miden las mismas unidades experimentales.

Dado que en este caso, el número de variables supera considerablemente el número de muestras analizadas; tendremos que tener esto en cuenta a la hora de realizar el análisis.

Comprobaremos que con los dos primeros pares de variables de correlación canónica podemos representar con gran exactitud la relación entre los dos conjuntos de variables.

Datos

En el siguiente ejercicio práctico trabajaremos con la base de datos `nutrimouse`, disponible en el paquete *CCA*.
En esta base de datos se recoge un estudio de la nutrición de ratones. Se estudiaron un total de 40 ratones, clasificados en función de dos factores:

1. Genotipo: El estudio se ha llevado a cabo en ratones de tipo salvaje y $PPAR_\alpha$ deficientes.

2. Dieta: Se usaron aceites para la preparación experimental de dietas. Acites de colza y maíz (50/50) fueron usados para una dieta de referencia (REF), aceite de coco hidrogenado para una dieta de ácidos grasos saturados (COC), aceite de girasol para una dieta rica en ácidos grasos $w6$ (SUN), aceite de semilla de lino para una dieta rica en ácidos grasos $w3$ (LIN) y aceites de maíz/colza/pescado enriquecido (42.5/42.5/15) para una dieta de pescado (FISH).

Mediremos 141 variables agrupadas en los siguientes conjuntos:

1. Los niveles de expresión de 120 genes medidos en células del hígado. Los genes elegidos se han seleccionado de un total de 30000, pues se consideran relevantes en el contexto de nuestro estudio.
2. Concentraciones de 21 ácidos grasos hepáticos medidos por cromatología de gases.

Paquetes

Para este ejemplo trabajaremos con el paquete *CCA* (disponible en "http://CRAN.R-project.org/"), que contiene las herramientas necesarias para llevar a cabo un Análisis de Correlación Canónica en conjuntos de datos con más variables que observaciones. Además, permite trabajar con valores perdidos.

5.4.1. ACC

En primer lugar, cargamos el paquete *CCA* previamente instalado.

```
require(CCA)
```

Una vez hecho esto, cargamos los datos contenidos en "nutrimouse". Con la función `View` podemos ver la matriz de datos en la consola de Rstudio.

```
data("nutrimouse")
```

Con `class` podemos ver la clase del objeto con el que trabajamos.

```
class(nutrimouse)

## [1] "list"
```

Es una lista de 4 elementos, que contiene:

1. "nutrimouse$gene", un data.frame de dimensión 40×120 que contiene los niveles de expresión de 120 genes en 40 muestras. Esta será una de las matrices con las que trabajaremos más adelante.
2. "nutrimouse$lipid", un data.frame de dimensión 40×21. Esta será la otra matriz utilizada en el análisis.

3. “nutrimouse$diet", un factor de longitud 40 indicando la dieta que sigue cada uno de los individuos.
4. “nutrimouse$genoType”, un factor de longitud 40 con distintos niveles que indican el tipo de ratón analizado en función de su genotipo.

Dado que los datos están almacenados en una lista, para evitar problemas que puedan surgir más adelante los convertiremos en una matriz, diferenciando las dos clases existentes entre las variables especificadas con anterioridad.

Los datos de los ratones a los que se han medido variables relacionadas con expresión génica se guardan en la matriz $\mathbf{X}$ y el resto de datos en la matriz $\mathbf{Y}$.
Con la función `dim` comprobamos que las dimensiones de las matrices se corresponden con lo explicado.

```
X <- as.matrix(nutrimouse$gene)
Y <- as.matrix(nutrimouse$lipid)
dim(X)

## [1]  40 120

dim(Y)

## [1] 40 21
```

Un paso preliminar para aplicar luego la técnica es conocer las correlaciones entre variables de ambas matrices. Haremos esto con la función `matcor`.

```
correl <- matcor(X, Y)
```

Podemos ver gráficamente esta matriz con la función `img.matcor`, donde “type=2” indica que tenemos en cuenta las 3 matrices generadas anteriormente; de dimensiones $p \times p$, $q \times q$ y $p \times q$ respectivamente.

```
img.matcor(correl, type = 2)
```

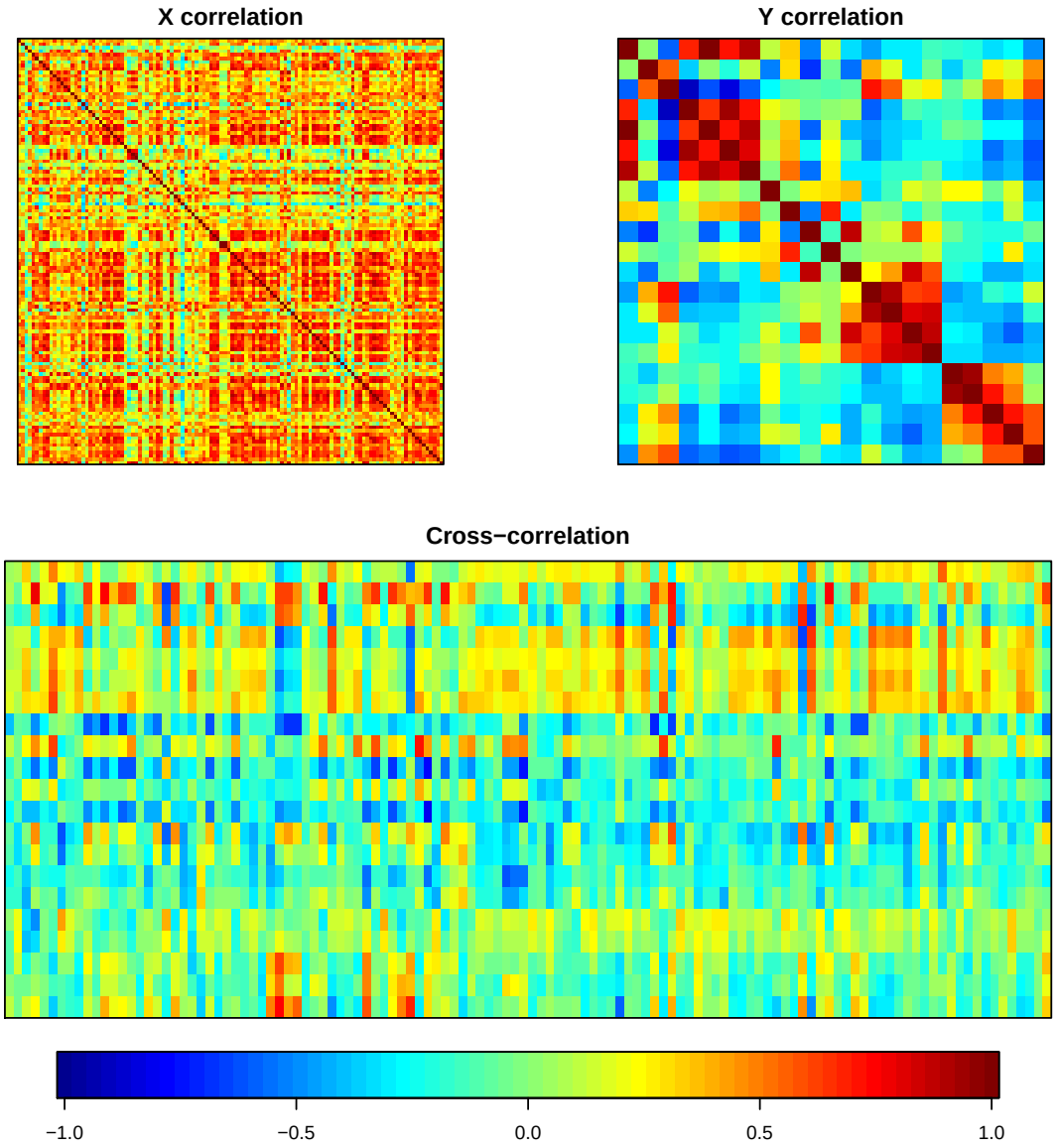


Figura 5.1: Matrices de correlación para las variables de **X** (lado superior izquierdo), para las variables de **Y** (lado superior derecho), para las variables de ambos conjuntos (parte inferior). Los valores se representan por colores que van desde azul (correlación negativa) hasta rojo (correlación positiva).

Para ilustrar la técnica se han seleccionado 10 variables de la matriz **X** con `sample` y el argumento "size=10". Hemos fijado una semilla para que siempre haga el muestreo seleccionando los mismos individuos.
En cambio, sí seleccionamos todas las variables de **Y**, ya que tan sólo son 21.

Aparecerán los valores que toman las variables de **X** seleccionadas al azar, así como las de **Y** para los 40 individuos.

Dado que se supone que las variables están estandarizadas cuando se aplica esta técnica, con la función `scale` estandarizamos los datos de ambas tablas de datos.

La función `cc()` se usa para llevar a cabo el análisis en este caso, donde hay que tener en cuenta que existen muchas más variables que individuos analizados.

```
set.seed(24)
Xr<- X[,sample(1:120, size=10)]
Xscale<- scale(Xr)
Yscale<- scale(Y)
res.cc <- cc(Xscale, Yscale)
```

Con la función `names` podemos ver todos los nombres de los elementos que se ha generado al aplicar ACC.

```
names(res.cc)

## [1] "cor"    "names"  "xcoef"  "ycoef"  "scores"
```

Si vemos los resultados de "res.cc" paso a paso, podemos observar que en primer lugar aparecen los valores de las correlaciones de los 10 índices resultantes $(cor(a_i'\underline{X}, b_i'\underline{Y})), i = 1, \ldots, 10$. La primera combinación de variables originales es la que tiene mayor correlación entre índices.

```
res.cc$cor

##  [1] 0.9888384 0.9795369 0.9409709 0.9218689 0.8894283
##  [6] 0.8042567 0.7679251 0.6598241 0.6176711 0.4321375
```

Cabe destacar que se han calculado 10 combinaciones posibles de las variables de cada matriz porque, como sabemos por teoría, se lleva a cabo el mínimo entre p y q.

Con `barplot`, respresentamos las correlaciones canónicas en función de las 10 variables de correlación canónica obtenidas en el análisis.

```
barplot(res.cc$cor, xlab = "Dimension", ylab =
    "Canonical correlations", names.arg = 1:10, ylim = c(0,1))
```

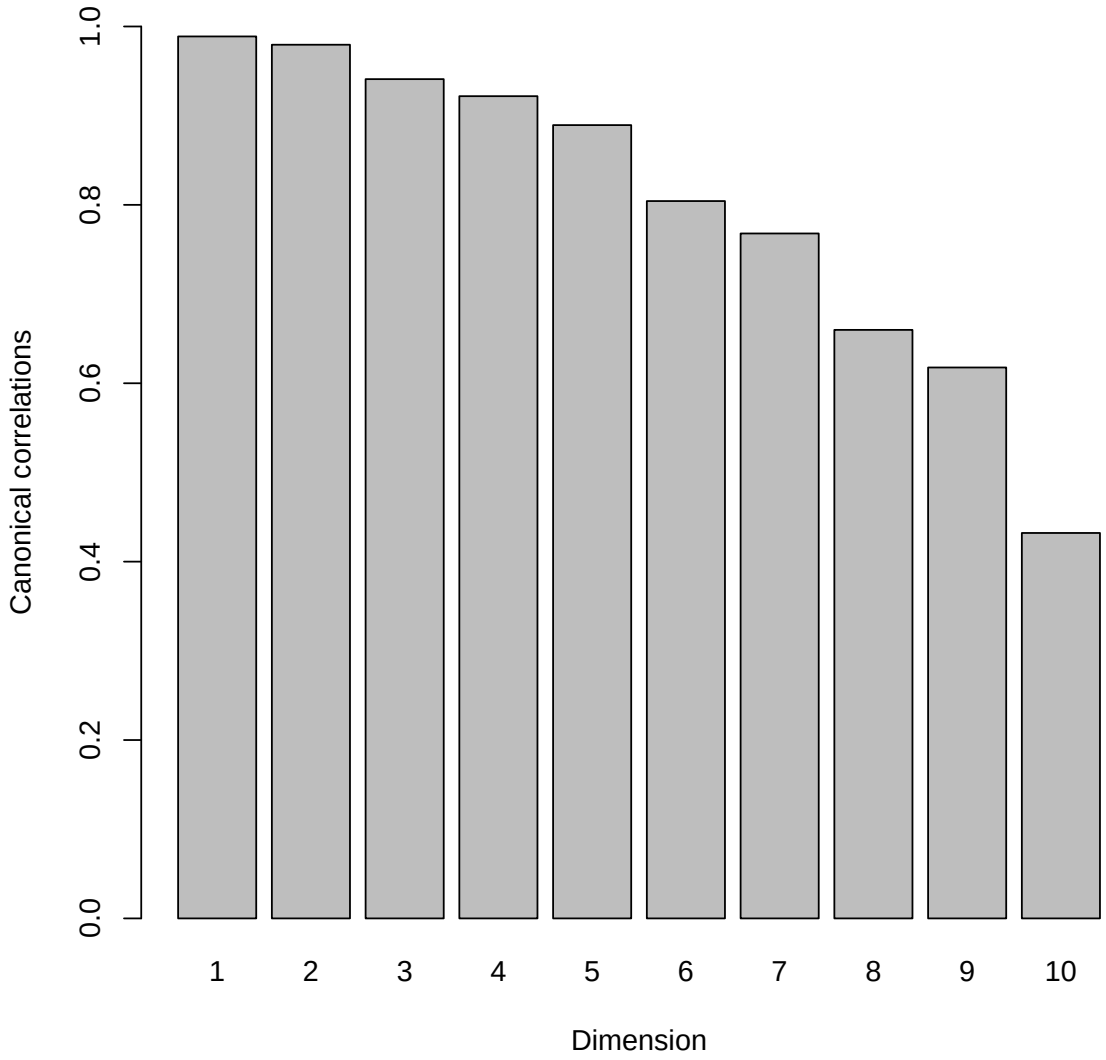


Figura 5.2: Correlación entre las variables canónicas, teniendo en cuenta los 10 índices calculados.

También podemos ver los nombres de las variables que se han seleccionado de ambos conjuntos para llevar a cabo el análisis.

```
res.cc$names

## $Xnames
##  [1] "CYP8b1"  "CYP27a1" "RARa"    "Lpin3"   "PMDCI"
##  [6] "apoA.I"  "CYP3A11" "RXRb2"   "SHP1"    "CYP2b10"
##
## $Ynames
##  [1] "C14.0"    "C16.0"    "C18.0"    "C16.1n.9" "C16.1n.7"
##  [6] "C18.1n.9" "C18.1n.7" "C20.1n.9" "C20.3n.9" "C18.2n.6"
## [11] "C18.3n.6" "C20.2n.6" "C20.3n.6" "C20.4n.6" "C22.4n.6"
## [16] "C22.5n.6" "C18.3n.3" "C20.3n.3" "C20.5n.3" "C22.5n.3"
## [21] "C22.6n.3"
##
## $ind.names
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11"
## [12] "12" "13" "14" "15" "16" "17" "18" "19" "20" "21" "22"
## [23] "23" "24" "25" "26" "27" "28" "29" "30" "31" "32" "33"
## [34] "34" "35" "36" "37" "38" "39" "40"
```

Tras esto, aparecen los coeficientes para cada conjunto de variables. Esto en teoría se corresponde con los valores de $\underline{a}_i$ y $\underline{b}_i$ para cada uno de los 10 índices que hemos obtenido, respectivamente.

Con la función `head` hemos reducido la salida de R para mostrar tan sólo los valores para las primeras variables de cada conjunto.

```
head(res.cc$xcoef)
head(res.cc$ycoef)
```

```
              [,1]        [,2]       [,3]        [,4]       [,5]
CYP8b1  -0.04317032 -0.13249783 -0.1225827  0.03851034  0.1431669
CYP27a1  0.13837072 -0.06704582  0.6847454 -0.90146905 -0.5080292
RARa    -0.06755834  0.36427522  0.6446277  0.74865435 -0.5921778

             [,1]        [,2]        [,3]        [,4]        [,5]
C14.0    -3.709529  -12.235375  -19.771392  -20.354826  -13.501884
C16.0   -20.181158  -52.688694  -86.390523  -97.289840  -71.384173
C18.0   -14.229198  -37.673860  -63.710750  -70.225616  -53.371176
```

A continuación, se muestran algunos de los valores que toman los individuos cuando introducimos los coeficientes obtenidos.

```
head(res.cc$scores$xscores)
```

```
       [,1]        [,2]       [,3]        [,4]       [,5]
1 0.6203762 -1.10043552 -0.8632726  0.28122447 -1.1996885
2 1.3464731 -0.29545361  0.2785762  0.37691778  0.6708977
3 1.0441307 -0.33255241 -0.4895570  0.06214316  1.8445435
```

```
head(res.cc$scores$yscores)
```

```
       [,1]       [,2]       [,3]       [,4]       [,5]
1 0.5162834 -1.3123080 -1.1923051  0.4132465 -0.8649199
2 1.3433274 -0.1746114  0.2325537  0.6365109  1.6488218
3 0.9831786 -0.3782187 -0.4989220  0.1365050  1.0234334
```

También se muestran las correlaciones entre las variables originales y las de correlación canónica. Se muestran algunos valores de $cor(\underline{X}, \underline{a}_i'\underline{X})$; aunque se pueden obtener también los de $cor(\underline{Y}, \underline{a}_i'\underline{X})$, $cor(\underline{X}, \underline{b}_i'\underline{Y})$, $cor(\underline{Y}, \underline{b}_i'\underline{Y})$.

```
head(res.cc$scores$corr.X.xscores)
```

```
##                [,1]         [,2]        [,3]        [,4]
## CYP8b1   0.09787891 -0.311267976 -0.02742548 -0.09783216
## CYP27a1  0.52252379 -0.227431346  0.47454092 -0.06563921
## RARa     0.07793188  0.237850465  0.42938647  0.45099119
## Lpin3   -0.03257523  0.262348515  0.10301516  0.40953864
## PMDCI    0.75313911 -0.601752574  0.03463925  0.24570849
## apoA.I   0.69139446 -0.000527109  0.03644501 -0.06229574
##                [,5]        [,6]        [,7]        [,8]
## CYP8b1   0.38206052 0.730094529 0.01819970 -0.42071311
## CYP27a1 -0.07135934 0.376644816 0.25243205  0.04629565
## RARa     0.14691865 0.219137273 0.37205274 -0.24975868
## Lpin3    0.19061760 0.114730180 0.22379031 -0.26325679
## PMDCI   -0.06771857 0.001600979 0.01355798  0.04560853
## apoA.I   0.35553712 0.137263174 0.58591644 -0.14603883
```

```
##                  [,9]        [,10]
## CYP8b1  -0.155085083  0.053087487
## CYP27a1 -0.005629969 -0.482592885
## RARa     0.260445559 -0.459725158
## Lpin3   -0.092898047 -0.757801903
## PMDCI    0.046948637 -0.006803365
## apoA.I  -0.034246900 -0.075619656
```

```
res.cc$scores$corr.Y.xscores
res.cc$scores$corr.X.yscores
res.cc$scores$corr.Y.yscores
```

Con la función `plt.cc` se representan variables e individuos.

Variables e individuos pueden representarse por el par (η_s, η_t) o bien por (φ_s, φ_t); con $1 \leq s < t \leq m$.

Dado que las correlaciones entre las dos primeras variables canónicas son cercanas a 1, las representaciones gráficas en los ejes definidos por ambas (η_1, η_2) y (φ_1, φ_2) son similares y hemos representado tanto variables como individuos haciendo uso de (η_1, η_2).

En el gráfico A se representan las correlaciones entre variables, diferenciando dos círculos (uno de radio 0.5 y otro de radio unidad) que revelen los patrones más destacados en el anillo entre las dos circunferencias.

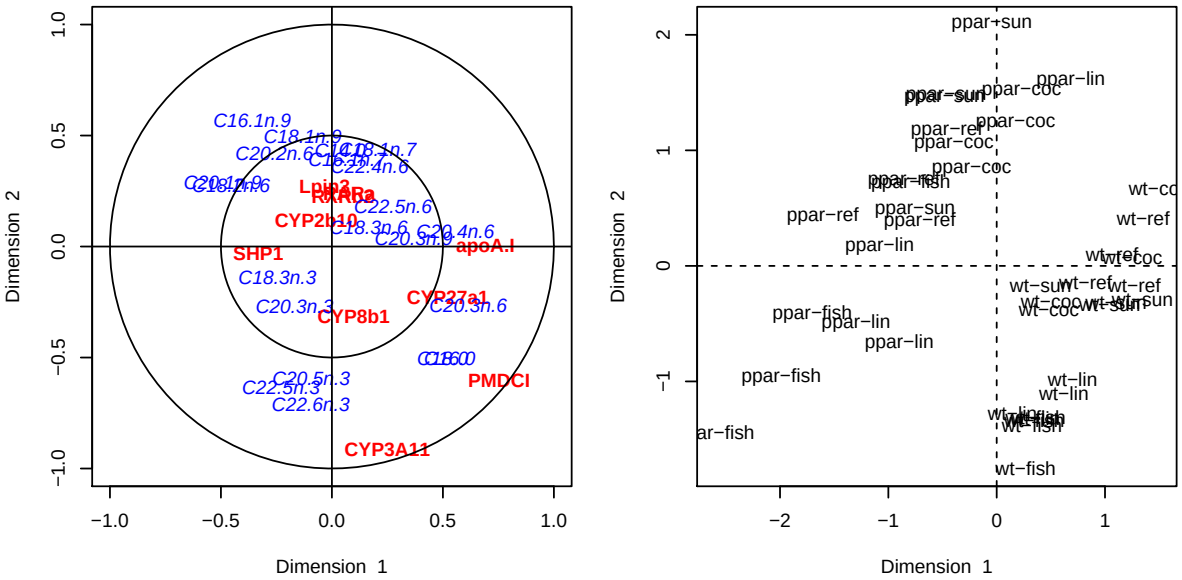
Los puntos representados en rojo indican las correlaciones entre las 10 variables de **X** y las dos primeras variables de correlación canónica $\underline{a}_1'\underline{X}$ y $\underline{a}_2'\underline{X}$, por tanto las coordenadas de cada punto son $(cor(\underline{a}_1'\underline{X}, \underline{X}), cor(\underline{a}_2'\underline{X}, \underline{X})) = (cor(\underline{\eta}_1, \underline{X}), cor(\underline{\eta}_2, \underline{X}))$.

Sin embargo, los triángulos indican la correlación de las 21 variables de la matriz **Y** con las variables de correlación canónica $\underline{a}_1'\underline{Y}$ y $\underline{a}_2'\underline{Y}$, siendo por tanto sus coordenadas:
$(cor(\underline{a}_1'\underline{X}, \underline{Y}), cor(\underline{a}_2'\underline{X}, \underline{Y})) = (cor(\underline{\eta}_1, \underline{Y}), cor(\underline{\eta}_2, \underline{Y}))$

Se asume que las variables X_j e Y_k tienen varianza unidad, sus proyecciones en el plano están dentro de un círculo de radio 1 centrado en el origen, llamado el círculo de correlación.
Variables con una fuerte relación se proyectan en la misma dirección desde el origen. Mientras más distancia hay al origen más fuerte es la relación.

En el gráfico B se representan los individuos caracterizados por el genotipo "ppar" o "wt" y la dieta que siguen. Son las puntuaciones de los individuos para la combinación de la variable $\underline{X}$, representadas en las dos primeras dimensiones ($\underline{a}_1'\underline{X}, \underline{s}_2'\underline{X}$)=($\eta_1$,$\eta_2$).

```
plt.cc(res.cc, var.label = TRUE, ind.names =
    paste(nutrimouse$genotype,nutrimouse$diet, sep = "-"))
```



Es sencillo observar que la variable de correlación canónica η_1 (dimensión 1) separa los individuos por genotipo ya que mientras que en el lado izquierdo están representados los ratones $PPAR_\alpha$ deficientes, en el derecho aparecen los ratones de tipo salvaje.
Por otra parte, la variable de correlación canónica η_2 (dimensión 2) separa los tipos de dietas en función de su composición. La separación de dietas en los ratones de tipo $PPAR_\alpha$ es menos precisa.

Las relaciones entre los dos gráficos pueden revelar asociaciones entre variables y unidades.

Capítulo 6

Paquetes de R

En este capítulo se hace un listado de todos los paquetes de R utilizados en las aplicaciones prácticas de la memoria, explicando sus características principales. También se han detallado las funciones de las que hemos hecho uso y los argumentos que se han considerado relevantes.

Paquete "base"

El paquete *base* es el que ejecuta las funciones básicas de R.

Funciones

1. `as.data.frame`: Comprueba si un objeto es data.frame y si no lo es, lo transforma.
2. `as.factor`: Comprueba si un objeto es un factor y si no lo es, lo transforma.
3. `as.matrix`: Comprueba si un objeto es una matriz y si no lo es, lo transforma.
4. `cbind.data.frame`: Toma la secuencia de los argumentos de un data.frame y combina sus columnas.
5. `class`: Muestra la clase a la que pertenece un objeto.
6. `colnames`: Conjunto de nombres de las columnas de un objeto de tipo matriz.
7. `cumsum`: Calcula la suma acumulada de los valores de su argumento.
8. `data`: Especifica un conjunto de datos.
9. `dim`: Dimensión de un objeto.
10. `file.path()`: Construye el camino a un fichero desde componentes en otra plataforma.
11. `getwd()`: Devuelve el directorio de trabajo que se está usando.
12. `length`: Longitud de un objeto.
13. `library` & `require`: Se cargan y adjuntan los paquetes especificados.
14. `names`: Nombres de un objeto.

15. `options`: Permite al usuario fijar funciones globales que afectan a la forma en que se computan los resultados.
    - $digits$: controla la cantidad de dígitos cuando se trabaja con valores numéricos.
    - $width$: controla el número de columnas máximo por línea para mostrar vectores y matrices.
16. `paste`: Concatena vectores tras convertirlos en caracteres.
17. `read.table`: Lee un fichero en formato tabla y crea un "data.frame" a partir de él, cuyos casos se corresponden con las filas y las variables con los campos en el fichero.
    - dec: El caracter a usar con los puntos decimales.
    - $header$: Valor lógico indicando si los nombres de las variables se asignan a la primera fila.
    - $row.names$: Puede ser un vector con los nombres de las filas o bien un número con la columna de la tabla que contiene los nombres de las filas.
    - sep: Separador de caracteres.
18. `rownames`: Conjunto de nombres de las filas de un objeto de tipo matriz.
19. `sample`: Toma una muestra de un tamaño especificado de los elementos de un objeto, con o sin reemplazamiento.
20. `sapply`: Devuelve un vector o matriz de la misma dimensión que la dada, resultado de aplicar una función especificada.
21. `scale`: Función genérica que centra y escala las columnas de una matriz numérica.
22. `str`: Alternativa a `summary` que muestra la estructura de un objeto.
23. `sum`: Calcula la suma de los valores de su argumento.
24. `summary`: Resúmenes de resultados de varios modelos.
    - $nbelements$: Número de elementos a mostrar.
25. `table`: Usa factores de clasificación cruzada para construir una tabla de contingencia de los elementos de cada combinación de los factores.
26. `View`: Permite visualizar cualquier objeto de R con forma de matriz.

Paquete "CCA"

CCA proporciona un conjunto de funciones que extienden la función `cancor()` del paquete *stats* con nuevas salidas numéricas y gráficas.
La función `cancor()` calcula las correlaciones canónicas entre dos matrices de datos. El paquete *CCA* incluye una extensión que permite trabajar con más variables que observaciones, así como con valores perdidos.

Funciones

1. `barplot`: Crea una representación gráfica con barras verticales u horizontales.
   - $names.arg$: Vector de nombres para ser representado bajo cada barra o grupo de barras.
   - $xlab$: nombre eje X.
   - $xlim$: límites eje X.
   - $ylab$: nombre eje Y.
   - $ylim$: límites eje Y.
2. `cc`: Ejecuta un Análisis de Correlación Canónica. Completa la función `cancor()`.
3. `img.matcor`: Muestra imágenes de las matrices de correlación entre 2 conjuntos de datos.
   - $type$: Caracter determinando la clase de plot. "type=1" representa una matriz $(p+q) \times (p+q)$ y "type=2" 3 matrices $(p \times q), (q \times q)$ y $(p \times p)$.
4. `matcor`: Matrices de correlación entre dos conjuntos de datos.
5. `plt.cc`: Utiliza la función `plt.var()`, `plt.ind()` o ambas para representar individuos, variables o ambos en un gráfico con ejes las variables canónicas.
   - $ind.names$: Vector conteniendo los nombres de los individuos.
   - $var.label$: Valor lógico indicando si el eje debe incluirse en la representación de variables.

Paquete "FactoMineR"

FactoMineR permite realizar análisis exploratorio de datos multivariantes y minería de datos.

Funciones

1. `PCA`: LLeva a cabo un Análisis de Componentes Principales de los datos.
   - $graph$: Toma valores lógicos. Indica si incluiremos gráficos o no.
   - $qualisup$: Variables cualitativas suplementarias.
2. `plot`: Representa gráficos.
   - $choix$: Elegimos qué representar (variables o individuos).
   - $habillage$: Damos color a los elementos representados en función del valor que tome este parámetro.
   - $lim.cos2.var$: Indica la calidad mínima de la proyección de los objetos que dibujaremos.
3. `plotellipses`: Dibuja las elipses de confianza para cada clase de una variable categórica. El nivel por defecto es $0{,}95$.
4. `substr`: Extrae o sustituye subcadenas de un vector de caracteres.

Paquete "graphics"

Graphics reune un conjunto de funciones para gráficos básicos.

Funciones

1. `hist`: Genera un histograma con los datos proporcionados.
   - $plot$: Toma valores lógicos. Con FALSE se devuelve una lista con los puntos para ambos ejes ("breaks" y "counts") en lugar de representar el histograma. TRUE es el valor por defecto.
2. `par`: Ajusta parámetros gráficos
   - $mfrow$: Indica cómo se dispondrán los plots por filas.
3. `pairs`: Produce una matriz de scatterplots.
   - cex: Valor numérico con la cantidad por la que se rigen el texto y los símbolos representados. Comienza con el 1.
   - $diag.panel$: Función para aplicar en las diagonales.
   - $panel$: Función que se usa para representar el contenido de cada panel.
   - pch: Símbolo para usar en el plot.
   - $xlim$: Límites eje X.
   - $ylim$: Límites eje Y.
4. `plot`: Representa gráficos.
5. `text`: Añade texto a un plot.

Paquete "HSAUR3"

Este paquete proporciona funciones, conjuntos de datos, análisis y ejemplos de la tercera edición del libro *A Handbook of Statistical Analysis using R* (Torsten Hothorn and Brian S. Everiitt, Chapman & Hall, CRC, 2014).

Paquete"knitr"

knitr se usa como herramienta alternativa a *Sweave*. Mientras que *Sweave* es un sistema que proporciona un marco de trabajo para generar documentos que combinan simultáneamente texto y código R, con el paquete *knitr* se consigue lo mismo mediante un diseño más flexible y características nuevas; entre ellas la posibilidad de tener un control más preciso de los gráficos.

Funciones

1. `opts_chunk`: Opciones para los diferentes chunk de R, siendo un chunk cada una de las partes de nuestro documento en las que introducimos código de R.
2. `opts_chunk$set()`: Suele utilizarse en el primer chunk del documento, donde se fijan las opciones globales para los siguientes chunk.
   - $concordance$: Toma valores lógicos. Permite escribir ficheros de concordancia, permitiendo así situar los números de las líneas del código con los de las líneas de los resultados, siendo esto de gran ayuda cuando tenemos un error al generar el PDF.

También podemos usar la sintaxis de este paquete para manejar cómo incluir el código R en nuestro documento. Las opciones disponibles son:

- $cache$: FALSE por defecto. Guarda los resultados en cache.
- $comment$: Caracter para comentarios.
- $echo$: TRUE por defecto. Muestra el código y los resultados.
- $error$: FALSE por defecto. Muestra errores.
- $eval$: TRUE por defecto. Evalúa el código e incluye los resultados.
- $fig.height$: 7 por defecto. Alto en pulgadas para las figuras.
- $fig.width$: 7 por defecto. Ancho en pulgadas para las figuras.
- $message$: TRUE por defecto. Muestra mensajes.
- $results$: "markup" por defecto. Controla la forma en que se muestran los resultados al compilar el documento a PDF. Puede ser "markup", "hide", "asis", y "hold".
- $tidy$: FALSE por defecto. Muestra el código de forma organizada.
- $warning$: TRUE por defecto. Muestra advertencias.

Paquetes "ade4" y "made4"

ade4 es un paquete desarrollado por el laboratorio de Biometría y Biología Evolucionaria (UMR 5558) de la Universidad de Lyon. Contiene funciones para analizar datos ecológicos y ambientales en el marco de métodos exploratorios euclídeos.
made4 necesita *ade4*. Además, facilita el análisis multivariante de datos de expresión génica obtenidos con microarrays.

Funciones

1. `cia`: Análisis de Coinercia con dos conjuntos de datos.
2. `coord.ellipses`: Construye elipses de confianza
   - $bary$: Toma un valor lógico. Se calculan las coordenadas de la elipse alrededor del baricentro de los individuos.

3. `heatplot`: Utiliza la función `heatmap2`, usando por defecto un rango de colores de rojo a verde. Además, dibuja dendogramas de los casos y las variables, usando como métrica similitud entre correlaciones y clustering. Es útil para análisis exploratorio de los datos.
   - $classvec$: Factor o vector que describe las clases en filas o columnas de los datos.
   - $dend$: Indica si los dendogramas se dibujan para filas, columnas o ambos. También puede no dibujarse.
   - $scale$: Indica la escala a utilizar en función de para qué se dibujen los dendogramas. Por defecto es fila. Puede ser "none".
4. `ord`: Lleva a cabo un ACP ("pca"), Análisis de Correspondencias ("coa") o Análisis de Correspondencias Asimétrico ("nsc") en datos de expresión génica.
   - $type$: Caracter "coa", "pca" o "nsc" indicando la transformación de datos requerida. Por defecto es "coa".
5. `overview`: Representa un boxplot, histograma y dendograma de análisis jerárquico.
   - $classvec$: Vector o factor que describe las clases en columnas del conjunto de datos.
   - $labels$: Vector, etiquetas para las muestras de los plots.
6. `plotarrays`: Gráfico de proyecciones de las muestras coloreadas por grupos. Útil para visualizar coordenadas de arrays resultantes de distintos análisis de datos microarray.

Paquete "mclust"

Basado en clustering, clasificación y estimación de la densidad, incluyendo regularización Bayesiana.

Paquete "MVA"

Proporciona funciones, conjuntos de datos, análisis y ejemplos del libro *An Introduction to Applied Multivariate Analysis with R* (Brian S. Everitt & Torsten Hothorn, Springer, 2011)

Funciones

1. `bvbox`: Representa boxplot en dos dimensiones.
   - add: Se añade a un plot existente.

Paquete "stats"

El paquete *stats* proporciona funciones estadísticas.

Funciones

1. `cor`: Matriz de correlaciones entre columnas de dos conjuntos de datos.
2. `lm`: Ajusta modelos lineales. Puede usarse para modelos de regresión, análisis de varianza y de covarianza.
3. `princomp`: Lleva a cabo un Análisis de Componentes Principales. Devuelve un objeto del tipo princomp.
   - *cor*: Valor lógico que indica si se usa o no la matriz de correlación.

Bibliografía

[1] Cuadras CM. (2014). Nuevos Métodos de Análisis Multivariante. CMC Editions.

[2] Culhane A. (2014). Introduction to Multivariate Analysis of Microarray Gene Expression Data using MADE4.

[3] Culhane A.(adapted by Sánchez A.) (2015) Cross-platform data integration using Coinertia Analysis.

[4] Culhane A., Perrière G, G. Higgins D. (2003) Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. BMC Bioinformatics.

[5] Culhane A., Thioulouse J., Perriere G., Higgins DG. (2005) MADE4:an R package for multivariate analysis of gene expression data Bioinformatics 21(11): 2789-90.

[6] Dufour A.B. (2008) Enseignements de Statistique en Biologie. URL: http://pbil.univ-lyon1.fr/R/pdf/course6.pdf

[7] Everitt B., Hothorn T. (2015). An Introduction to Applied Multivariate Analysis with R. Springer.

[8] González I., Déjean S., G.P. Martin P., Baccini A. (2008) CCA: An R Package to Extend Canonical Correlation Analysis. Journal of Statistical Software.

[9] González I., Déjean S. (2012). CCA: Canonical correlation analysis. R package version 1.2. http://CRAN.R-project.org/package=CCA

[10] Härdle W.K., Simar L. (2015). Applied Multivariate Statistical Analysis. Springer-Verlag.

[11] Husson F., Josse J., Le S., Mazet J. (2015). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining. R package version 1.31.4. http://CRAN.R-project.org/package=FactoMineR

[12] Jobson J.D.(1992). Applied Multivariate Data Analysis.
Volume II:Categorical and Multivariate Methods. Springer-Verlag.

[13] Johnson R. A., Wichern D.W. (2002). Applied Multivariate Statistical Analysis. Pearson Education, Prentice Hall.

[14] Lê Cao K., GP Martin P., Christèle-Granié, Besse P. (2009) Sparse canonical methods for biological data integration: application to a cross-platform study. BMC Bioinformatics.

[15] Lê S., Josse J., Husson F. (2008). FactoMineR: an R package for multivariate analysis. Journal of Statistical Software.

[16] Mardia K.V., Kent J.T., Bibby J.M. (1979) Multivariate Analysis. Academic Press.

[17] Peña D.(2002) Análisis multivariante de datos. MC GRAW HILL INTERAMERICANA.

[18] 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/.

[19] R Studio URL: http://www.rstudio.org/

[20] Sánchez A. Multivariate Methods for the integrate analysis of omics data. (1): Exploratory Data Analysis. URL: http://eib.stat.ub.edu/Integrative+Analysis+of+Omics+Data

[21] Sánchez A. Multivariate Methods for the Integration and Visualization of Omics Data. URL: http://eib.stat.ub.edu/Integrative+Analysis+of+Omics+Data

[22] Santamaria R. Análisis de Datos de Microarray. Universidad de Salamanca.